Universitat Pompeu Fabra Barcelona

Neural Networks For Singing Voice Extraction In Monaural Polyphonic Music Signals

Pritish Chandna

TESI DOCTORAL UPF / 2021

Thesis Directors:

Dr. Emilia Gómez

Music Technology Group

Dept. of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona, Spain

Dissertation submitted to the Department of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfillment of the requirements for the degree of

DOCTOR PER LA UNIVERSITAT POMPEU FABRA

Music Technology Group (http://mtg.upf.edu), Department of Information and Communication Technologies (http://www.upf.edu/dtic), Universitat Pompeu Fabra (http://www.upf.edu), Barcelona, Spain.

The doctoral defense was held on ..... at the Universitat Pompeu Fabra and scored as .....

Dr. Emilia Gómez Gutierrez

(Thesis Supervisor)

Universitat Pompeu Fabra (UPF), Barcelona, Spain

Dr. Antoine Liutikus

(Thesis Committee Member)

Inria, Montpellier, France

Dr. Marius Miron

(Thesis Committee Member)

Universitat Pompeu Fabra (UPF), Barcelona, Spain

Dr. Estefanía Cano

(Thesis Committee Member)

AudioSourceRE, Singapore

This thesis has been carried out at the Music Technology Group (*MTG*) of Universitat Pompeu Fabra in Barcelona, Spain, from Feb. 2017 to Mar. 2021. It is supervised by Dr. Emilia Gómez.

This work has been supported by the Department of Information and Communication Technologies (*DTIC*) *PhD* fellowship (2017-21), Universitat Pompeu Fabra, the *TROMPA* (TROMPA H2020 770376) Project.

Acknowledgments

The PhD journey over the last four years was a transformative and challenging one. I am grateful to have received support and guidance from a plethora of people both academic and personal during this period. I am extremely grateful to Dr. Emilia Gómez for having recognized within me a potential that. even I never thought possible and to Dr. Xavier Serra for accepting me into the masters' program where the journey began. I was lucky to share an office with the brilliant minds of Dr. Jordi Bonada and Merlijn Blaauw, from whom I tried to absorb as much knowledge as I could.

The Music Technology Group is filled with exceptional talent that I had the privilege to interact with over the last six years. Such interactions fostered friendships which will last a lifetime and which I will always be thankful for. It is an impossible task to comprehensively list each individual, but I will always remember the time spent with Pablo Alonso, Dmitry Bogdano, Xavier Favory, Helena Custa, Marius Miron, Lorenzo Porcaro, Olga Slizovskaia, Jordi Pons, Rong Gong, Antonio Ramires and Furkan Yessler, among many many others.

Friendships are the strongest pillars of support that carries one through life and I would like to express my gratitude to all my friends in Barcelona and Delhi who have constantly supported me throughout the crests and troughs of life. To name but a few would be unfair and for that I will, with sincerest apologies, forgo the tradition of listing out names.

Finally, I would like to thank my parents, who encouraged and supported my decision to quit a job in the high paying financial sector pursue higher studies. My mother, Dr. Nirupma Chandna, is by far the kindest and strongest person I have come come to know in my life. Her love, nurturing and intellectual influence has made me everything I am

and will be in the future. Words fail me as I try to express my gratitude towards her and my father, Dr. Chandesh Chandna, to whom I would like to dedicate this thesis. The man was there for me, each step of my life from the day I was born, through my intolerable adolescence and the early stages of my adult life. Not a single day goes by that I do not remember his smiling face, wishing to see it again. His exceptional intellect and knowledge could only be matched by his compassion and values. I only hope to live up to be half the man he was. I remember his joy when I told him that I would try to add the doctor prefix to my name and it is a shame that he will not be with us when the day comes to pass. I hope not to disappoint and to honour and cherish his memory forever.

Abstract

This thesis dissertation focuses on singing voice extraction from polyphonic musical signals. In particular, we focus on two cases; contemporary popular music, which typically has a processed singing voice with instrumental accompaniment and ensemble choral singing, which involves multiple singers singing in harmony and unison.

Over the last decade, several deep learning based models have been proposed to separate the singing voice from instrumental accompaniment in a musical mixture. Most of these models assume that the musical mixture is a linear sum of the individual sources and estimate time-frequency masks to filter out the sources from the input mixture. While this assumption doesn't always hold, deep learning based models have shown remarkable capacity to model the separate sources in a mixture.

In this thesis, we propose an alternative method for singing voice extraction. This methodology assumes that the perceived linguistic and melodic content of a singing voice signal is retained even when it is put through a non-linear mixing process. To this end, we explore language independent representations of linguistic content in a voice signal as well as generative methodologies for voice synthesis. Using these, we propose the framework for a methodology to synthesize a clean singing voice signal from the underlying linguistic and melodic content of a processed voice signal in a musical mixture.

In addition, we adapt and evaluate state-of-the-art source separation methodologies to separate the soprano, alto, tenor and bass parts of choral recordings. We also use the proposed methodology for extraction via synthesis along with other deep learning based models to analyze unison singing within choral recordings.

Resum

Aquesta tesi se centra en l'extracció de veu cantada a partir de senyals musicals polifònics. En particular, ens centrem en dos casos; música popular contemporània, que normalment té una veu cantada processada amb acompanyament instrumental, i cant coral, que consisteix en diversos cantants cantant en harmonia i a l'uníson. Durant l'última dècada, s'han proposat diversos models basats en l'aprenentatge profund per separar la veu de l'acompanyament instrumental en una mescla musical. La majoria d'aquests models assumeixen que la mescla és una suma lineal de les fonts individuals i estimen les màscares temps-freqüència per filtrar les fonts de la mescla d'entrada. Tot i que aquesta assumpció no sempre es compleix, els models basats en l'aprenentatge profund han demostrat una capacitat notable per modelar les fonts en una mescla. En aquesta tesi, proposem un mètode alternatiu per l'extracció de la veu cantada. Aquesta metodologia assumeix que el contingut lingüístic i melòdic que percebem d'un senyal de veu cantada es manté fins i tot quan es tracta d'una mescla no lineal. Per a això, explorem representacions del contingut lingüístic independents de l'idioma en un senyal de veu, així com metodologies generatives per a la síntesi de veu. Utilitzant-les, proposem una metodologia per sintetitzar un senyal de veu cantada a partir del contingut lingüístic i melòdic subjacent d'un senyal de veu processat en una mescla musical. A més, adaptem i avaluem metodologies de separació de fonts d'última generació per separar les parts de soprano, contralt, tenor i baix dels enregistraments corals. També utilitzem la metodologia proposada per a l'extracció mitjançant síntesi juntament amb altres models basats en l'aprenentatge profund per analitzar el cant a l'uníson dins dels enregistraments corals.

(*Translated from English by Helena Cuesta*)

Resumen

Esta disertación doctoral se centra en la extracción de voz cantada a partir de señales musicales polifónicas de audio. En particular, analizamos dos casos; música popular contemporánea, que normalmente contiene voz cantada procesada y acompañada de instrumentación, y canto coral, que involucra a varios coristas cantando en armonía y al unísono.

Durante la última década, se han propuesto varios modelos basados en aprendizaje profundo para separar la voz cantada del acompañamiento instrumental en una mezcla musical. La mayoría de estos modelos asumen que la mezcla musical es una suma lineal de fuentes individuales y estiman máscaras de tiempo-frecuencia para extraerlas de la mezcla. Si bien esta suposición no siempre se cumple, los modelos basados en aprendizaje profundo han demostrado tener una gran capacidad para modelar las fuentes de la mezcla.

En esta tesis proponemos un método alternativo para extraer voz cantada. Esta técnica asume que el contenido lingüístico y melódico que se percibe en la voz cantada se retiene incluso cuando la señal es sometida a un proceso de mezcla no lineal. Con este fin, exploramos representaciones del contenido lingüístico independientes del lenguaje en la señal de voz, así como metodos generativos para síntesis de voz. Utilizando estas técnicas, proponemos la base para una metodología de síntesis de voz cantada limpia a partir del contenido lingüístico y melódico subyacente de la señal de voz procesada en una mezcla musical.

Además, adaptamos y evaluamos metodologías de separación de fuentes de última generación para separar las voces soprano, alto, tenor y bajo de grabaciones corales. También utilizamos la metodología propuesta para extracción mediante síntesis junto con

otros modelos basados en aprendizaje profundo para analizar canto al unísono dentro de grabaciones corales.

(*Translated from English by Pablo Alonso-Jiménez*)

Contents

List of Figures

List of Tables

Part I

Introduction

List of symbols

a A representation a general signal.

A Spectogram of the signal denoted by **a**.

c A representation of a general signal, different from **a**.

C Spectrogram of the signal denoted by **c**.

s The time domain waveform of an arbitrary source mixed in a musical mixture.

S Spectrogram of the source denoted by **s**.

x Time-domain waveform of voice signal, could be speech or singing.

X Spectrogram of the voice signal denoted by **x**.

$\mathbf{X_{voc}}$ Compressed spectral envelope pertaining the voice signal denoted by **x**.

$\mathbf{X_{mel}}$ Mel-scale spectrogram pertaining to the voice signal denoted by **x**.

y Time-domain waveform of voice signal with modulations added.

Y Spectrogram of the voice signal with modulations added, **y**.

$\hat{\mathbf{x}}$ Time-domain waveform of an output voice signal.

$\hat{\mathbf{X}}$ Spectrogram of the output voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{X}}_{\mathbf{voc}}$ Compressed spectral envelope pertaining to the voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{X}}_{\mathbf{mel}}$ Mel-scale spectrogram pertaining to the voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{y}}$ Time-domain waveform of an output voice signal, which has effects and modulations added.

$\hat{\mathbf{Y}}$ Spectrogram of the output voice signal with modulations added, $\hat{\mathbf{y}}$.

b Time-domain waveform of musical instrumental backing track.

B Spectogram of musical instrumental backing track

m The mixture signal formed by mixing **y** with **b**, the mix does not necessarily have to be a linear mixture.

M Spectogram of the mixture signal denoted by **m**.

enc The encoder network of an autoencoder.

dec The decoder network of an autoencoder.

V The latent embedding of an autoencoder.

gen The generator network of a GAN.

dis The discriminator network of a GAN.

Z The linguistic content of the voice signal, **x**.

η The melodic content of the voice signal, **x**.

ψ A representation of a singer or speaker, who is the source of **x**.

ω A soft-mask or Wiener filter used for source separation.

Chapter 1

Introduction

The voice has been one of the fundamental means of communication between humans since the emergence of our species. The ability to produce and distinguish distinct sequences of sounds has allowed for the formation of language and the subsequent the evolution of society and technology. While language and speech are the de facto means of human communication, there are some abstract concepts, like emotions (Scherer et al., 2003), that are not well encapsulated by language alone, leading to a so called *semantic gap*. Music, and in particular singing, provides a bridge across this gap and has been practiced both as a means of social entertainment and the passage of ideas across geographies and generations.

Various forms of singing exist across cultures around the world; from the *katajjaq* singing of the *Inuit* tribes of the northern Tundra regions of Canada to the polyphonic yodeling of the Ituri people in the forests of the Democratic Republic of Congo (Potter & Sorrell, 2012). Evidence has been found that singing traditions existed since ancient times, allowing for communication of ideas from mouth to mouth before the invention of written textual culture. While song and music are generally dismissed as a leisurely or extracurricular activities with limited value, singing has often served as a powerful means of spreading revolutionary propaganda especially in opposition to oppression. The French revolution in the 1790s was inspired by several songs including the famous

La Marseillaise, and other hymns such as *Chant du départ* and *Carmagnole*. The Spanish socialist revolution was accompanied by the *A Las Barricadas* anthem, whereas the American struggle for independence was inspired by songs such as *Free America* and *Poor Old Tory*. The Indian equivalent of the same also had its revolutionary counterparts including *Aye mere watan ke logo* and *Ye desh hai veer jawaano ka*. Even in modern times, singing retains an important cultural significance. Various artists like the *Sex Pistols*, *Bob Dylan* and more recently, *Donald Glover* and *Kendrick Lamar* have used singing and music to bring attention to social injustice and invoke political change through massive public movements. Such artists use the singing voice as an instrument within the music they compose to convey information that words alone are not sufficient to transmit. In addition, music and singing has been noted to have therapeutic effects related to speech deficits associated with conditions such as Parkinson's disease, autism and brain lesions (Monroe et al., 2020; Wan et al., 2010). Such activity has been linked to the brain's emotional and cognitive functioning, and has been explored in the Banda Sonora Vital (Personal Life Soundtrack) project (Navarro, 2013).

As such, analysing the singing voice provides an avenue into understanding human nature. Singing and music have had an important role to play in man's evolution. While Charles Darwin proposed the the role of music was to attract the opposite sex, philosophers such as Pythagoras believed that studying music and the mathematical structures within was fundamental to understanding the fundamental concepts of reality. Indeed the first analysis of the human voice was of a ten-second fragment of the French folk song *Au Clair de la Lune*, by Edouard-Leon Scott (de Martinville, 1860) in 1860. The French inventor used stenographic device to inscribe the waveform of the sound produced while singing the song on a glass plate. Since then, audio recording and analysis technologies have improved leaps and bounds and have evolved into a field of research and application known commonly as *audio signal processing* or audio processing (AP). Researchers have used audio processing to represent and understand intrinsic aural structures present in the voice, providing insight into the perceptual qual-

ities of the voice. This understanding has opened the door for several practical branches of audio processing for the **voice** including:

- *Source separation* which deals with separating the voice signal from other signals like noise or musical instruments. The sub-branch applied to musical signals is often referred to as *musical source separation* and involves separating the various instruments in a musical mixture. *Speech source separation* aims the signals pertaining to the voices to two or more speakers speaking asynchronously at the same time.

- *Voice synthesis* is the task of generating a speech signal given certain parameters like linguistic content and a speaker identity. It includes *text to speech* (TTS) synthesis and *singing voice synthesis* (SVS).

- *Music Information Retrieval* is the branch of signal processing the extract information from a musical signal in a way that it can be interpreted by humans. In the case of the singing voice, this can include information like the melody, linguistic information and the identity of the speaker.

- *Voice conversion* is the task of changing the perceived speaker of the voice signal while retaining the intelligible linguistic content.

While this thesis utilizes concepts from each of these branches, it primarily focuses on source separation. Source separation is a particularly important and well researched branch of audio signal processing that aims to separate the individual sound sources from a mixture of sources. Although the field has received attention from researchers through the last century, data-driven deep learning approaches have led to a significant improvement in performance and results over the last decade. Such algorithms generally assume that the mixture is a linear sum of the individual sources and use spectral filtering to separate the sources. Such a process is based on the auditory masking process, wherein an audio source that has dominant energy in a frequency band masks

other sources within the band (Moore et al., 2009). The limitation of such methodologies is that they can only extract the processed form of the vocal signal present in the mixture, which may or may not be desirable for the end application.

This thesis aims to overcome this limitation for singing voice extraction in two different contexts; contemporary popular music which includes the singing voice with effects combined with an instrumental accompaniment and ensemble choral singing. In Part II, we propose a methodology to synthesize the vocal signal present in the input mixture such that the perceptual content of the signal is retained in the output. This methodology is motivated by the analysis by synthesis theory of speech perception (Stevens, 1972, 1960), which states that humans use mental synthesis to identify the linguistic content of a speech signal. We assume that the perceived content of a singing voice signal remains unchanged when it is mixed with instrumental accompaniment. To this end, we study speech and singing voice synthesis methodologies, representations of content used for synthesis and information retrieval techniques to extract such content from a signal.

In Part III, we apply state-of-the-art (SOTA) source separation algorithms in conjunction with our proposed methodology to separate the individual voices in ensemble choral singing in the soprano, alto, tenor bass format. This format of choir singing involves groups of multiple singers singing simultaneously in four parts which are arranged in harmony. Within each part, there might be multiple singers singing in unison. As the mixture of the 4 distinct parts can be assumed to be a linear mixture, we adapt mask-based source separation algorithms to separate these parts from a mixture recordings. The individual voices within a unison mixture are indistinguishable, but are perceived to be singing the same melodic and linguistic content. As such, we use the methodology proposed in Part II to model the unison singing signal.

In the rest of this chapter, we look at past analysis of the voice including the production mechanism, tools used for analysis, distinguishable features, differences between speech and singing and methodologies for separating the voice from other sources. We

then discuss the motivation for the research carried out in this thesis and the context in which it can be applied. This is followed by a list of contributions and the objectives and structure of the rest of the thesis.

1.1 The voice as a signal

The media for distribution and transmission have evolved over the centuries. While early music was transmitted purely by word-of-mouth means, written representations were invented and iteratively formalized to represent the content of musical pieces known as songs. Such a representation is called a *score*. The oldest known form of a score can be dated back to 1400 BC (Kilmer & Civil, 1986) and provides instructions for performing a piece of music in an organized melodic format. The commonly accepted modern form a score developed around the 14th century and provides a system for representation of melodic, rhythmic and linguistic content to be sung by the singer or played on various instruments for a performance of a song. Scores generally provide guidelines for singing and can be interpreted by artists.

More recent advancements of technology have allowed us to represent audio and music in a reproducible machine readable format, known as a **signal**.The act of encapsulating the audio information in such a format is called **recording**. Music and singing it self has evolved with such technology, which allows opens new avenues for understanding and producing audio, voice and music. In the next section, Section 1.1.1, we briefly discuss the production mechanism of the human voice, followed by a summary of the methodologies used for analysis of the same.

1.1.1 Voice production mechanism

The human voice is generated by a combination of the lungs, the larynx, the pharynx, the nose and the mouth. The lungs start out by generating an air-stream through an excess of pressure (Sundberg & Rossing, 1990). The larynx contains mucous membrane lining, which are commonly known as vocal cords or folds and the opening between

them is known as the glottis. These folds open and close as the air-stream passes through them, creating a pulse, with a frequency that depends on the air pressure in the lungs as well the vocal folds. This pulse passes through the vocal tract, consisting of the larynx, the pharynx and the mouth. The tract acts as a resonance chamber, which amplifies certain frequencies. This leads to a periodic waveform with a distinct harmonic spectral structure. Sounds created in this manner are generally termed as *voiced* sounds and include vowel sounds and can be sustained over long periods of time.

Other mechanisms for voice sounds creation include closing various parts of the mouth structure like the tongue, alveolar ridge and the lips, to stop the flow of air. Such sounds are termed as *plosives*. Sounds which are created using partially blocked air flow include *fricatives* and *affricates*. *Nasal* sounds are created by diverting the flow of air through the nose instead of the mouth while flow of air through the sides of the tongue generates *lateral* sounds. *Approximant* sounds are created with interactions between the tip of the tongue and the alveolar area of the mouth. Such sounds are typically classified as *consonants*. While most consonant sounds are *unvoiced*, some like the fricatives are classified as voiced.

1.1.2 Analyzing the voice

Like all sounds, the human voice propagates through the air through fluctuations in air pressure. When these fluctuations reach a human ear, they cause the tympanic membrane inside to vibrate. These vibrations are transmitted through the cochlea, which contains a Reissner's membrane and basilar member, the later of which transduces the vibrations into neural activity, via inner hair cells, different groups of which react to distinct frequencies present in the vibrations, through a phenomenon known as *phase locking*. Following the encoding of the physical vibrations by the ear mechanism, it has been observed that the human central nervous system uses a tonotopic (Young, 2008) representation for interpretation of the audio received by the ear. Such a representation consists of frequency-based clustering in the activation of the neural population in the

system.

A similar mechanism for recording and analysing sounds is used technologically. A microphone used for recording sounds consists of a membrane that vibrates, deviating from the central position. Such deviations are sampled at regular intervals [1] and the resulting recorded signal is termed as a waveform. Signals are sequences of observations which are distributed over time. In signal processing, it is common to use $n \in \{1,2,3...N\}$ as a representation of discrete time and $t \in \{1,2,3...T\}$ as that of continuous time. For the context of this thesis, we will cover discrete-time signals, generally represented by the symbols **a** and **c**. The signal pertaining to the clean voice is represented by **x**. A sample of the voice signal, **x** at time n is represented as $\mathbf{x}(n)$. However, for the sake of convenience, we will use the shorthand **x**.

Modelling the human nervous model, a computationally recorded waveform is often analyzed through convolutions with complex sinusoidal basis functions. This operation, known as the Fourier Transform results in the projection of the signal onto the basis function, and gives us a distribution of the audio signal over various frequencies. The same operation carried over overlapping windowed frames allows us to analyze the evolution of the frequency distribution in the signal over time and is known as the Short-Time Fourier Transform (STFT) (McAulay & Quatieri, 1986). The resulting complex matrix has one axis representing time and the other frequency and is often called the **spectrogram**. We refer to the spectrogram of a signal **a** as **A** and for the voice signal, **x** as **X**. Generally, the magnitude part of the spectrogram, $|\mathbf{A}|$, can be used to visualize the distribution of energy across frequency bands and time in a signal. The logarithm of the spectrogram is often used for visualisation and analysis as the logarithmic scale has been shown to be closely related to human perception (Stevens et al., 1937; Sundberg & Rossing, 1990). In thesis, we will refer to the linear spectrogram of a signal as its spectrogram.

[1] Sampling frequencies of 44.1 kHz or 48kHz are used for musical signals whereas 8kHz or 16kHz are used for transmission of speech in telephony.

Various structures can be seen in frequency distribution of a voice signal. The voiced parts of a voiced signal show a *harmonic* nature with energy distribution over multiples of a frequency, known as harmonics. The lowest common denominator of the harmonics is often termed as the **fundamental frequency** (f0). This frequency is determined by the vibration of the vocal folds. The perceived pitch of a signal has been shown to be related to the logarithm of the fundamental frequency (Stevens et al., 1937; Sundberg & Rossing, 1990).

The frequencies amplified by the human resonant tract and the articulators can be observed as peaks within the frequency spectrum of the voice signal and are often term as **formants**. The relative position of the two lowest formants in a segment of a speech signal has been shown to be a determinant factor in the perception of the vowel sound for that segment. Non-voiced parts of human speech are generally less harmonic in nature and their perception is usually affected by the transition from a voiced to unvoiced segment to speech or vice-versa (Sundberg & Rossing, 1990). It can be seen from Figure 1.1 that the human voice is a combination of harmonic and inharmonic parts. As such, the voice is often modeled as a combination of pure tone sinusoids, a sample for which can easily be determined from knowledge of previous samples and noise, any sample of which is completely independent of past samples (Rafii et al., 2018).

A common model used for the human voice is called the source-filter model (Fant, 2001; Dudley, 1939). This model comprises of an excitation signal, representing the vibration of the vocal folds. Such an excitation signal might have a periodic nature, representing voiced parts of the human speech, or be more noise-like, representing unvoiced parts of speech. The excitation signal is passed through a series of band-pass filters (Dudley, 1939), mimicking the filtering process of the vocal tract. This gives a distinct spectral structure to the generated signal, with certain frequencies representing formants being amplified. Such a structure is often termed as the *spectral envelope* of the voice signal and is used in both analysis of voice recordings and the synthesis of

new voice signals. Models used for synthesis of speech in this manner are known as *vocoders* and are discussed in Section 2.4.1.

Analysis of the voice is also commonly done via **Mel-frequency cepstral coefficients** (MFCCs) (Davis & Mermelstein, 1980), which make use of a logarithmic scale of frequencies known as the mel-scale (Stevens et al., 1937). The mel-scale has been closely linked to human perception of audio, with consecutive units in the scale being perceived at an equal distance of separation by listeners. MFCCs are coefficients of the of the *mel-frequency cepstrum* (MFC) calculated by taking the discrete cosine transform (DCT) of the of the logarithm of the spectrogram of a signal on a mel-scale. These coefficients model the response of the human auditory system and have proven to be useful in several applications such as voice recognition (Chakraborty et al., 2014), sound classification and instrument recognition (Müller, 2007) among others. The mel-scale spectrogram is often termed as the *mel-spectrogram* and is also useful for speech analysis (Qian et al., 2019).

1.1.3 Components of voice signals

Humans can interpret speech signals to derive information pertaining to the speaker's identity, emotional state and physical location (Holt & Lotto, 2010). However, the primary function of speech is to transmit the intended information of the speaker (Moore et al., 2009), typically through the medium of language. Speech and language have been the focus of much research, through two distinct branches of study. The branch of auditory speech perception deals with perceptual mapping from acoustic signals to a sequential and categorical representation such as words, syllables or phonemes. Such a representation is referred to as the linguistic content of a speech signal. Psycholinguistics is the study of interpretation of meaning from the linguistic content of a signal.

Much research in the field of auditory speech perception has been dedicated to the study of *phonemes*, which are defined as *"the smallest linguistic unit that changes meaning*

within a particular language" (Holt & Lotto, 2010). The auditory mapping process from acoustic signal to a phonetic representation is quite complex as the acoustic signal is subject to high variability both inter-speaker and intra-speaker. Humans are capable of speech perception, under robust conditions including in the presence of noise and distortions (Diehl, 2008). This process of perception takes into account acoustic features in the speech signal such as changes in formant patterns, fundamental frequency, formant transition duration, voice onset time (VOT), among others to form a representation of linguistic features.

Phonemes include voiced vowels and voiced and unvoiced consonants, as discussed in Section 1.1.1. Each language uses distinct combinations of various voiced and unvoiced phonemes. If two speakers of similar linguistic background were to read the same line of text at a constant pace, then the phonetic sequence and the phonological content for both speakers would be the same (Raphael et al., 2007). The sequence of phonemes is parsed by the human cognitive mechanism to form meaningful language representations like words. Such a parsing process is language dependent and requires knowledge of the language being heard. While full details of linguistic perception are outside the scope of this thesis, we define the **linguistic content** as the speaker independent phonological content of a voice signal. In the context of this thesis, the linguistic content of a signal is denoted by $\mathbf{Z}$.

As seen in Figure 1.1, the signals for different speakers reading the same text is significantly different. This leads to the second element of speech that we consider, **prosody**. This is a combination of elements like *intonation*, *emphasis* and *rhythm* (Silverman et al., 1992), which are almost completely dependent on the speaker and may be influenced by variable factors such as emotions of the speaker at the time of speech. Intonation is defined is that variation in the **pitch** of a voice, which is a perceptual quality of sounds that humans can perceive as being *high* or *low*. The sensation of pitch is closely related to the **fundamental frequency** (f0) of the air pressure pulses created in the vocal folds, particularly for voiced sounds. In general, the f0 is defined as the lowest rate

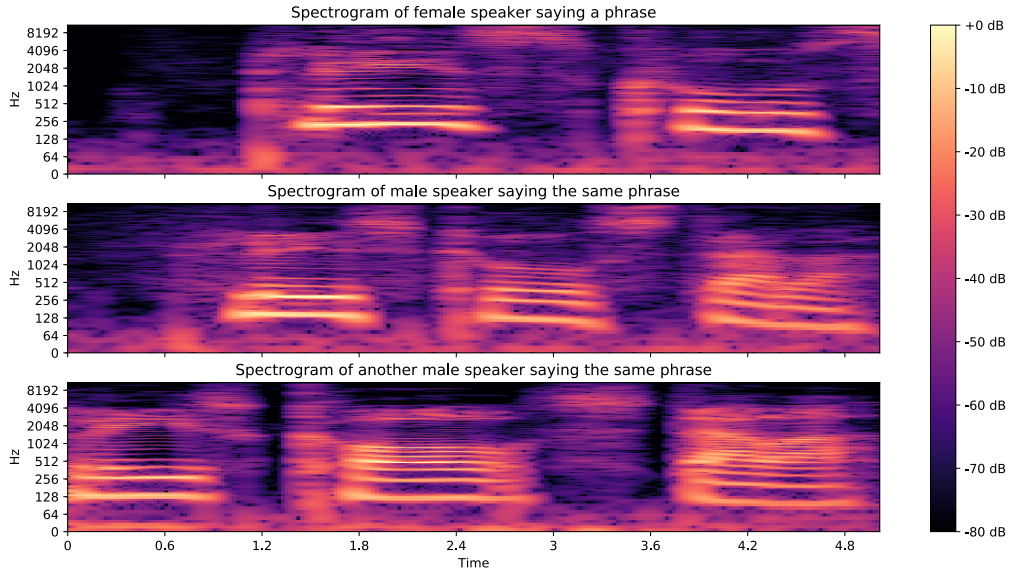

Figure 1.1: Log-scale spectrograms of the speech signals of three speakers, a female and 2 males saying the phrase "Please call Stella". The speech samples were taken from the VCTK corpus (Yamagishi et al., 2019).

of repetition of the cycles of air pressure in a sound signal and is generally expressed in **Hertz**, which is the number of cycles of a periodic signal per second. The perceived pitch of a signal has been shown to be correlated to the logarithm of the f0, defined in Hertz (Stevens et al., 1937). A signal with a higher f0 is generally perceived to have a higher pitch than a signal with lower f0. The f0 of a voice signal can be calculated through mathematical tools such as *auto-correlation*, discussed in Section 2.5.2.

The f0 has been shown to be linguistically insignificant for speech perception (Klatt & Klatt, 1990) although relative pitch changes do have an influence on acoustic perception (House & Fairbanks, 1953). The intrinsic f0 or the intrinsic pitch is also a physiologically driven factor effecting the f0 of a speech signal (Chen et al., 2021). An f0 curve, as the one shown in Figure 1.3 represents the evolution of the f0 across time and by doing so, also captures rhythmic information, which is the temporal structure of the sound. The emphasis component of prosody can be measured by the *energy* of the signal or by its **loudness**.

Like musical instruments, the voice signal has a certain distinguishable quality which

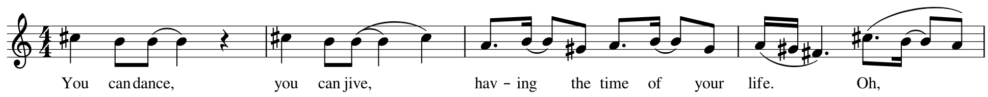

Figure 1.2: A score providing the vocal melody and the lyrics for a popular song.

allows one to distinguish between speakers, not taking into account the linguistic features and the quantifiable elements of prosody, like the f0. This distinctive feature is often termed as the voice quality or **timbre** of the voice, and is generally speaker specific. Formant frequencies higher than the first two formants typically contribute to this feature, although influence the perceived quality of voice (Sundberg & Rossing, 1990). Discounting emotional changes and changes in the pace of speaking, the timbre is a characteristic feature of the identity of the speaker. In this thesis, we will use the symbol ψ to denote the identity of a person, who can be a speaker or a singer.

1.1.4 Differences between speech and signing

Singing generally takes advantage of the harmonic nature of the voiced human sound, for musical effect. The fundamental frequencies of the singing voice signal show much higher variance than those for a speech signal. The range of fundamental frequencies for speech signals is around 110 Hz - 200 Hz for males and 200 Hz - 350 Hz for females. Whereas for singing, the fundamental frequency can go as high as 1400 Hz for soprano singing and 523 Hz for tenors (Sundberg & Rossing, 1990). As such, singing generally also has longer vowel lengths as compared to speech (Duan et al., 2013). While 60% of speech signals comprise of harmonic voiced sounds, the singing voice signal has been shown to be comprised of 90% voiced sounds (Sundberg & Rossing, 1990; Sundberg, 1987). This can be seen in Figure 1.4, where a phrase is both spoken and sung by the same person. It can be seen that the duration of the voiced phonemes is substantially longer in the sung version of the phrase than in the spoken version.

While the prosody and f0 of speech is speaker dependent, the pitch and emphasis of the singing voice is guided by **melody** provided in a musical score. As shown in Figure

1.2, a musical score generally provides information pertaining to linguistic information, pitch and timing. Linguistic information in a score is generally termed as the lyric and the pitch and timing information comprise the melody. As it is related to the perceived pitch, the f0 of a singing voice signal provides an estimation of the melody of the signal. In a real performance, the pitch or the f0 of singer can deviate from the melodic guideline provided by the score, both in terms of the pitch and timing. These deviations provide both an artistic outlet for the interpreter and are dependent on the singer. Artistic deviations to the pitch for example include *vibrato* (Seashore, 1932), which is a rhythmic and periodic fluctuation of the sung pitch from the pitch indicated in a score. Other identified deviations from the score include the *overshoot* (De Krom, 1995), which occurs just after a change in note, a *preparation*, which is a change in pitch opposite to the direction of a note change, right before the change and *fine fluctuation* (Akagi & Kitakaze, 2000), which includes irregular fluctuations higher than 10 Hz. These deviations are shown in Figure 1.3.

The f0 curve of a singing voice signal, as shown in Figure 1.3, captures the temporal evolution of the perceived pitch of an unaccompanied **a capella** singing voice signal. In doing so, it also represents the rhythmic information present in the signal and thus the main melody of the same. In this thesis, we will use the f0 curve to represent the melodic content of a singing voice signal, denoted by the symbol η.

In addition, the singing voice signal also has a slightly different spectral structure from the speech signal. In particular, the frequency of the first formant, which is crucial for vowel identification, has been shown to be varied by singers, according to the fundamental frequency being sung. This is generally done through a change in articulations by the singer. Despite this, vowel intelligibility has been shown to be high for the singing voice (Sundberg & Rossing, 1990). A spectral peak at 3 kHz, often termed as the *singing formant* (Sundberg & Rossing, 1990) has also been observed in recordings of the singing voice. Additionally, it has been shown that the formant amplitude of a singing voice signal modulates in synchronization with modulations in frequency (On-

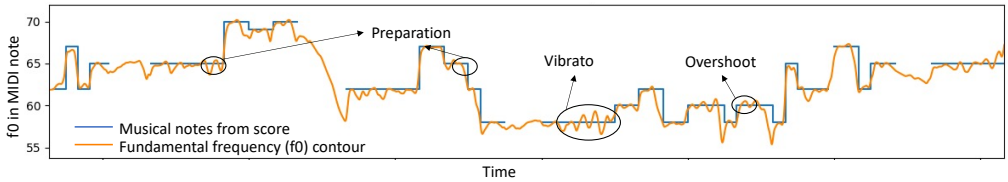

Figure 1.3: The identifiable deviations of a singing voice f0 contour from the melodic guidelines provided by the score. The f0 contour is shown in MIDI notes, using the formula $\eta = 12 \cdot log_2 \frac{\eta_{hertz}-69}{440}$, where η_{hertz} is the value of the f0 in Hertz.

cley, 1971), especially when the singer is performing a vibrato. Other contemporary artistic modulations can be added by the singer including variations in dynamics like tremolo or timbre via techniques like *growling* or *screaming*. These modulations allow the singer to communicate abstract information like emotion which cannot fully be encoded by the lyrics or melody and are discussed in Section 1.2.1. We note that despite these deviations in spectral characteristics, the perceived linguistic content of the singing voice is quite similar to that of normal speech. This can be observed in Figure 1.4, which shows the spectrogram of the phrase *"It's late in the evening, she's wondering what clothes to wear"* being spoken and sung by the same speaker. While there are clear differences in the vowel duration, spectral structures and the fundamental frequencies, the perceived linguistic content, $\mathbf{Z}$, from both audio examples is the same.

We thus define the singing voice signal to be composed of three key elements, the linguistic content, $\mathbf{Z}$, the melodic content, η, and the timbre. While the linguistic and melodic content are to a large degree independent of the singer, the timbre is a personal quality of the singer, ψ.

1.1.5 Separating the voice signal from other sources

In the early 20th century, Hermann von Helmholtz remarked upon what is called the *cocktail party problem* (Von Helmholtz, 1912; Cherry, 1966; Bregman, 1994). This problem constitutes of the interior of a ball room with multiple speakers speaking simultaneously, along with musical instruments and other noises commonly heard during

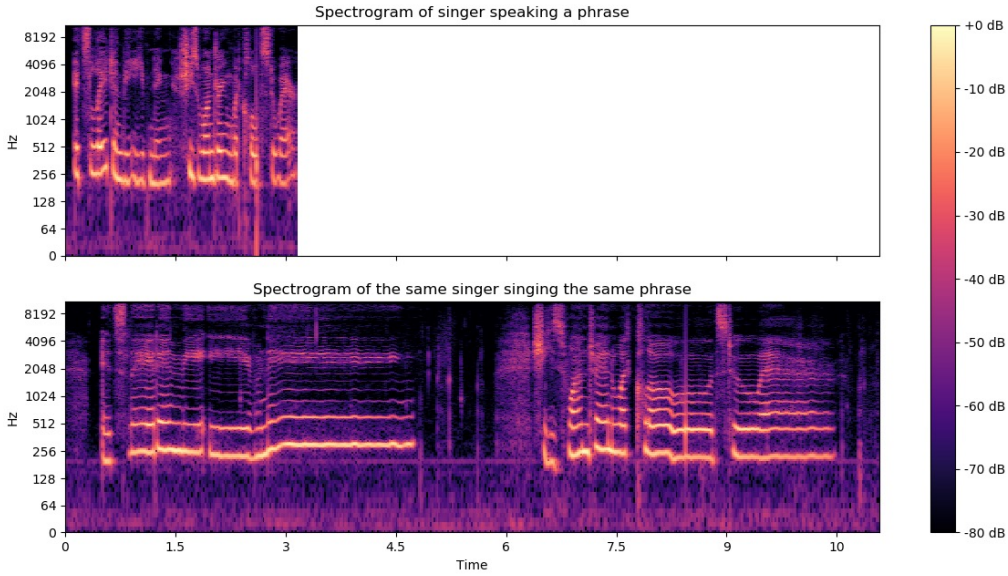

Figure 1.4: Log-scale spectrograms of Speech and singing voice experts of the same phrase, taken from (Duan et al., 2013)

a party of the era. Helmholtz states that the most important human facility is to distinguish various sounds in such an environment. Colin Cherry (Cherry, 1966) also commented on the human ability to comprehend speech even with the presence of other sound sources. The human brain's ability to distinguish individual sound sources within a complex acoustic environment has been researched over the decades. It has been postulated that the human auditory system first segments the auditory information received by the ear and then groups the segments into individual stream in a process termed as auditory scene analysis (ASA) (Bregman, 1994). The ASA process uses auditory cues like proximity in time and frequency, harmonicity, onsets and prior knowledge such as language.

Attempts to computationally replicate this process fall in the field of computational auditory scene analysis (CASA) (Weintraub, 1985; Bregman, 1994). In the later part of the 20th century, several models were proposed for grouping and segmenting auditory streams, particularly in single microphone **monaural** audio recordings. A fundamental principle used in CASA is that of auditory masking (Moore, 2012), which

suggests audio sources with higher energy content in a frequency band tend to mask other sources which might have lower energy content in that band. This has led to the proposition of an ideal binary mask or time frequency (TF) mask (Weintraub, 1985; Brown & Cooke, 1994; Wang, 2005). Such masks are calculated for each of the source to be separated from the mixture and are generally applied to the spectrogram of a mixture signal. The mask for a source has a value of either 0 or 1 for each frequency bin of the mixture at each time step, indicating whether the bin contributes to the source the mask corresponds to or not. Applied as such, binary masks were used by early CASA models estimate binary TF masks to separate speech signals from noise (Brown & Cooke, 1994) for **speech enhancement**.

Binary masks impose strict constraints on the separation process; a TF bin of a spectrogram either pertains to a source or it does not. However, such an approach adds artifacts to the output signal as it leads to abrupt changes in amplitude and phase. To alleviate this problem, soft masks like the generalized Wiener filter (Liutkus & Badeau, 2015a; Wiener et al., 1949; Wiener, 1950) have been proposed. Unlike a binary mask, a soft mask allows continuous values between 0 to 1 for each time frequency bin. The ideal ratio mask (Liutkus & Badeau, 2015a) (IRM) for a source is typically calculated as the ratio of the spectrogram of that source over the sum of the spectrograms of all sources in the mixture. Source separation algorithms typically estimate a soft mask emulating the IRM for separating the sources from an audio mixture. In this thesis, we use ω to represent a general **soft** mask. Each TF bin of a soft mask has a value in the range of 0 to 1 and the sum of all the masks for a mixture is 1 for each bin. Such a filtering approach assumes that the mixture is a linear sum of the individual sources to be separated.

The field of separating audio sources from a mixture is known as **audio source separation**. Along with the CASA inspired TF mask estimation approach detailed above, the field uses statistical approaches like independent component analysis (Hyvarinen, 1999; Hyvärinen & Oja, 2000) (ICA) and principal component analysis (Candès et al.,

2011; Huang et al., 2012; Sprechmann et al., 2012; Recht et al., 2010) (PCA). Innovations in the field of audio source separation have inspired research in the related field of **music source separation**. This field focuses on extracting the various musical sources in a musical recording. Due to its importance, extracting the singing voice from the musical mixture has received much interest in particular. Knowledge based algorithms have been applied to this task, exploiting the unique harmonic nature of the human voice using an analysis-synthesis approach (Miller, 1973; Maher, 1989; Wang, 1994, 1995). Other methodologies incorporating musical knowledge into the extraction process include the repeating pattern extraction technique (REPET) (Rafii & Pardo, 2011, 2012; Rafii et al., 2014) and its generalization, kernel additive modeling (KAM) (Liutkus et al., 2014b,a). Non-negative matrix factorization (Lee & Seung, 1999; Févotte & Idier, 2011; Vembu & Baumann, 2005; Virtanen, 2007; Févotte et al., 2009; Ozerov et al., 2012), which decomposes the musical signal into basis and activation functions has seen particular success for musical source separation. Closely related to the field of music source separation is the field of **speech source separation**, which involves separate the individual voices in a mixture of 2 or more speech recordings.

Over the last decade, deep learning based algorithms for music and speech source separation have led to significant improvements to the performance of source separation algorithms (Rafii et al., 2018). Such algorithms, discussed in Section 2.1, typically use deep neural networks to estimate TF masks given the magnitude component of the spectrogram of the mixture signal. The limitation to this data-driven filtering approach is that it can only extract what the models have been trained to extract, (Narayanaswamy et al., 2020) i.e, a deep learning model trained on separating vocals from an instrumental backing will extract the vocal stem signal, $\hat{\mathbf{y}}$, with any effects and spectral distortion from the singer added during the mixing stage. Also, there is dependency on the training data used for the training phase, in that a model trained to separate the vocal, bass and drum stems from a contemporary musical mixture cannot separate the soprano, alto, tenor and bass voices from an SATB recording. This leads to the mo-

tivation for the work presented in this thesis. We propose a system to synthesize the voice signal given a mixture spectrogram, allowing us to extract an estimation of the raw vocal signal, $\hat{\mathbf{x}}$, present in a polyphonic mixture. The proposed methodology can be applied to both contemporary popular music to remove the effects discussed in Section 1.1.3 and to synthesize a prototypical single voice signal representative of unison singing typically seen in choirs, as discussed in Section 1.2.2.

1.2 Motivation

Extracting the vocal signal from a mixture is the first step towards many applications such as analysis, improved listening through hearing aids (Pons et al., 2016; Edwards, 2007; Reindl et al., 2010), active music listening (Goto, 2007), generating new data for artistic purposes (Fitzgerald, 2011), karaoke and practice. For many of these applications, the processed stem signal, $\hat{\mathbf{y}}$, extracted by the filtering approach discussed in the previous section are acceptable. However, for several purposes, such as modulations, re-mixing with effects, melody enhancement, the processing effects retained in the extracted stem might be undesirable.

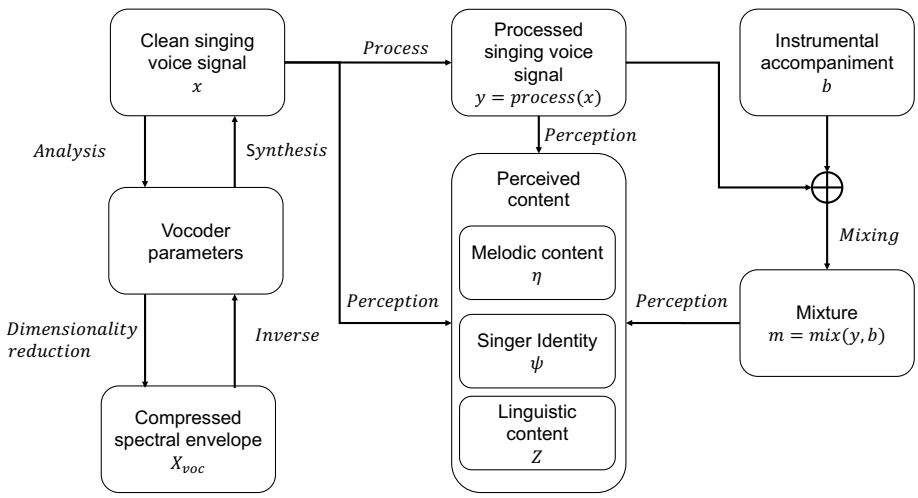

Figure 1.5: The analysis background for Part II of the thesis.

To extract a clean vocal signal free from modulations and effects from a polyphonic mixture, we propose a methodology to synthesize a signal $\hat{\mathbf{x}}$. The methodology aims to retain the linguistic and melodic content of the vocal signal present in a musical mixture. The approach is inspired by auditory scene analysis (ASA) (Bregman, 1994) and the analysis by synthesis theory of speech perception (Stevens, 1972, 1960).

We hypothesize that the perceived content consists of the linguistic content, the melodic content and the timbre and personal variations of the singer. As shown in Figure 1.5, we assume that this content, represented by $\mathbf{Z}, \eta and \psi$, respectively, remains the same for a singing voice signal even when the signal is processed and mixed with instrumental accompaniment. In this dissertation, we focus primarily on **monaural** signals which have a single channel in the audio mixture. We aim to computationally replicate human perception by deriving the linguistic and melodic content as well as the singer identity directly from a mixture signal and synthesize the clean vocal signal based on this content. While interesting research in itself, it can also have several practical applications such as:

- Synthesized versions of the vocal signal in contemporary musical mixtures could allow for more detailed analysis of the spectral distortions in the signal that vocal techniques such as *growling* produce as well as easier transcription and lyrics alignment.

- For enhanced hearing via audio aids, a vocal signal without effects might allow for easier melody and lyrical following, especially for people with hearing disorders.

- Active listening, which involves focusing attention on a particular element of music for enhanced appreciation, can benefit from mixing a clean version of the vocals with the song, especially for contemporary pop music.

- Teaching applications, such as the one we propose in this paper can be designed

allowing teachers or students to modify recorded practice corpora to the student's ability.

We apply this methodology to two main contexts in which the singing voice is present; contemporary popular music and ensemble singing.

1.2.1 Contemporary popular music

The singing voice is usually accompanied by music. Ancient forms of flutes and drums have been found by archaeologists, showing the capacity for humans to produce music of both a melodic and percussive nature. Through the ages, the instruments accompanying the voice have evolved with changes in technology. In contemporary music, we see various kinds of musical instruments including acoustic instruments like the violin, instrument utilizing electronic amplification like electric guitars, synthetic instruments like synthesizers and percussive instruments like drums. Musical composition and production using such instruments takes into the abstract characteristics of the sound, often called *timbre* along with the rhythmic, harmonic and melodic qualities for an arrangement. In a contemporary popular music recording, such signals from recordings or synthesis of such instruments are mixed along with the voice signal. This is a dedicated process which involves attention to detail in many aspects including frequency spectral balance between instruments, for which equalization is often used. While mixing, mastering and production are dedicated art forms in themselves, for the purpose of this study, we will call the entire process as *mixing*.

Signal processing has also opened avenues for using the singing voice as an instrument, whose sound can be modified. Some common effects used include *reverb*, *echo* and *delay*, but more artistic effects are also commonly used in modern music. From the 1970s, artists like *Led Zeppelin* and *The Beatles* utilized production techniques such as double tracking and rotating speakers to increase the *fatness* of the sound. In the 1980s, artists like *Bon Jovi* used effects like the *talk box* to merge the timbre of the voice with the sound of other instruments. More contemporary artists from the 21st century like

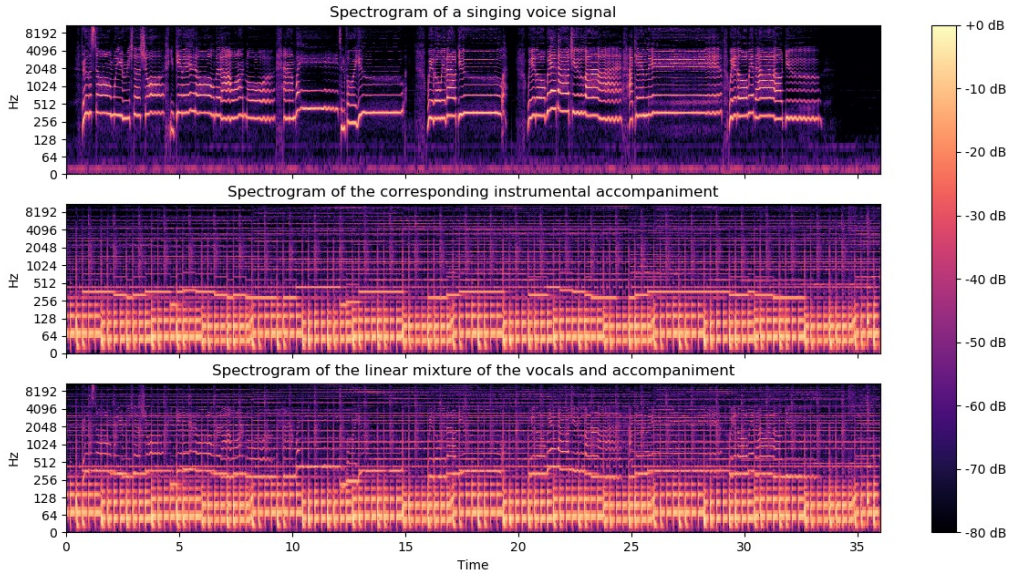

Figure 1.6: The mix for a contemporary popular music song. Log scale spectrgrams for the respective signals can be seen in the Figure, with a) showing the specrogram of the clean vocal signal, b) showing the spectogram of the backing track and c) showing the spectrogram of the mixture generated by a linear sum of the two tracks.

Muse and *Childish Gambino* use complex effects like *overdrive*, *ring modulators*, *pitch shifting* and *formant shifting* to shape the sound of their voice. The signal processed in such a way is often called as the *stem*. The stem signal is mixed with sounds from other instruments which may or may not have a harmonic structure. This process is commonly known as *mixing*. Along with effects, several non-traditional techniques are commonly applied by the singers themselves to alter the timbre or the spectral shape of the voice. In cases such as *growling*, *grunting* or *screaming* the harmonic structure of the voice, as described in the previous section is no longer retained. Artists such as *Rage Against The Machine* and *Opeth* have used such vocal techniques as a tool for expression within the song that cannot be covered by melody or lyrics alone.

From a spectral point of view, it can be observed from Figure 1.7 that the harmonic structure for processed vocals deviates from structure observed for the traditionally studied for the singing voice. Despite the spectral alterations associated with such techniques, effects and mixing, the intelligibility of the vocals is usually retained in a

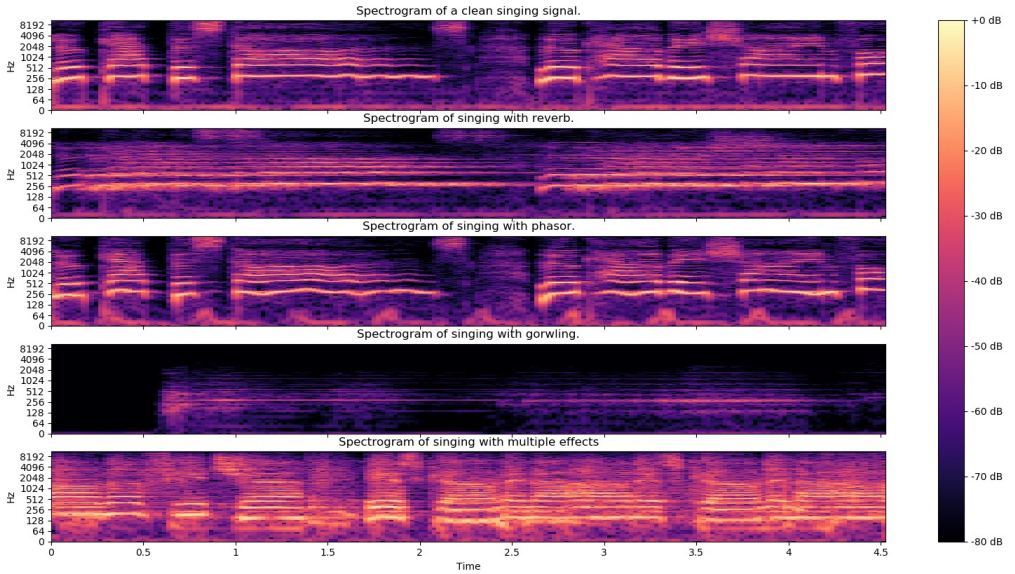

Figure 1.7: Log-scale spectrograms of vocals with various effects, including phasor, reverb, growling and multiple effects. The last two samples shown were taken from real world commercial recordings.

contemporary popular music mixture. For the rest of this thesis, we will use the symbol $\mathbf{y} = process(\mathbf{x})$ to denote the processed vocal stem signal, where $process()$ is a function covering the covering the effects discussed in this section. The backing instrumental track is represented by $\mathbf{b}$ and the mixture, $\mathbf{m}$, is considered to be a sum of the backing track and the processed vocal signal, $\mathbf{m} = mix(\mathbf{y}, \mathbf{b})$, where mix is the mixing process which may or may not be a linear sum.

The methodology we propose for synthesizing the voice signal assumes that the linguistic content of x is retained even though modulations are added to the signal. This linguistic content remains the same even for the mixture signal. i.e.: $\mathbf{Z_x} = \mathbf{Z_y} = \mathbf{Z_m} = \mathbf{Z}$. The melodic content can also be considered to be consistent, $\eta_\mathbf{x} = \eta_\mathbf{y} = \eta_\mathbf{m} = \eta$, although this is not necessarily always the case as many vocal effects like growling and formant shifting effect the perceived pitch of a signal even though the linguistic content is maintained.

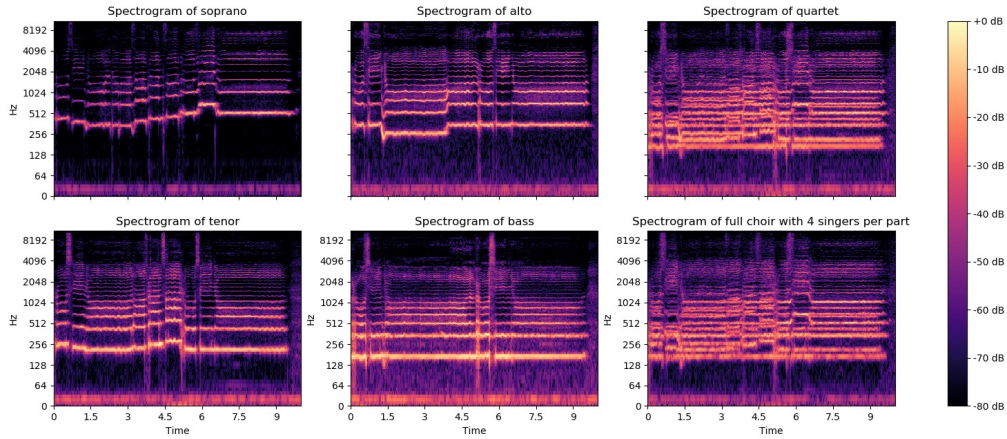

Figure 1.8: The log-scale spectrograms of the soprano, alto, tenor and bass parts of an SATB choir, along with the quartet mixture and the full choir mixture.

1.2.2 Ensemble singing and choirs

Singing has often been practiced as a social activity, with groups of multiple people getting together to sing in harmony. Such co-operative activity is termed as ensemble singing. Choral singing is the most popular format of ensemble singing, particularly in the western culture. As a social and at times religious activity, choral singing is studied and practiced in conservatories across the world. There are several formats for this type of singing, with different singing voices classified by the range of the singers. The most typical format of choral singing involves involves four distinct voices, with complimentary melodies and an associated vocal range for each voice.

As they sing simultaneously, the melodies for each of the voices are composed so as to be complementary. While the content sung by the distinct voices might differ, they are generally complementary; harmonically, rhythmically and lyrically. This means that the audio signal for the different voices has overlapping energy content in both the frequency and time domain, making source separation task for the case of choirs more challenging than for speech or music source separation. However, the difference in the vocal range between the voices makes for an important distinction that can be modelled by sufficiently power deep learning models.

Within each voice, there can be multiple singers singing the same melodic and lyrical content, simultaneously. Although choral singers practice to blend together as a single voice, there are some deviations in pitch and timing, which along with the ensemble of timbres leads to the perception of a *unison*. The singers in the unison are perceived to be singing a single pitch contour (Ternström, 1991) and the linguistic content is also consistent throughout. The similarity in timing, pitch, spectral characteristics and content makes the task of separating the individual voices in the unison all but impossible for source separation methodologies. However, we hypothesize that the intelligibility and the perceived pitch of the unison can be modelled via deep learning models to synthesize a single voice representative of the perceived pitch of the unison.

An SATB choir with just one singer per voice is often termed as a *quartet*. But an SATB choir could also have multiple singers in unison across each voice. AS seen in Figure 1.8, a recording of a full choir would have overlapping harmonics in the frequency dimension and blended temporal cues across time.

1.2.3 The TROMPA project

The Towards Richer Online Music Public-domain archives (TROMPA H2020 770376) or TROMPA is a multi-disciplinary project funded by the European Commission under the Horizon 2020 Research and Innovation program. The project aims to leverage large scale public musical data and state-of-the-art Music Information Retrieval (MIR) technology with the goal of democratizing publicly available European cultural heritage.

Various partners across academia and industry are involved with the project, with the goal of utilizing state-of-the-art technology across modalities to connect five distinct classes of users; music scholars, music enthusiasts, choir singers, content owners and instrument players

This thesis falls withing the choir use case, which aims to assist choir singers in their individual practice. We propose the framework for a methodology to allow choir singers to practice at home when they do not have a digitized score of the choral song but

rather a recording of a chorus. We propose a methodology incorporating blind source separation to separate the soprano, alto, tenor and bass voices from an SATB mixture. These voices are then analyzed to extract the linguistic and melodic content, which are used to synthesize alternate versions of the voices that can be modified and remixed according to the users' ability.

1.3 Research questions

We hypothesize that the vocal signal present within a musical mixture can be synthesized by extracting the underlying perceptual features from the mixture signal. Such a framework can also be applied to synthesizing a single voice signal from a unison singing signal commonly seen in ensemble singing. This thesis aims to answer the following research questions:

- Can a singing voice signal be synthesized from a musical mixture by using language independent representations of the perceived content of the signal?
    - Is it possible to extract synthesis parameters pertaining to the singing voice from a polyphonic contemporary music mixture?
    - How can the voice signal extracted using such a methodology be evaluated?
    - How can a feedforward neural network be used for singing voice synthesis given an input of linguistic content, singer identity and the f0 curve?
    - Can the linguistic content of a singing voice signal be represented in a language independent manner from which a voice signal can be synthesized?
    - Is it possible to extract such a representation of the linguistic content from a polyphonic contemporary music mixture?
    - How can we derive a representation of the singer identity for the voice synthesis process?
    - What are the potential applications of such a methodology?

- Can the individual voices in an ensemble choral singing recording be separated, given limited training data?
    - Can waveform based source separation algorithms work as well as spectrogram based models for choral part separation?
    - Are music source separation algorithms better suited to choral voice separation or should speech source separation algorithms be used?
    - How can we curate data from varied datasets which has been recorded under different conditions?
    - Can quartet based data with a single singer per part be used to train deep learning based algorithms for voice separate even with multiple singers per part in unison?
    - Is it possible to separate a single voice from within the unison singing signal?
    - What are the perceptual qualities of a unison signal that distinguish it from a signal voice singing signal?
    - How can choral source separation be useful?

1.4 Structure of the thesis

The rest of the thesis is structured as follows:

- A summary of musical source separation, voice synthesis and music information retrieval methodologies, particularly for the linguistic and melodic content and the singer identity is presented in Chapter 2.
- Chapter 3 lists the datasets relevant to our study as well as evaluation strategies used for source separation and voice synthesis algorithms.

- Part II of the thesis presents the framework for a methodology to synthesize a clean singing voice signal from a popular music mixture, using the underlying content. Chapter 4 provides an introduction to this part of the research.

- We present a deep learning based methodology to extract voice synthesis parameters from a contemporary popular musical mixture in Chapter 5.

- A multi-singer singing voice synthesizer based on a generative network in presented in Chapter 6. This synthesizer takes as input the linguistic and melodic content pertaining to the voice signal as well as the identity of the singer as a one-hot vector.

- Chapter 7 presents a methodology to extract the underlying language independent linguistic content and singer identity from mixture signal and synthesize the clean singing voice signal based on this content.

- Part III presents the application of source separation algorithms applied to choral singing. An introduction to this part of the research is provided in Chapter 8.

- We adapt and evaluate some the best performing methodologies for speech and musical source separation to separate the individual parts in an SATB choir. This study is presented in Chapter 9.

- We apply the methodology presented in Chapter 7 to synthesize a single voice signal from a unison mixture as well as to generate a unison signal from a single voice. This research is presented in Chapter 10,

- Applications for the methodologies presented in this thesis are discussed in Chapter 11. These include a choir practice tool using source separation, described in Section 11.2 and the application of analysis and synthesis to modern musical elements like percussive sounds and loops, presented in Section 11.3.

Audio examples associated with this thesis are presented at https://pc2752.github.io/thesis_examples/.

Chapter 2

Scientific background

This chapter provides a brief overview of the scientific background for the rest of this thesis. We start with a brief look at audio source separation through knowledge based algorithms in Section 2.1, following which we introduce some deep learning concepts in Section 2.2 and the application of such to audio source separation in Section 2.3.

We then take a look at voice synthesis algorithms, which are used for generating speech and singing voice signals from text and scores in Section 2.4. Singing voice synthesis algorithms are particularly pertinent to our task and we observe that they typically derive linguistic and melodic information from the input score and use this information along with a singer identity representation to synthesize a voice signal.

Section 2.5 discusses music information retrieval (MIR) techniques relevant to our subject, particularly looking at the extraction of linguistic content in Section 2.5.1, melodic content in Section 2.5.2 and singer identity representation in Section 2.5.2.

2.1 Knowledge based source separation

As discussed in Section 1.1.5, source separation is the task of separating the individual sources in a mixture of the same. Musical source separation and speech source separation are two fields of audio source separation, which focus on musical mixtures and

asynchronous speech mixtures, respectively. Statistical algorithms such as independent component analysis (Hyvärinen & Oja, 2000) (ICA) are typically used for source separation in several audio fields. ICA assumes statistical independence amongst the source signals in the mixture. This assumption applies to speech separation but not necessarily to music source separation. Musical mixtures typically consist of harmonic instruments including the singing voice, guitars, piano and bass as well as percussive inharmonic instruments like drums, which are synchronized in time, leading to frequency and temporal correlations between the individual signals.

While the unique nature of music hinders the application of source separation methodologies like the ICA, it opens up avenues for the use of innovative algorithms specifically tailored to the musical domain. The repeating pattern extraction technique (REPET) (Rafii & Pardo, 2012) is one such algorithm which leverages the fact that the instrumental accompaniment in a musical song has a short time repetitive nature, while the singing voice is generally more robust over this short period of time. Algorithms using REPET typically use MIR methodologies to identify repetitions like the beat spectrum (Foote & Uchihashi, 2001) within the mixture. Such repetitions are then used to estimate the accompaniment by averaging, typically over spectrograms. Extensions to the basic REPET technique have been proposed to handle non-periodic structure by using a self-similarity matrix (Rafii et al., 2014), in a methodology known as REPET-SIM. A generalization of the REPET methodology is the use of kernel additive modelling (Liutkus et al., 2014b,a) (KAM), which uses source specific kernels to model a source at various points in the spectrogram, allowing the identification of multiple repeating patterns within the accompaniment.

Comb-filtering and synthesis based approaches specifically for extracting the singing voice from a musical mixture exploit the harmonic nature of the singing voice. Synthesis based methods typically identify the fundamental frequency of the singing voice within the and use source-filter models to synthesize the singing voice signal. Such models are discussed further in Chapter 4. Comb-filtering involves a similar procedure,

but uses the voice model to create a filter for separating the voice and accompaniment signals.

One of the most successful methodologies applied to musical source separation is non-negative matrix factorization (NMF) (Lee & Seung, 1999; Févotte & Idier, 2011; Vembu & Baumann, 2005; Virtanen, 2007; Févotte et al., 2009; Ozerov et al., 2012). NMFs exploit low-rank assumptions to separate components of the mixture using non-negative constraints. The algorithm used for source separation via NMF involves decomposing the input mixture spectrogram, $\mathbf{M}$, into two non-negative matrices, known as the basis or spatial matrix, N, and the temporal or activation matrix, Q, as $\mathbf{M} = NQ$. Given an input spectrogram matrix with Υ frequency bins across T time frames, basis matrices for each instrument to be separated are calculated, with a shape of $\Psi x \Upsilon$, where Ψ represents the number of bases computed. As shown in Figure 2.1, the time wise activation for each of these bases is represented by the activation matrix, which has a shape of $T x \Psi$. For musical source separation, the basis function is often assumed to represent the pitch of the instrument to be separated while the activation matrix represents the onset and offset time (Carabias-Orti et al., 2013). Such factorization allows for the estimation of Weiner filters, which can be applied to the mixture spectrogram to separate the individual sources.

NMFs have been applied to the task of musical source separation, particularly for separating the singing voice and instrumental accompaniment (Vembu & Baumann, 2005). Probabilistic latent component analysis (Smaragdis et al., 2007) (PLCA), an equivalent of NMF has also been proposed for musical source separation. NMFs have also been supplemented with pitch and timing information provided by a musical score (Joder & Schuller, 2012; Zhao et al., 2014) along with source-filter models of the voice (Durrieu et al., 2011, 2009; Janer & Marxer, 2013). The Flexible Audio Source Separation Toolbox (FASST) (Ozerov et al., 2012; Salaün et al., 2014) provides a version of the NMF algorithm for source separation using an algorithm for generalized expectation-maximization (GEM) from incomplete data (Dempster et al., 1977). Application of

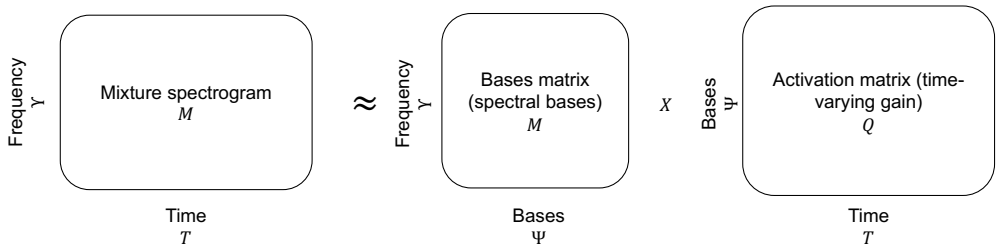

Figure 2.1: Source separation via Non-Negative Matrix Factorization (NMF) involves decomposing the mixture spectrogram into bases and activation matrices.

NMFs to musical source separation makes a low-rank assumption for both the vocals and accompaniment. However, models proposing a sparsity constraint on the vocals have been proposed for source separation (Huang et al., 2012; Sprechmann et al., 2012; Jeong & Lee, 2014; Yang, 2013) using robust principal component analysis (Candès et al., 2011) (RPCA). Additionally, NMFs have been applied forsource separation of other instruments in a mixture (Miron et al., 2016).

Over the last decade, data-driven deep learning (DL) methodologies have been applied to the source separation problem and have shown to outperform other methodologies. Deep learning is a term generally given to Artificial Neural Networks with sufficient depth. Such networks are able to internally model patterns in data given a sufficiently large amount of data to learn from as well as a training objective. Knowledge intervention is minimal in this case, but the models are generally domain specific, i.e. they are only able to separate the sources that they have been trained to separate and need adaptations for different sources.

2.2 A brief introduction to deep learning

Deep Learning is name given to data-driven machine learning algorithms that are based on the biological neural networks used by the human brain. In general, the interconnected neuron graph structure that is observed in the brain is computationally replicated in an artificial neural network in a series of nodes representing neurons. The simplest

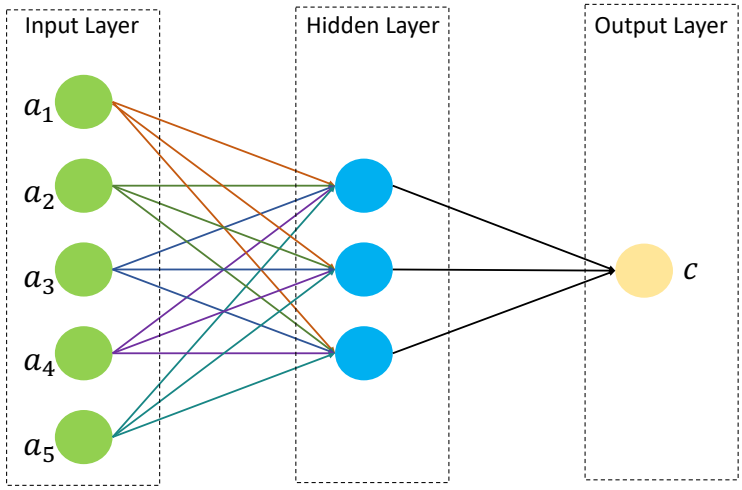

Figure 2.2: Basic Neural Network: connections are shown between neurons for the input layer, the hidden layer and the output layer. Note that each node in the input layer is connected to each node in the hidden layer and each node in the hidden layer is connected to each node in the output layer.

artificial neural network is consists of three layers, an input layer, a hidden layer and an output layer.

The input layer contains as many nodes has the dimensions of the input features. As can be seen in Figure 2.2, each node of of the hidden layer is connected to each layer of the input layer and calculates a weighted sum of the input nodes, in a topology often referred to as the **Restricted Boltzmann Machine** (RBM). The weights of these layers are often known as the parameters of the network and can be modified and learned using an optimization algorithm. As such, such a network topology is called a *fully-connected* network. The number of nodes in the hidden layer is often termed as the width of the network at the layer. A non-linearity such as the **Rectified Linear Unit** (ReLU) can be applied to the weighted sum. This non-linearity is often termed as an activation function or a transfer function. Other non-linearities that are often used include the **tanh** activation function, the **sigmoid** activation function and variations of the ReLU function like the **Leaky-ReLU** and the **parameterized-ReLU**. A similar

connection exists between each node of the hidden layer and the nodes of the output layer, along with an output.

Such a connected graph structure allows the network to express complex non-linear functions of the input features. The universal approximation theorem shows that such neural networks can approximate continuous functions. I has been shown that any Lebesgue integrable function can be approximated by a width bound neural network with ReLU activations of sufficient depth (Lu & Lu, 2020). As such, a network with sufficient depth is commonly known as a Deep neural Network, leading to the Deep Learning nomenclature.

Generally, neural networks are used to learn mappings between an input data distribution and a target distribution. To do so, the weights of the neural network are modified using optimization algorithms to minimize the error between the output of the network and the target data. Such an error is expressed as **loss function**, which can express the difference between the output of the network and the target. The main aim of these loss functions is to model the output distribution given the input distribution, or maximizing the probability of the output distribution, $\hat{\mathbf{c}}$, matching the target distribution, $\mathbf{c}$, given the input distribution, $\mathbf{a}$ by optimizing the model parameters, θ. The most intuitive way to do this is to represent the target distribution as a normal distribution with mean μ and standard deviation σ, $a \sim \mathcal{N}(\mu, \sigma)$ and maximize the likelihood, $\prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\mu-\mathbf{c})^2}{\sigma^2}}$. Minimizing the log-likelihood is one way to do so as it allows the loss function to be expressed as a sum, as shown in Equation 2.1. Another approach is to assume unit variance, $\sigma^2 = 1$, which gives us the **Mean Squared Error** (MSE), shown in Equation 2.2.

$$\mathcal{L}_{nll} = \sum \left(log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \frac{(\mu - \mathbf{c})^2}{\sigma^2} \right) \tag{2.1}$$

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{c}_i - \hat{\mathbf{c}}_i\|^2 \tag{2.2}$$

Where N is the total number of samples in the distribution of $\mathbf{a}$ and $\mathbf{a}_i$ and $\mathbf{c}_i$ refer to the ith samples of the respective distributions. Another commonly used loss function for continuous data used in the case of **regression** is the **Mean Absolute Error** (MAE), shown in Equation 2.3.

$$\mathcal{L}_{MAE} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{c}_i - \hat{\mathbf{c}}_i\| \tag{2.3}$$

Binary Cross Entropy is often used when the target can be expressed as a binary (yes/no) value. Such data is often used in binary classification and the output layer in this case usually has a sigmoid activation. Multi-class classification is also done using neural networks, using a **softmax** activation in the final layer. Categorical Cross Entropy is the loss function used in this case. These can also be shown to be forms of maximizing the log-likelihood of the output distribution given the input distribution.

Backpropogation of gradient of the loss function through the layers is used to optimize the parameters of the networks over numerous iterations (Rumelhart et al., 1985). This optimization phase is generally called the **training** phase and is generally done in batches of input and target pairs using an algorithm known as **stochastic gradient descent** (SGD). The basic optimization algorithm can be augmented using momentum (Qian, 1999), which takes into account prior optimization steps while updating the network parameters for the current step. Adagrad (Duchi et al., 2011), Adadelta (Zeiler, 2012) and Adam (Kingma & Ba, 2014) algorithms have been proposed for parameter optimization. Normalization techniques such as **batch normalization**, **instance normalization** and **layer normalization** are often applied during the training phase. Normalization involves shifting and scaling the internal representations of the neural network using the mean and standard deviation of the the features calculated across various dimensions. Such normalization technique have been introduced to mitigate training problems such as the *internal covariate shift* (Ioffe & Szegedy, 2015) in the data and also lead to a smoother objective function used for optimization of the network (Santurkar et al., 2018). Normalization also helps with over-fitting and vanishing and

exploding gradients (Salimans & Kingma, 2016).

For evaluation and application, the input can be fed through the network in a **feedforward** manner. This process where the inputs are fed through the network to generate outputs is known as **inference**. More hidden layers can be added to the basic structure shown in Figure 2.2, adding *depth* to the network. Such feedforward networks are often termed as multi-layer perceptrons (MLPs)

2.2.1 Unsupervised learning

Deep learning can be **supervised** or **unsupervised**, both terms pertaining to the training stage. In the supervised case the desired target for a given input is explicitly used for optimizing network parameters, whereas in the unsupervised case, the network has to implicitly learn desirable features from the data presented. One of the most successfully applied unsupervised methodologies is the autoencoder network (Rumelhart et al., 1985; Baldi, 2012). The input and target for an autoencoder network are the same. Constraints such as width limitation, shown in in Figure 2.3, are imposed on the hidden layer allowing it to learn implicit structures within the data that can be used for a representation of the data from which it can be reconstructed. The design of the autoencoder architecture is important for the learned representation, often termed as the latent representation, to be meaningful for the desired task of the autoencoder.

The architecture for most deep neural networks used for processing audio is based on the concept of convolutional neural networks (CNNs) and recurrent neural networks (Medsker & Jain, 2001) (RNNs).

2.2.2 Recurrent neural networks

Recurrent neural networks (Rumelhart et al., 1985; Jordan, 1997) (RNNs) are neural networks that take a time series as input and output a time series. Considering a simple three layer architecture, each node of the the hidden layer maps a function of the input nodes at the current time step and the hidden node of the previous time step, as shown

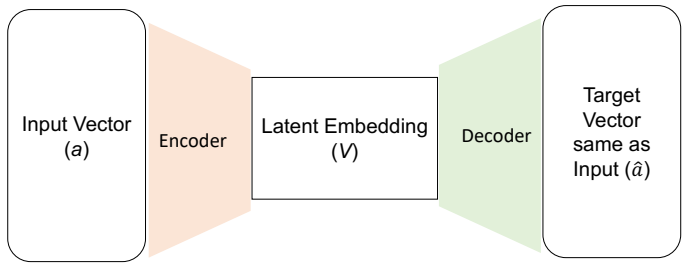

Figure 2.3: The framework of an autoencoder, with an encoder and a decoder. The input and the target vector are the same data and the latent embedding has some restrictions imposed to allow it to learn meaningful structures from the data

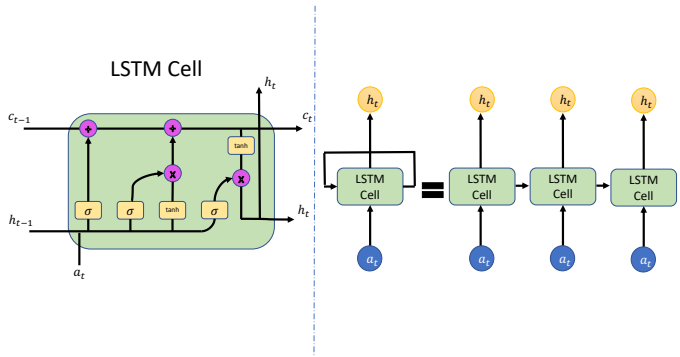

Figure 2.4: The cells used for Recurrent Neural Networks (RNNs) with Long Short Term Memory Networks (LSTMs) and the RNN unrolled.

in Figure 2.4. Since an equal weight is given to each of the time steps, this network has infinite memory, but is vulnerable to exploding or vanishing gradients due to propagation of information over a long time series. One of the solutions to such problems is to use learnable gated weights for the previous steps. **Long short term memory networks** (Hochreiter & Schmidhuber, 1997) (LSTMs) and **gated recurrent unit** (GRUs) are often used to this effect. The architecture for LSTMs includes a cell state, shown in Figure 2.4, the weight parameters of which decide the degree of influence a time step in a series has on other time steps. Bidirectional LSTMs or BLSTMs have also been proposed. Such networks process the time series in both the forward and backward dir-

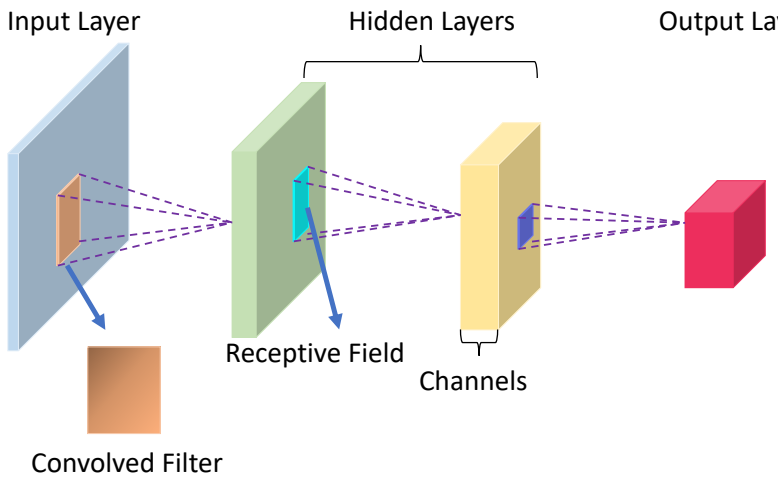

Figure 2.5: Convolutional Neural Networks: Local Receptive Field

ection and have been used for various audio processing applications including source separation.

2.2.3 Convolutional neural networks

Convolutional neural networks (Fukushima, 1988) (CNNs) are inspired by the human visual cortex, convolving a bank of two-dimensional filters on the two-dimensional input, commonly seen in images. The coefficients of the filter kernels are optimized in the same manner as the weights of an artificial neural network. The distinction between a fully connected network and an convolutional neural network is that the each neuron of the hidden layer is only connected to a localized set of the nodes of the input layer, known as the *local receptive field*. Multiple filters of the same shape and size are convolved on each layer, the output feature map formed by each such filter is commonly known as a *channel*.

CNNs have been adapted to the audio domain as well, either by using two-dimensional filters over time-frequency representations like the sepctrogram of an audio or by using one-dimensional filter convolutions directly over the waveform. Temporal convolu-

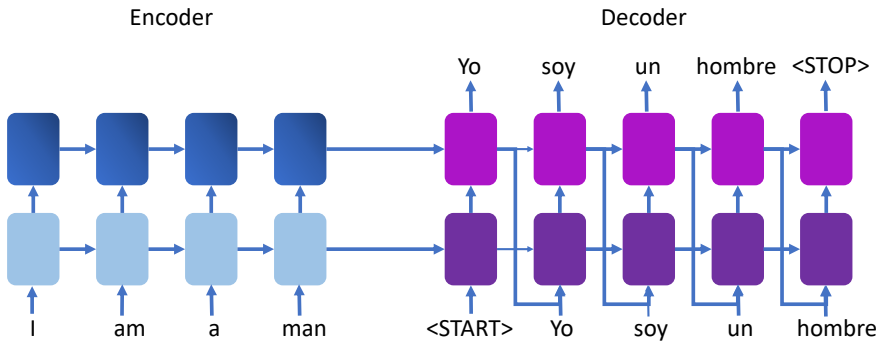

Figure 2.6: Sequence to Sequence modelling

tional network (TCNs) (Lea et al., 2016) use causal convolutions to model time series. Such networks have been used for source separation and speech enhancement. An autoregressive neural network is one such example of a CNN, which uses a series of causal convolutions over time to model the time series in a manner similar to the RNNs. The receptive field of an autoregressive CNN is increased using dilated convolutions and with skip and residual connections.

The limitation of using such models for temporal modelling is that the length of the output sequence along time is dependent on the temporal dimension of the the input sequence. However, this condition does not always hold for all temporal sequences that can be mapped. For example, for text to speech (TTS) synthesis, the input sequence of linguistic features like phonemes has a smaller length than the output sequence pertaining to acoustic features [2]. For such sequences, a technique called sequence-to-sequence (seq2seq) modelling has been proposed. Initially used for machine translation (Bahdanau et al., 2015), the basic structure of the model is shown in Figure 2.6 and includes an RNN based encoder which generates a summary vector that is fed to an RNN based decoder. The decoder generates the output sequences until a $< STOP >$ character is generated. This technique can be extended further using **attention** (Vaswani et al., 2017). As seen in Figure 2.7, using attention involves calculating a weight matrix for

[2] TTS synthesis is discussed in section 2.4

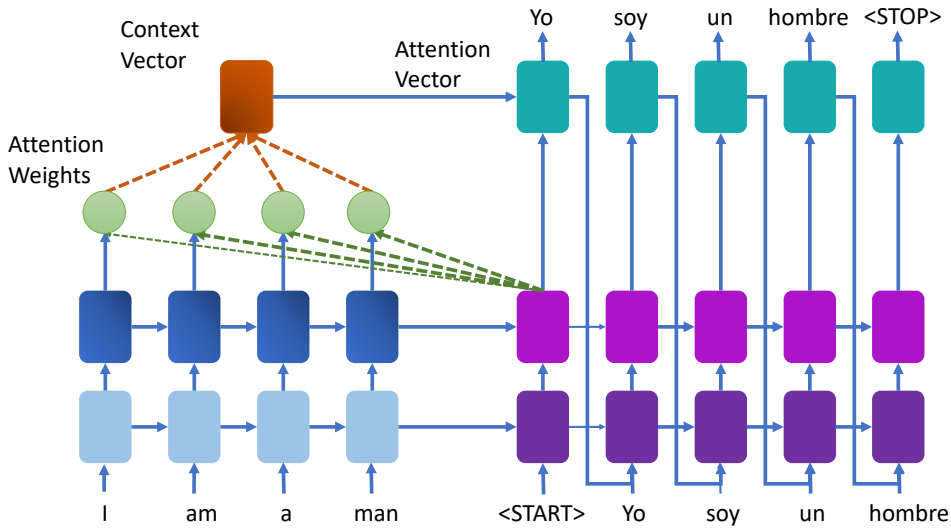

Figure 2.7: Sequence to Sequence modelling with attention

the input sequence which assigns a different weight to influence each time step of the input sequence has on a step of the output sequence.

2.2.4 Generative networks

Generative modeling is a form of probabilistic modeling which can define the underlying probability distribution of data. In general, generative models are used for generating new data from the probability distribution, but also help discover underlying correlations and structures of interest within the data. within the data. Statistical models like *Gaussian Mixture Models* and *Hidden Markov Models* have been used in the past for generative modeling, in particular for speech synthesis.

With deep neural networks, new possibilities have opened up for generative modelling. One of the most common methodologies used for generative modelling is the **variational autoencoder** (Kingma & Welling, 2014) (VAEs). While various architectures have been used within the VAE framework, the common structure consists of an encoder and a decoder, both of which use neural networks as parameterized function estimators. The encoder part of the network models a posterior distribution of a random

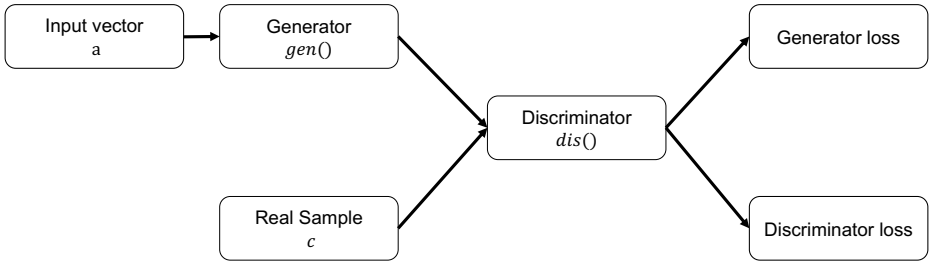

Figure 2.8: The Generator and Discriminator networks used in Generative Adversarial Networks (GANs).

latent variable, $\mathbf{V}$, over input data, $\mathbf{a}$, as $p(\mathbf{V}|\mathbf{a})$. The latent variable is typically of lower dimensions than the input data and is assumed to be standard normal with a diagonal covariance, $\mathbf{V} \sim \mathcal{N}(0,1)$. The decoder learns a distribution of the input data over the latent variable. The latent variable can then be sampled to generate new data from the distribution. Normalizing flows offer an extension of this generative methodology.

The **generative adversarial network** (Goodfellow et al., 2014) (GANs) is another generative model that is based on an adversarial training scheme. The methodology was initially proposed for generating image samples and consisted of a generator network, *gen* and a discriminator network, *dis*, which are trained simultaneously. The generator takes a noise vectors as input and generates image samples while the discriminator is trained to distinguish between the images generated by the generator and a set of real image samples, as shown in Figure 2.8. The authors showed that the networks can be optimized through backpropogation to an equilibrium between the networks wherein the Generator is able to model the probability distribution of the real image data such that the output of the discriminator is $1/2$ for all samples.

The two-player non-cooperative training that tries to minimize the divergence between a parameterized generated distribution ρ_g and a real data distribution, ρ_r, as shown in Equation 2.4.

$$\begin{aligned}\mathcal{L}_{GAN} = \min_{gen}\max_{dis} &\mathbb{E}_{\mathbf{c}\sim\rho_r}[\log(dis(\mathbf{c}))] \\ &+\mathbb{E}_{\mathbf{a}\sim\rho_\mathbf{a}}[\log(1-dis(gen(\mathbf{a})))]\end{aligned} \tag{2.4}$$

Where **c** is a sample from the real distribution and **a** is the input to the generator, which may be noise or conditioning as in the Conditional GAN (Mirza & Osindero, 2014) and is taken from a distribution of such inputs, $\rho_{\mathbf{a}}$.

While GANs have been shown to produce realistic images, there are difficulties in training including vanishing gradient, mode collapse and instability. To mitigate these difficulties, several adaptations of the GAN methodology have been proposed like the Laplacian pyramid GAN (Wang et al., 2019) (LAPGAN), the deep convolutional generative adversarial network (DCGAN) (Radford et al., 2016), generative recurrent adversarial network (GRAN) (Wang et al., 2020a) and the Wasserstein GAN (WGAN) (Arjovsky et al., 2017).

Another form of generative modelling that has been particularly useful for audio applications is the autoregressive framework, which has long been used to model. time series in fields like economics. In the adaptation for audio, either RNNs or causal CNNs are used to model one sample of the audio time series as a function of the previous samples. Conditional autoregressive models are often used in speech synthesis, particularly for inverting time-frequency representations like the mel-spectrogram to the corresponding waveform. The WaveNet (van den Oord et al., 2016a), described in Section 2.4.4 is an example of an autoregressive convolutional network used for speech synthesis. Generative modelling has been applied to the audio and music domain (Dieleman et al., 2018; Zukowski & Carr, 2018; Engel et al., 2020b; Défossez et al., 2018).

2.3 Data-driven source separation with deep learning

While signal processing based algorithms like those described in Section 2.1 have performed well, data-driven deep learning based models have led to significant improvements in the field of audio source separation. Several models using recurrent neural networks, LSTMs or convolutional neural networks (CNNs) to model time-frequency correlations in the signal have been proposed for the related tasks of speech enhance-

ment, musical source separation and speech source separation. While several models have been proposed, we look briefly at some of the models that are used within the scope of this thesis.

The commonly used deep learning pipeline involves the use of TF masks that are applied to the input mixture signal, with the assumption that the mixture, $\mathbf{m}$ of sources is a linear sum of the sources $\mathbf{s}_i$; $\mathbf{m} = \sum_{i=1}^{K} \mathbf{s}_i$, where i is the index of the K sources. Deep neural networks are then trained to estimate soft TF masks, ω_i, that are generally applied to magnitude component of the spectrogram, $\mathbf{M}$, for each of the sources to be separated. This results in estimates, $\|\hat{\mathbf{S}}_i\|$, as shown in Equation 2.5. As the mixture is a linear sum, the sum of the masks for the individual sources is 1. The separated sources are synthesized with the phase component of the mixture spectrogram, using the Inverse Short Time Fourier Transformation (ISTFT).

$$
\begin{aligned}
|\hat{\mathbf{S}}_i| = \omega_i \bigodot |\mathbf{M}|, \text{for i = \{1,2...K\}} \\
\sum_{i=1}^{K} \omega_i = 1 \\
\hat{\boldsymbol{s}}_i = ISTFT(|\hat{\mathbf{S}}_i|, \angle \mathbf{M})
\end{aligned}
\tag{2.5}
$$

Where $ISTFT()$ represents the Inverse Short Time Fourier Transformation. For most source separation algorithms pertaining to contemporary polyphonic music, the mixture, $\mathbf{m}$, of processed vocals, $\mathbf{y}$ and a backing track, $\mathbf{b}$ is approximated as a linear mixture, $\mathbf{m} \approx \mathbf{y} + \mathbf{b}$. Deep neural network models, represented by $nn()$, are generally used to estimate a TF mask for the vocal stem, ω_{vocal}, given the mixture signal as an input. The mask is applied to the mixture signal to generate the separated output. The resulting separated vocal track is an approximation of the magnitude component of the processed vocal stem signal, $|\mathbf{Y}|$, as shown in Equation 2.6. The phase component of the mixture spectrogram is typically used to synthesize the waveform of the estimated sources, using the Inverse Short Time Fourier Transformation (ISTFT), as shown in Figure 2.9.

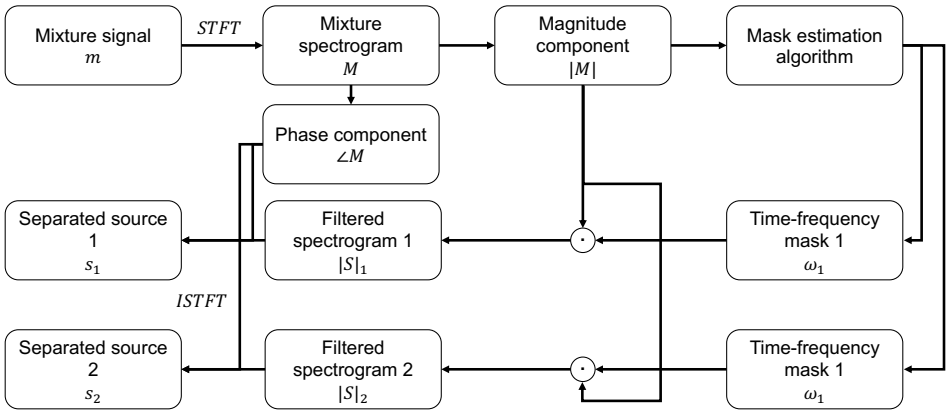

Figure 2.9: The pipeline for source separation using TF masks.

$$
\begin{aligned}
\omega_{vocal} &= nn(|\mathbf{M}|) \\
|\hat{\mathbf{Y}}| &= \omega_{vocal} \bigodot |\mathbf{M}| \\
|\hat{\mathbf{B}}| &= (1-\omega_{vocal}) \bigodot |\mathbf{M}| \\
\hat{\mathbf{y}} &= ISTFT(|\hat{\mathbf{Y}}|, \angle \mathbf{M}) \\
\hat{\mathbf{b}} &= ISTFT(|\tilde{\mathbf{B}}|, \angle \mathbf{M})
\end{aligned}
\tag{2.6}
$$

Where $ISTFT()$ represents the Inverse Short Time Fourier Transformation. For training, the ground-truth sources to be estimated are required. On inference, the Time-Frequency masks are computed via the neural network, given the mixture as an input. The general sequence of deep learning based source separation algorithms is shown in Figure 2.9. Deep neural networks using the phase information (Roux et al., 2018) as well as directly operating on the waveform (Stoller et al., 2018) have also been proposed. But even such algorithms work on a form of filtering in the waveform domain assuming that the mixture is a linear sum of the sources.

A deep learning based for estimating Ideal Binary Masks (IBMs) to separate speech signals from a noisy mixture was proposed by (Wang et al., 2014). While one of the

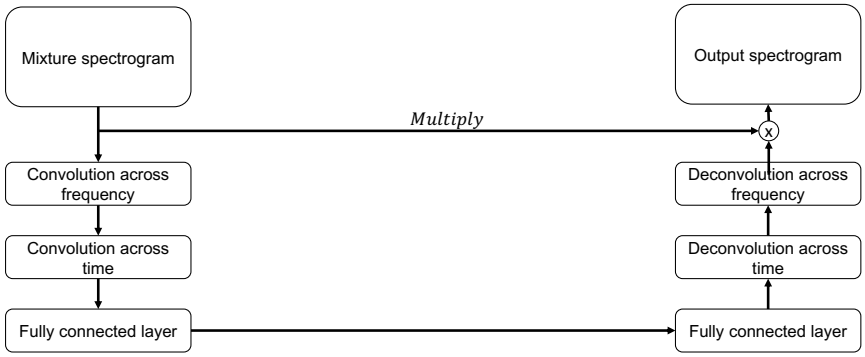

Figure 2.10: The DeepConvSep architecture for source separation (Chandna, 2016), utilizing a convolutional encoder and a decoder to generate soft TF masks for source separation.

first applications of Deep Learning to musical source separation used a simple architecture comprising of 3 layers (Huang et al., 2014). The first layer is a fully connected layer, used as a feature extractor over the frequency dimension of the input mixture sepctrogram. This feature map is passed through a Recurrent Neural Network to model the temporal dependencies with past samples and finally another fully connected layer is used to estimate TF masks with the same dimensions as the input.

One of the first models using Convolutional Neural Networks, commonly known as *DeepConvSep*, was proposed by us (Chandna, 2016). It used an autoencoder inspired bottleneck. Convolutions across frequency and time are used to learn a compressed representation of the input spectrogram, which is then upsampled by the corresponding convolutions to generate TF masks for the sources to be separated. As seen in Figure 2.10, the architecture is composed of two stages, namely the encoder, which encodes the input spectrogram into a lower dimension or latent representation and the decode, which generates the masks from the latent representation. The intuition behind this model is that the lower dimension representation can be used by the network to distinguish between the individual sources. The output of the network, γ_n, is used to estimate a soft mask, $\omega_i^{DeepConvSep}$ as shown in Equation 2.7:

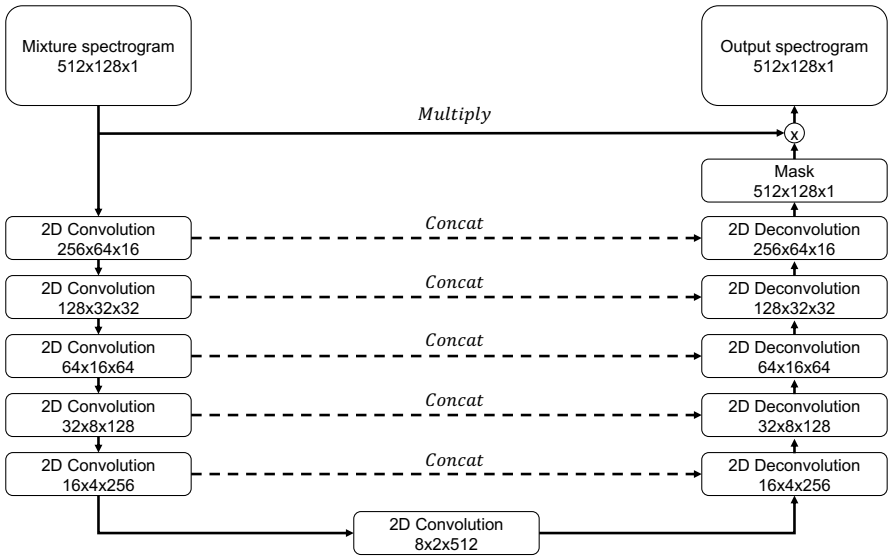

Figure 2.11: The U-Net architecture for source separation (Jansson et al., 2017)

$$\omega_i^{DeepConvSep} = \frac{|\gamma_i|}{\sum_{i=1}^{K} |\gamma_n|} \tag{2.7}$$

where γ_i represents the output of the network for the i^{th} source and K is the total number of sources to be estimated.

The estimated mask is then applied to the input mixture signal to estimate the sources $\hat{\mathbf{s}}_n$.

$$|\hat{\mathbf{S}}_i^{DeepConvSep}| = \omega_i^{DeepConvSep}.|\mathbf{M}| \tag{2.8}$$

The model was trained to minimize the MSE between the estimated magnitude of the spectrograms of the respective sources and the corresponding ground truths.

$$\mathcal{L}_{DeepConvSep} = \mathbb{E}||\hat{\mathbf{S}}_i^{DeepConvSep}| - |\mathbf{S}_i||^2 \tag{2.9}$$

A similar methodology was used by the the U-Net architecture The architecture, initially proposed for medical image segmentation (Ronneberger et al., 2015), includes skip connections between the corresponding layers of the encoder and decoder stages of the convolutional neural network, thus allowing for propagation of information and

gradients between the encoder and decoder layers. When applied to source separation (Jansson et al., 2017), the model takes as input the mixture spectrogram and produces the corresponding masks to be applied for extracting the sources. The intermediate layers of the model consist 2D convolutional layers, as seen in Figure 2.11. The convolutional filter in the first layer is applied across both the frequency and time dimensions of the input spectrogram, with successive layers building on top of the representation learned by this layer. The model is trained to minimize the MAE between the estimated sources, $|\hat{\mathbf{A}}_i^{UNet}|$, and the corresponding ground truths, as shown in Equation 2.10.

$$\mathcal{L}_{UNet} = \mathbb{E}[||\hat{\mathbf{S}}_i^{Unet}| - |\mathbf{S}_i||] \tag{2.10}$$

The U-Net model was originally proposed for separating the singing voice from a musical mixture and used two separate networks to predict the masks for the vocal and instrumental accompaniment stems (Jansson et al., 2017). *Deezer* has also applied this model for a product known as **spleeter** (Hennequin et al., 2020), that is commonly used for source separation amongst artists. The model follows the U-Net architecture (Jansson et al., 2017) with the specifications published by (Prétet et al., 2019) and has been shown to perform at very high speeds and efficiency. Conditional (Meseguer-Brocal & Peeters, 2019) versions of the U-Net have also been proposed, using a **Feature-wise Linear Modulation** (FiLM) (Perez et al., 2018a) layers.

A limitation of such methodologies using the spectrogram was that they only modelled the magnitude part of the complex spectrograms, disregarding phase information. A few models have been proposed for including phase information in the source separation process (Rennie et al., 2005; Liutkus et al., 2018; Roux et al., 2018; Williamson et al., 2015). One methodology for modelling the complex spectrogram is to use complex ratio mask, wherein the TF masks applied to the mixture spectgrogram can take complex values (Williamson et al., 2015; Roux et al., 2018). Other models estimate the magnitude spectrogram and the phase difference between consecutive frames of the sources to generate the individual signals (Afouras et al., 2018). The PhaseNet model

(Takahashi et al., 2018a) estimates a discrete representation of the phase of the output spectrogram.

The Wave-U-Net (Stoller et al., 2018), as seen in Figure 2.12 is one such model, applying the U-Net architecture to the waveform of a musical mixture signal. The model introduces learned upsampling using linear interpolation in the decoder allowing the feature maps to have meaningful representations. To this end, an interpolated feature $f_{t+0.5}$ is computed between neighbouring features, f_t and f_{t+1} for each time step in each feature map of the decoder layers, using a parameter, w constrained by a sigmoid non-linearity as shown in Equation 2.11. The last layer of the Wave-U-Net model uses a *tanh* non-linearity and enforces an energy-conserving criteria using a difference output layer. This is done by estimating $K-1$ source signals and estimating the last signal as $m - \sum_{i=1}^{K-1} \hat{a}_i$. In doing so, the model maintains the linear sum assumption and replicates the TF masking process that is used in the models discussed previously, while estimating the waveform of the signals to be separated.

$$f_{t+0.5} = \sigma(w) \bigodot f_t + (1 - \sigma(w)) \bigodot f_{t+1} \tag{2.11}$$

The learned interpolation is implemented as a 1D convolution with constraints and allows convex combinations of weight, leading to a generalization of simple linear interpolation wherein $w = 1$.

While several such models for source separation have been proposed, the data-driven nature of such methodologies requires that the datasets used for training and evaluation be standardized, along with the methodologies used for evaluation. As such, a community based effort for standardization of the source separation paradigm has taken shape. This led to asks within Music Information Retrieval Evaluation eXchange (Downie, 2008) (MIREX) campaign and Signal Separation Evaluation Campaign (Vincent et al., 2009, 2012; Araki et al., 2012; Liutkus et al., 2017; Stöter et al., 2018) (SISEC), which is conducted every few years. The musical source separation task of the

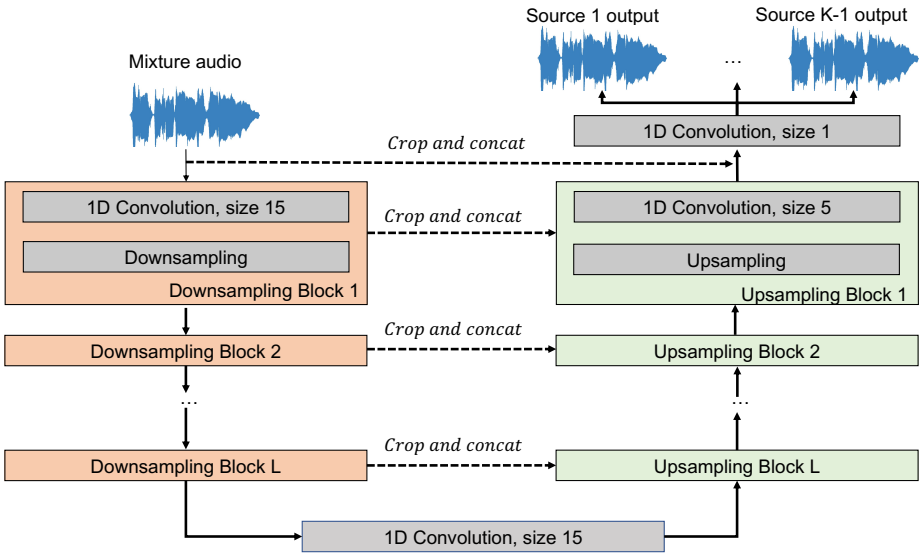

Figure 2.12: The Wave-U-Net architecture (Stoller et al., 2018)

SiSEC campaign consists of two sub tasks; separating the *vocals* and *accompaniment* stems from a stereo mixture of contemporary popular music and separating the *vocals*, *drums*, *bass* and *others* stems from a mixture of the same. The *MUSDB18* dataset, discussed in Section 3.1, is used in the more recent iterations of the campaign and the performance of the algorithms is evaluated using the *bss_eval_sources* set of metrics (Vincent et al., 2006), discussed in Section 3.3.2. Source separation algorithms submitted in the campaign are compared against **oracle** source separation methods which use the ground truth stems to estimate TF masks like the **Ideal Binary Mask** (Wang, 2005) (IBM), the **Ideal Ratio Mask** (Liutkus & Badeau, 2015b) (IRM) or the α-Wiener filter and the Multichannel Wiener Filter (MWF) (Duong et al., 2010).

Such campaigns have fuelled research in the domain of music source separation, and have seen deep learning based models for source separation outperform classical signal processing based models over the last few years. Many of these algorithms use combinations of fully-connected layers, LSTMs and CNNs. Over the last few iterations, a network blending a feedforward network with a BLSTM to predict MWFs for the input mixture spectrogram has consistently been amongst the best performing algorithms

evaluated in this campaign (Uhlich et al., 2017). However, the algorithm is also trained on additional proprietary data. The same applies to other algorithms that have performed well in the campaign, including the use of a densely connected convolutional network (DenseNet) (Takahashi & Mitsufuji, 2017) and the MMDenseNet (Takahashi et al., 2018b).

Along with the use of external data for training another problem faced by researchers building on the SiSEC campaign is that many of the best performing algorithms are not accompanied by open source implementations. As such researchers in the field face issues while using a benchmark for their own algorithms. This had a negative impact on research as researchers have had to compare their own proposals with open source implementation of models which were not quite the state-of-the-art (Stöter et al., 2019). Open-Unmix (Stöter et al., 2019) is an open source model proposed for providing the state-of-the-art benchmark in the musical source separation domain in 2019. As shown in Figure 2.13, the model uses a combination of fully connected time-distributed layers, along with skip connections, LSTMs and batch normalization to estimate multichannel Wiener filters (MWFs) (Nugraha et al., 2016) for each of the sources from the input mixture spectrogram . The MWFs are calculated by combining the output of all estimated sources to filter the input mixture spectrogram. This model outperformed the best models proposed in the 2018 SiSEC campaign, establishing the state-of-the-art open source source separation system for musical signals.

The Conv-TasNet (Luo & Mesgarani, 2019) was proposed for speech source separation in the waveform domain and was shown to outperform the ideal TF mask. Adapting the idea from the TasNet (Luo & Mesgarani, 2018), the model applies 1-D convolutions, called encoder basis functions, to overlapping chunks of the input waveform. This generates an intermediate representation similar in form to the spectrogram estimated via the convolutions of the STFT. A ReLU non-linearity is used to ensure non-negativity in the basis functions. A temporal convolutional network (Lea et al., 2016) (TCN) is used to estimate masks which are applied to the intermediate representation, similar to the

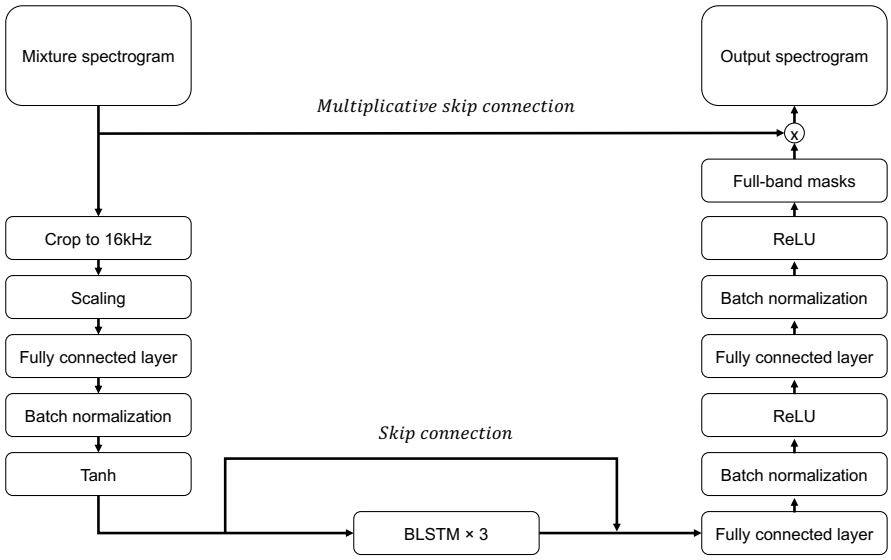

Figure 2.13: The Open Unmix architecture (Stöter et al., 2019)

TF masks estimated for spectrogram based source separation. The resulting representations are inverted to waveforms for the sources to be separated using a 1-D convolution with decoder basis functions, as shown in Figure 2.14. The weights of the encoder and decoder basis functions are learned during training. The networks are optimized by using the scale-invariant source-to-noise ratio (SI-SNR) loss function instead of the MAE or MSE loss functions that are typically used for source separation algorithms. The Conv-TasNet model has been adapted to music source separation (Samuel et al., 2020), through the use of a meta-nueral network.

$$\begin{aligned} \mathbf{s}_i^{target} &= \frac{\langle \hat{\mathbf{s}}_i, \mathbf{s}_i \rangle \mathbf{s}_i}{\|\mathbf{s}_i\|^2} \\ e_i^{noise} &= \hat{\mathbf{s}}_i - \mathbf{s}_i^{target} \\ \mathcal{L}_{sisnr} &= 10 log_{10} \frac{\|\mathbf{s}_i^{target}\|^2}{\|e_i^{noise}\|^2} \end{aligned} \tag{2.12}$$

Demucs (Défossez et al., 2019) is another deep learning based algorithm for source separation that uses a variant of the Wave-U-Net (Stoller et al., 2018) with BLSTM layers the middle of the U-Net architecture. The Demucs model has an encoder comprising of

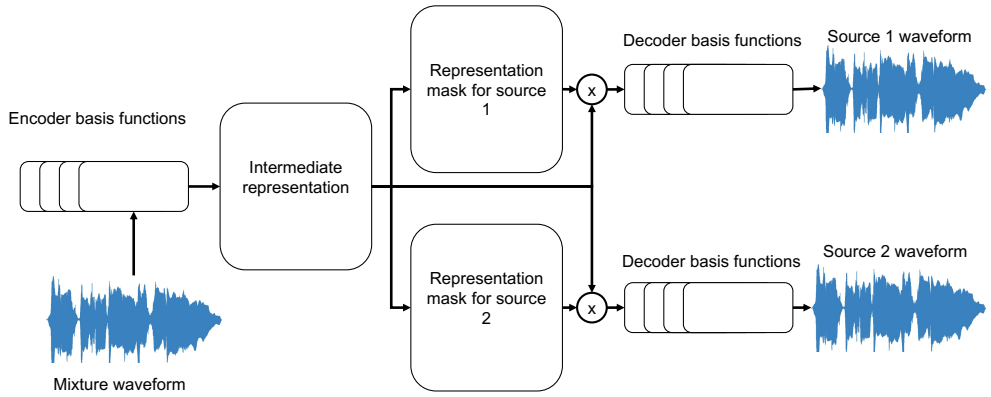

Figure 2.14: The basis functions framework used for the TasNet (Luo & Mesgarani, 2018) and Conv-TasNet (Luo & Mesgarani, 2019) models.

6 1D convolutional layers, each with a gated linear unit (GLU) (Dauphin et al., 2017) activation function. The encoder is followed by 2 BLSTM layers, the dimensions of the output of which are reduced using a fully connected layer. This output is then passed through a decoder network using transposed convolutions. The decoder mirrors the encoder network and there are skip connections between the corresponding layers of the two networks. Ablation studies with the model showed that optimizing the network with either the MAE or the MSE was nearly equivalent in terms of performance. The significance of the GLU non-linearity was also highlighted in this study.

Generative models have been used for audio source separation as well. In particular for unsupervised audio source separation using generative priors instead of predefined ones as used by most of the algorithms listed above (Narayanaswamy et al., 2020). The SVSGAN (Fan et al., 2018) model has been used to separate the singing voice from instrumental accompaniment using a GAN based training methodology. However, the system itself is not generative and uses the adversarial loss to compliment the MSE while estimating TF masks for the individual sources to be separated.

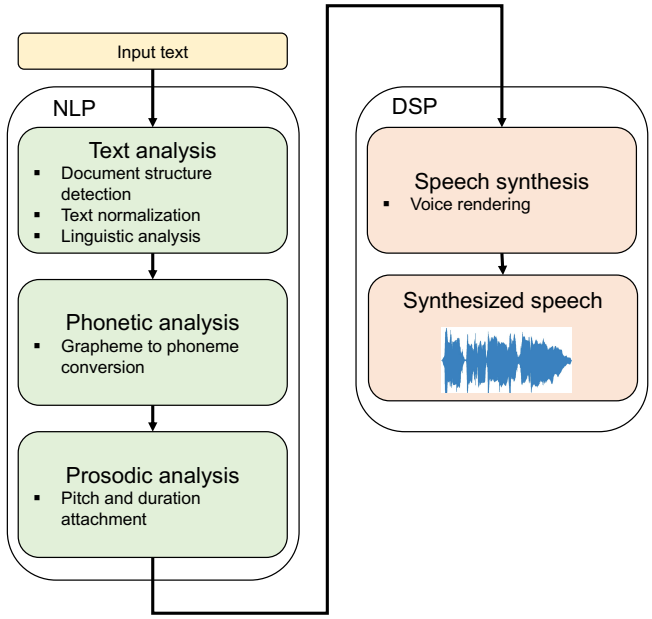

Figure 2.15: The Text-to-Speech (TTS) synthesis pipeline

2.4 Voice synthesis

Along with speech, written text is the most common means of communication for humans and both can be seen as representations of the same information. Due to the importance of the two forms of media, inter-conversion via machines between them has been a field of interest for researchers for decades. Text-to-Speech (TTS) synthesis is the field of research which aims to generate a voice signal from textual data, stored in machine readable form. TTS uses a combination of *Natural Language Processing* (NLP) and *Digital Signal Processing* (DSP). NLP is used for processing the linguistic information in text and consists of three phases, as shown in Figure 2.15

1. *Text analysis*: Linguistic information in the text can be deciphered in the form of graphemes, which are the smallest contrasting units in the writing system. Syntactic and semantic analysis are used to normalize the text and to model the context underlying the input text and generate a sequence of graphemes.

2. *Phonetic Analysis*: The pronunciation for grapheme units units is context dependent, both long-term, influenced by semantics and syntax and short-term, with the various combinations of the grapheme units. The phonetic analysis stage of the TTS pipeline formalizes the representation of the pronunciation of the normalized text with grapheme-to-phoeneme conversion. The output of this stage is a sequence of phonemes.

3. *Prosodic Analysis*: As discussed in Section 1.1.3, the prosodic component of the speech signal includes the fundamental frequency, the dynamics and the duration of the individual phonemes in sequence. The prosodic analysis of the TTS pipeline models these elements from the sequence of phonemes generated by the previous stage. The output of the prosodic analysis is generally a sequence of phonemes with duration assigned and a F0 curve.

For the context of this thesis, the DSP part of the TTS is the most pertinent and is discussed in greater detail. In this phase, synthesis algorithms like concatenate or parametric synthesis are used to generate the speech signal from the features derived by the prosodic analysis. Concatenate synthesis using Unit Selection based Synthesis (USS) methods have long been used for speech synthesis. USS consists of a pre-recorded set of speech recordings corresponding to various phonological units like phonemes, diphones, syllables, words, phrases and even sentences. For synthesis, such. units are concatenated together , often with modifications, to generate the desired speech content. Forced alignment of linguistic units to the corresponding speech segments is often done through Hidden Markov models (HMMs). Such models are used to select the sequence of units closest in correspondence to the sequence of linguistic and prosodic information provided as input (Gonzalvo et al., Proc. Interspeech 2016). This process is often termed ass unit selection.

Parametric synthesis (Black et al., 2007; Zen et al., 2007) has been successfully used for TTS in the past. Such synthesis uses generative models like HMMs to generate a set

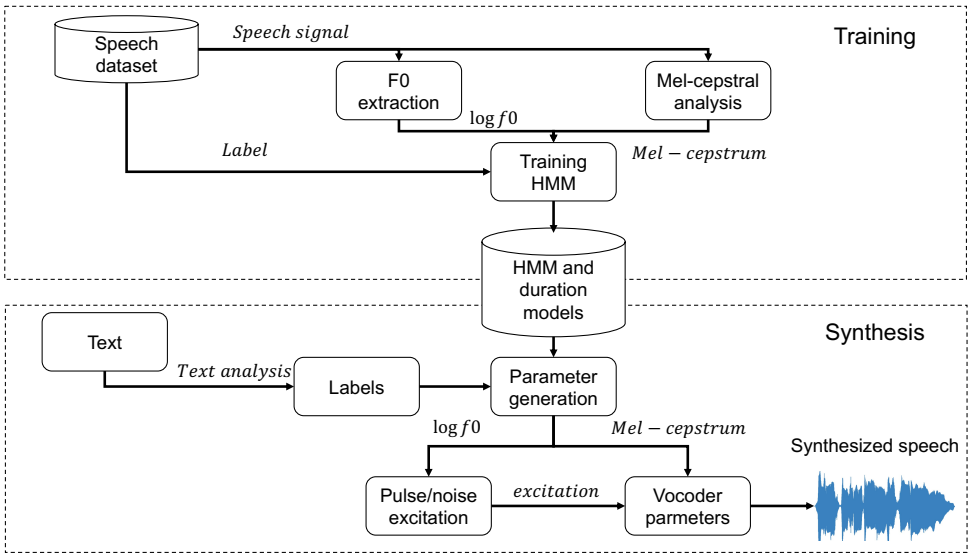

Figure 2.16: The framework for training and synthesis from Hidden Markov models (HMMs)

of acoustic parameters given linguistic and prosodic information. As shown in Figure 2.16, for training an HMM-based TTS system, typically mel-cepstral parameters and the log f0 is extracted from the ground truth voice signal and used to fit the parameters of a combination of context-dependent HMMs using the maximum log likelihood criteria. Such HMMs are used to model various combinations of sequences in the input linguistic information. Clusters of Gaussian mixture models (GMMs), assigned using top-down decision-tree-based context clustering are used to reduce the total number of combinations of input contexts. For synthesis, linguistic and melodic information from the musical score to be synthesized is converted to a context-dependent sequence of linguistic content labels. This sequence is used to generate a state sequence by concatenating the corresponding context-dependent HMMS, the duration of each of which are determined using the previously trained models. The sequence of HMMs is then used to generate synthesis parameters often termed as *acoustic parameters* or *vocoder parameters*. Vocoders are used for compact representation of the speech signal and are inspired by the source-filter model of human voice production (Dudley, 1939, 1938; Dudley et al., 1939; Dudley, 1958; McAulay & Quatieri, 1986).

2.4.1 Vocoder parameters for voice synthesis

Acoustic features or parameters used for synthesizing the human voice are often known as vocoders. Such parameters are often used for speech transmission (Dudley, 1940; Atal, 2018), analysis and synthesis (Dudley, 1939). The idea of the vocoder is based on the human vocal system, described in section 1.1.1, and the source-filter model of the human voice. The first vocoder was developed by Homer Dudley (Dudley, 1939, 1938; Dudley et al., 1939; Dudley, 1958) in the 1930s, using a series of band-pass filters to model the spectral envelope of the speech signal. In doing so, the vocoder system was able to isolate the f0 and the spectral envelope components of the signal in a manner that signal could be re-synthesized from information about the components. It was also shown that the spectral envelope was correlated to the perceived linguistic content of the speech signal the f0 to the emotional content (Dudley, 1939). The basic system has some key drawbacks; the spectral envelope estimated in this case was not completely robust to changes in the f0 and the unvoiced segments of the voice signal were not fully represented. However, the system proposed by Dudley led to conception of the idea of channel vocoders for speech transmission and synthesis.

Claude Shannon (Shannon, 1949), advocated the use of white Gaussian noise to model the aperiodic elements of speech and the use of all-pole filters for modelling the spectral envelope. This lead to the code-excited linear prediction (CELP) and the idea of linear predictive coding (Makhoul, 1975) (LPC). Within an LPC model, a speech signal is represented by the convolution of a source or excitation signal and a series of all-pole infinite impulse response (IIR) filters which models the resonance characteristics of the human vocal tract as the spectral envelope. The excitation signal is modeled via an impulse train or white noise with certain characteristics. The filter is used to shape the spectral characteristics of the signal. Such a system allows for paramterized speech encoding and synthesis as the coefficients of the all-pole filters are used as parameters for synthesis. Linear regression is typically used to estimate these parameters for a voice signal by minimizing the least squares error between the signal estimated by the

model. Since the output of the IIR filters depends on the current and previous values of the excitation signal, the system is often termed as auto-regressive.

Modern vocoder systems build on LPC idea, particularly for improving the quality of unvoiced sounds using dynamic excitation modelling (Chung & Schafer, 1990) and mixed excitation (McCree & Barnwell, 1995). Such systems include the cepstral vocoder (Vích & Vondra), the homomorphic vocoder (Chung & Schafer, 1989; Weinstein & Oppenheim, 1971) the STRAIGHT vocoder (Banno et al., 2007), the TANDEM-STRAIGHT vococer (Kawahara et al., 2008) and the WORLD vocoder (Morise et al., 2016).

The WORLD vocoder has been shown to be particularly effective for singing voice synthesis (Blaauw & Bonada, 2016, 2017) and is used throughout this thesis. The analysis phase of this vocoding algorithm involves estimation the fundamental frequency of the speech signal to be analyzed. The original framework for the vocoder proposes the use of an algorithm known as DIO. (Morise et al., 2009) for f0 estimation. The f0 is then used to estimate the spectral envelope using a pitch adaptive analysis (Mathews et al., 1961) based model known as CheapTrick (Morise, 2015a). The model estimates the power spectrum of a windowed waveform using pitch information to negate the effects of spectral leaks at the boundaries of the windowed frame. This power spectrum is referred to as the harmonic component of the vocoder parameters and is used to filter the excitation signal.

Using this information, the next step of the analysis is to model the aperiodic elements of the speech signal. This is done using an algorithm known as Definitive Decomposition Derived Dirt-Cheap (D4C) (Morise, 2016). This algorithm estimates the band aperiodicity for a voice signal using a sinusoidal model based on pitch synchronous analysis (Mathews et al., 1961) and a parameter based on the temporally static group delay (Kawahara et al., 2012). This model provides the periodic power component for each of the frequency bands analyzed. A ratio between the total power and the periodic power components is used to calculate the band aperiodicity.

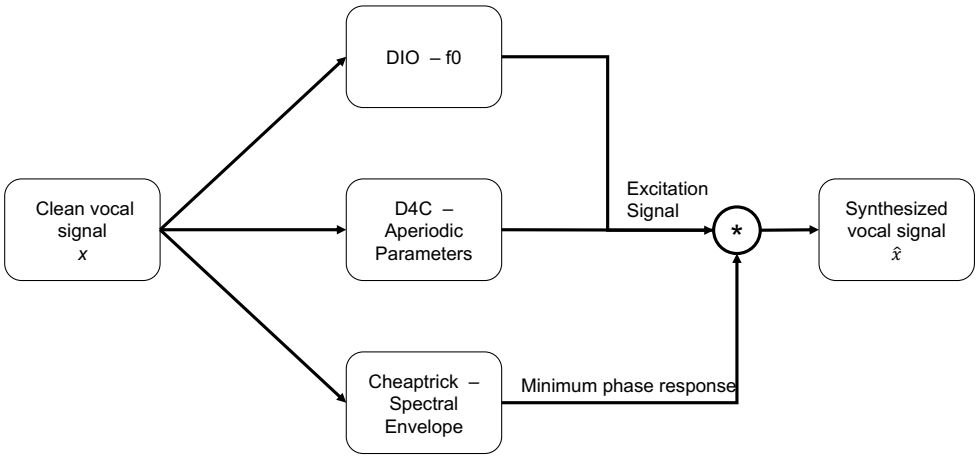

Figure 2.17: The frameworkrk for the WORLD vocoder (Morise et al., 2016)

For synthesis, the band aperiodicity coefficients are used along with the f0 to generate the excitation signal for the vocoder model. A convolution between the excitation signal and the harmonic spectral envelope is used to generate the speech signal. As such, the vocoder allows a parametereized representation of the voice signal, with three components; the f0, the harmonic component and the aperiodic component. This process is shown in Figure 2.17.

The WORLD vocoder has been effectively used for speech and singing voice synthesis. For singing voice synthesis with the WORLD vocoder, it is common to compress the harmonic and aperiodic components (Blaauw & Bonada, 2016) of the vocoder parameters. The dimensionality of the harmonic component is reduced using truncated frequency warping in the cepstral domain (Tokuda et al., 1994) with an all-pole filter with warping coefficient of $\alpha = 0.45$. This leads to 60 log Mel-Frequency Spectral Coefficients (MFSCs), representing the harmonic component of the WORLD vocoder parameters. Bandwise aperiodic analysis is used to reduce the dimensions of the aperiodic component of the parameters to 4. The compressed harmonic and aperiodic components are concatenated together to form a 64 dimension feature that we utilize

throughout this thesis, referred to as the **compressed spectral envelope**, $\mathbf{X_{voc}}$.

Other methodologies have been proposed to improve the quality of the synthesis of the channel vocoder, including the phase vocoder (Flanagan & Golden, 1966), Pitch-synchronous overlap add (PSOLA) (Moulines & Charpentier, 1990) and sinusoidal models of the voice (McAulay & Quatieri, 1986).The magnitude component of the spectrogram is also used as a type of vocoder and can be inverted to the waveform using the Griffin Lim (Griffin & Lim, 1984) algorithm. Over the last few years, deep neural networks have also been used as vocoder models.

2.4.2 Neural vocoders

The idea of learning low dimension representations of the acoustic features has also been explored using Deep Learning based algorithms, often termed as Neural Vocoders. The first such neural vocoder (Shen et al., 2018) used an adaptation of the WaveNet (van den Oord et al., 2016a) architecture, with upsampling, to invert spectrograms on a mel-frequency scale to corresponding speech waveform. In doing so, the WaveNet vocoder models the auto-regressive nature of the LPC based vocoder.

A number of neural vocoders have been proposed, including ones that use the source-filter model as described in the previous sections. Such neural vocoders include the LPCNet (Valin & Skoglund, 2019), GELP (Juvela et al., 2019a) , GlotGAN (Juvela et al., 2019b), Neural Homomorphic Vocoder (Liu et al., 2020) and Differentiable Digital Signal Processing (DDSP) (Engel et al., 2020a).

While such models have been quite effective for speech synthesis, their use for singing voice synthesis was still under investigation during the course of this thesis. The methodology we present in this thesis uses the WORLD vocoder, but is vocoder agnostic and can be replaced by a neural vocoder in the future.

2.4.3 Singing voice synthesis

TTS synthesis has several applications in *public communication*, *pedagogy* and *accessibility for disabled persons*. Artistic applications have also been researched, particularly for the singing voice. Generating a voice signal replicating the sound of a particular singer singing from a score is a research field onto itself, known as **singing voice synthesis** (SVS). Like its textual counter-part, SVS typically uses linguistic information in the form of a sequence of phonemes. However, the pitch is not dependent on the prosody, but is guided by the melodic information provided by the score. As discussed in Section 1.1.4, the f0 curve of a singing voice signal follows the melodic guidelines of the score with added embellishments and natural deviations of the singer. SVS aims to model both the f0 curve and the timbre of a singer given an input score in a machine-readable format such as *MusicXML*. SVS has been applied to commercial applications like *Vocaloid* (Kenmochi & Ohshita, 2007), *Sinsy*[3], *Melodyne* [4], *Utan* [5] and *CeVIO* [6]. Methodologies applied in this field include unit selection and concatenation (Bonada et al., 2016), HMM based synthesis (Saino et al., 2006; Oura et al., 2012) and in more recent years, deep learning based models like the Neural Parametric Singing Synthesizer (NPSS) (Blaauw & Bonada, 2017). Deep learning based SVS algorithms are discussed in detail in Section 2.4.5.

2.4.4 Synthesis with deep learning based models

The modelling capability of deep learning based models has opened by new avenues in the field of TTS and SVS. End-to-end TTS synthesizers capable of mapping the text directly to the speech waveform have been proposed, along with several other methodologies for modelling the various stages of the TTS pipeline.

Parametric TTS using deep learning (Zen et al., 2016; Ze et al., 2013) uses neural net-

[3] http://www.sinsy.jp/
[4] https://www.celemony.com/en/melodyne/what-is-melodyne
[5] http://utau.us/
[6] https://cevio.jp/

works to map a sequence of input linguistic and prosodic information to acoustic parameters used for synthesis of the voice signal. The WaveNet model (van den Oord et al., 2016a) introduced the concept of autoregressive convolutional neural networks. The model was used for waveform generation for speech and music. In the text-conditioned form of the model, *local* conditioning in the form of linguistic information and *global* conditioning the form of a vector representing the speaker identity is provided to the autoregressive network. The linguistic features are represented by a sequence of phonemes across time, while the speaker identity is represented as a one-hot vector that is assumed to be constant throughout the time. The global conditioning vector was broadcast throughout the time context of the linguistic features. The concatenated vector containing the global and local conditioning is up-sampled to the frequency to the desired waveform using transposed convolution. The output of the network at each time-step is conditioned on the conditioning vector for that time step and the output of the network at a series of previous time steps, the number of which pertains to the *receptive field* of the network. The output vector is a softmax distribution, representing the probability distribution of the output frame over 256 possible values of the μ law companding transformation of the waveform pertaining to the speech signal.

DeepVoice (Arik et al., 2017b) is a TTS model uses deep neural networks to model the various stages of the traditional TTS pipeline; it first converts the sequence of graphemes found in text to a sequence of phonemes. For the training phase, this sequence of phonemes is temporally aligned to the spectrogram of the audio using **connectionist temporal classification** (CTC) (Graves et al., 2006) loss. This alignment is used to train the subsequent components of the model, which include a phoneme duration model, which assigns a duration to the sequence of phonemes found in the input using the previous alignment. A fundamental frequency (f0) curve corresponding to the sequence of phonemes is also generated in this phase. This model is trained on the ground truth f0 of the speech signal, extracted via an external model. Finally, a synthesis model, that uses an architecture. similar to the WaveNet is used to generate the

Spectogram of the speech signal. The Griffin-Lim algorithm (Griffin & Lim, 1984) is used to synthesize the speech waveform from this output. DeepVoice 2 (Arik et al., 2017a) follows a similar structure to DeepVoice, except that the phoneme duration model and the f0 curve generation models are separated, with the output of the phoneme duration model being fed to the f0 curve generation model. The synthesis part of the architecture generates the waveform from the Spectrogram using an autoregressive network instead of the Griffin-Lim algorithm. Such a network is often called a neural vocoder and is discussed in detail in Section 2.4.2. DeepVoice 3 (Ping et al., 2018) utilizes an internal representation of linguistic features learned from the input sequence of characters and decodes this representation via an autoregressive network. The architecture used in this model is comprised solely of convolutional layers.

The Tacotron (Wang et al., 2017) model and Tacotron 2 (Shen et al., 2018) models introduced end-to-end TTS synthesis, synthesizing speech signals directly from text. The first of these takes as input a sequence of text characters and passes them through an encoder network, comprising of a series of LSTMs. A decoder layer which uses an attention mechanism is used to generate the linear scale spectrogram from the output of the encoder. Griffin-Lim (Griffin & Lim, 1984) is used to synthesize the waveform for the speech signal from the spectogram output. Tacotron 2 utilizes a neural vocoder based on the autroregressive architecture to synthesize the waveform from the Mel scale spectrogram that is generated by the decoder network.

Other deep learning based TTS system include the Char2Wav (Sotelo et al., 2017) model which is also an end-to-end TTS synthesis system that generates neural vocoder parameters given an input sequence of characters, and FastSpeech (Ren et al., 2019), which uses a feedforward transformer network for generating a Mel-Spectrogram from an input sequence of phonemes.

2.4.5 Neural networks for singing voice synthesis

Deep learning architectures have been applied to singing voice synthesis (SVS) over the last few years. One of the first such models (Nishimura et al., 2016)uses a series of fully-connected layers to map frame-wise linguistic and melodic features like current phoneme identity, absolute pitch of current musical note to *STRAIGHT* (Kawahara et al., 1999) vocoder parameters. These parameters are used to synthesize the waveform of the singing voice signal.

The Neural Parametric Singing Synthesizer (NPSS) (Blaauw & Bonada, 2016) used a WaveNet (van den Oord et al., 2016a) style autoregressive architecture to map frame-wise contextual linguistic features like the current phoneme identity, the previous phoneme identity and the next phoneme identity. These features are represented as one-hot encoded vectors. The normalized position of the current input frame is also input to the system, represented by a 3-state coarse coded vector. A multi-stream architecture is used to predict parameters of the WORLD (Morise et al., 2016) vocoder, which is used for synthesis of the singing voice signal. Unlike the WaveNet, which uses a categorical distribution as the output, the NPSS output was modeled as a continuous mixture density output (Salimans et al., 2017), with a combination of four continuous Gaussian components constrained to four parameters. The model was optimized to maximize the log-likelihood of the output distribution given the ground truth target distribution of vocoder features.

An extension of the model with phoneme duration and pitch prediction models was also proposed by (Blaauw & Bonada, 2017), which information from the score to generate a sequence of phonemes and a f0 curve that was fed into the previously described model. The full model, shown in Figure 2.19 was one of the first end-to-end Deep Learning based singing voice synthesis models, capable of generating a singing voice signal directly from an input score. Other singing voice synthesizers based on convolutional autoregressive networks have also been proposed by (Bous & Roebel, 2019; Yi et al., 2019).

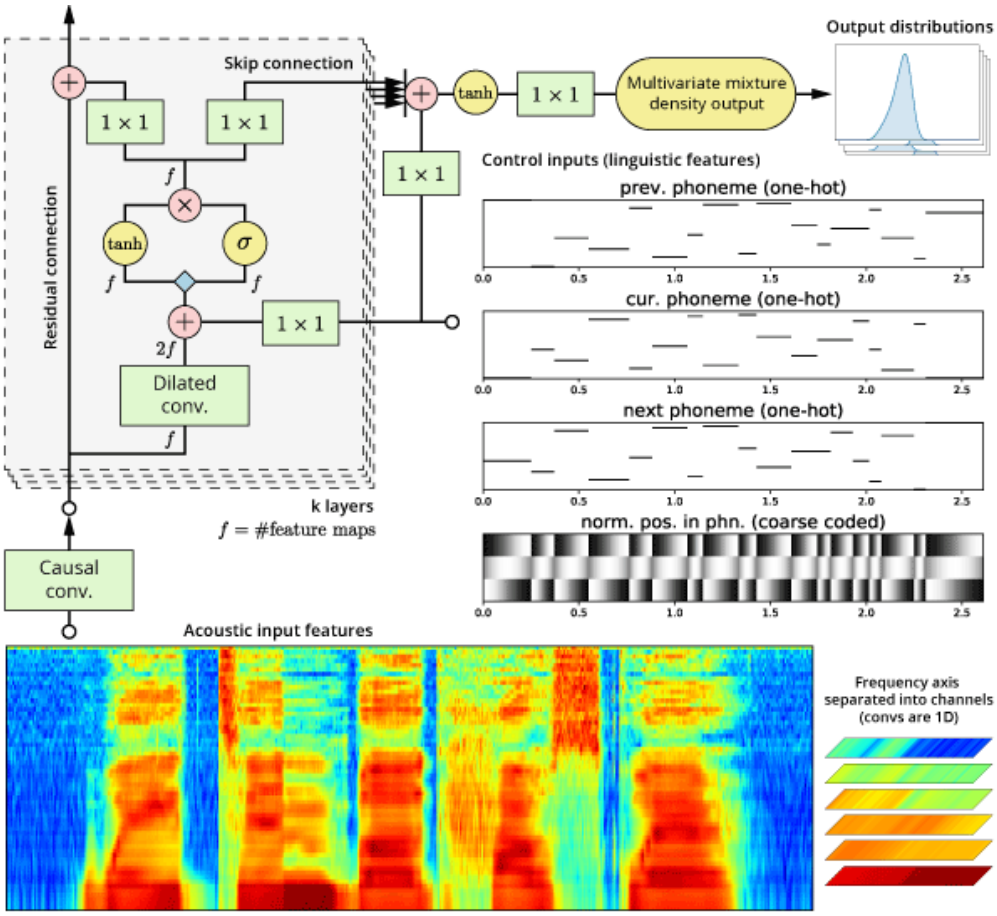

Figure 2.18: The Neural Parametric Singing Synthesizer (NPSS) proposed by (Blaauw & Bonada, 2016)

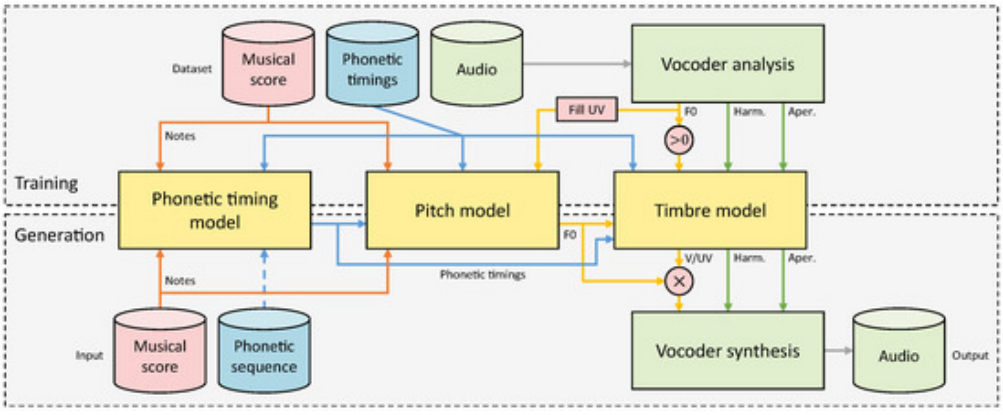

Figure 2.19: The full Neural Parametric Singing Synthesizer (NPSS) with phonetic timing and pitch prediction models proposed by (Blaauw & Bonada, 2017)

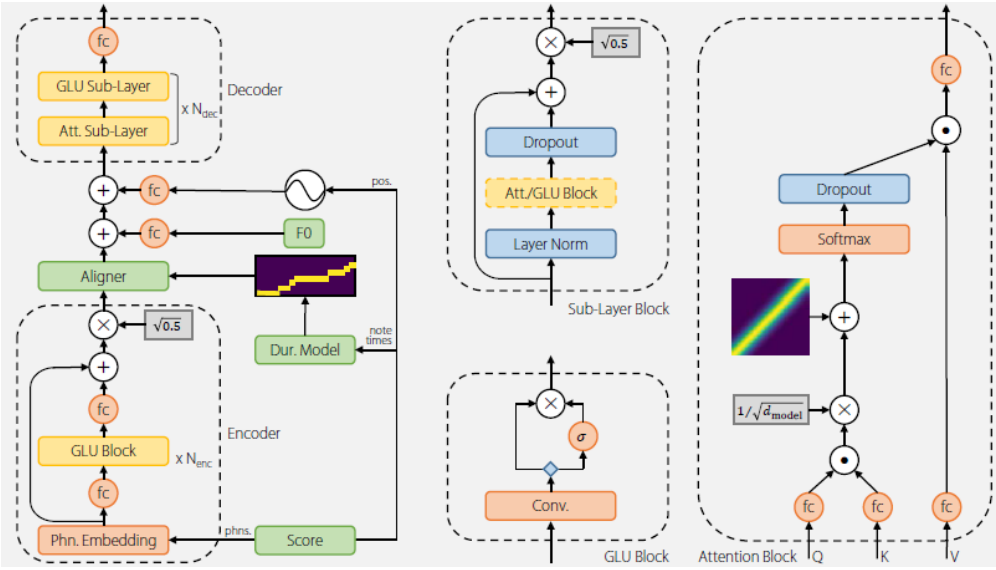

Figure 2.20: A transformer (Vaswani et al., 2017) based sequence-to-sequence model for singing voice synthesis, proposed by (Blaauw et al., 2019)

Feedforward convolutional neural networks have also been used for singing voice synthesis, principally by (Nakamura et al., 2019). The model proposed by the researchers models segments of several seconds of input frames containing linguistic and melodic information derived from the score using a series of convolutional layers. The CNN maps these input segments to the corresponding STRAIGHT (Kawahara et al., 1999) vocoder parameters, which are used for synthesis of the singing voice signals.

Sequence-to-sequence (seq2seq) models for singing voice synthesis have been proposed; (Blaauw et al., 2019) proposed a model based on the transformer network (Vaswani et al., 2017) with self-attention and convolutional layers, as shown in Figure 2.20. Such models do not require a pre-allignment for the linguistic and acoustic features and can produce the singing voice signal in an end-to-end manner. Other sequence-to-sequence SVS systems include the model proposed by (Lee et al., 2019a), which uses an adversarial trained system exploiting the syllable structure specific to the *Korean* language. Adversarial training methodology was also used by (Hono et al., 2019) to training a network for singing voice synthesis.

2.5 Music information retrieval

Music information retrieval (Müller, 2007) (MIR) is the field of research that leverages advancements in fields like signal processing, musicology, machine learning, psychology and psycho-acoustics to computationally extract information from musical representations in various forms. In the context of this thesis, we focus on retrieval of linguistic, melodic and singer identity related information given a polyphonic musical mixture as an input.

2.5.1 Linguistic features

Linguistic features pertain to the content of speech that humans can interpret when listening to each other talk. As discussed in Section 1.1.3, the study of linguistic features in speech perception has been dedicated to the identification of phonemes. Each language has its own set of phonemes, with their own pronunciation, although overlap has been found. The pronunciation of certain phonemes within a language might also depend on the accent of the speaker and emphasis that the speaker might be putting on certain parts of the speech. Lexicons like the *Carnegie Mellon University (CMU) Pronouncing Dictionary*[7] provide standardized definitions for phonetic pronunciations of words for some of the well studied and documented languages around the world like English, Spanish and Japanese.

Such lexicons are used in **automatic speech recognition** (ASR), which is a field of research that aims to derive the textual transcription from a speech signal. The corresponding field for the singing voice signal is known as **automatic lyrics transcription** (ALT). ASR and ALT methodologies use to derive sequences of phonemes from acoustic representations of the voice signal, like MFCCs. Phoneme sequences can then be mapped to words using lexicons. Such phoneme sequences are also used for voice synthesis, discussed in Section 2.4. Gaussian mixture model-Hidden Markov

[7] The CMU Pronouncing Dictionary provides definitions for pronunciations of North American English

Models (GMM-HMM) trained with Maximum Linear Likelihood Regression (MLLR) (Mesaros & Virtanen, 2009) have been used for both ASR and ALT. For the singing voice, most of the research haw been applied to monophonic a capella signals, with some researchers using source separation as an intermediate step (Mesaros, 2013). Over the last few years, Deep Learning based models have been applied to ALT, including bootstrapping combinations of HMMs and DNNs (Kruspe & Fraunhofer, 2016; Gupta et al., 2018a,b). Other Deep Learning based methodologies use Dilated Convolutional Neural Networks with Self-Attention (Demirel et al., 2020) and TDNN-BLSTM neural networks (Tsai et al., 2018). Training such systems requires expert aligned lexical annotations, typically in the form of phonemes.

Alternate representations of linguistic representation have been explored in the related tasks of zero resource synthesis (Jansen et al., 2013; Glass, 2012; Dunbar et al., 2019) and voice conversion (VC) (Mohammadi & Kain, 2017). Both tasks involve the extraction of speaker independent linguistic content from a speech signal and the subsequent synthesis of an intelligible speech signal with the linguistic content.

The task of zero resource synthesis is motivated by the linguistic representation learned by children learning to talk, without lexical knowledge. While not being able to completely understand language, infants are able to distinguish phonological sub-word units of speech (Kuhl et al., 2008), often termed as proto-phonemes (Dunbar et al., 2019). Discovering such units for synthesis from a speech signal is a task addressed by the The Zero Resource Speech Challenge. This challenge is typically organized in conjunction with the *INTERSPEECH* conference (Versteegh et al., 2015; Dunbar et al., 2017, 2019, 2020). Participants in the challenge are required to propose a synthesis system that can synthesize a speech signal from a language different to the one used for training the system. The proposed system is also required to synthesize a speech signal which retains the linguistic content of the input speech signal while changing the perceived identity of the speaker of the signal. The re-synthesized signal is evaluated through MOS based subjective listening tests with judges with native-level language

competency. The criteria for evaluation includes *intelligibility*, *naturalness* and *speaker similarity*.

Voice conversion is a closely related task aims to modify a voice signal in such a way that the perceived source speaker is changed to a target speaker, which maintaining the linguistic information that can be interpreted while listening to the audio. Contemporary methodologies for voice conversion. While voice conversion can be viewed as signal manipulation, most of the recently proposed algorithms use synthesis as a part of the process (Wu & Lee, 2020; Wu et al., 2020; Qian et al., 2019; Ganin et al., 2016; Nachmani & Wolf, 2019; Chou et al., 2018). The Voice Conversion Challenge (Yi et al., 2020; Lorenzo-Trueba et al., 2018; Wester et al., 2016) is also organized as a satellite workshop of the INTERSPEECH conference. Over the last few iterations, the task has evolved from parallel to non-parallel conversion, with multiple speakers. Most of the models proposed over the last few iterations have used deep learning based models. The evaluation of the participating system in this challenge is also via subjective listening tests.

The idea behind zero resource synthesis (Glass, 2012; Jansen et al., 2013) is that decipher based speech processing can be used to identify recurring structures or *motifs* within a speech signal. Such structures can can be used to define an appropriate set of sub-word units within a languages for which expert based phonetic pronunciation annotations are not easily available. Several methodologies have been proposed to discover such units (Chiu et al., 2003; Park & Glass, 2007; Zhang & Glass, 2010; Jansen et al., 2010; Siu et al., 2011), often referred to as **self-organizing units** (Glass, 2012) (SOUs). Such units form abstract representations of linguistic content that can be discreet or continuous in nature (Dunbar et al., 2019). Such representations have been used for applications like spoken query retrieval (Muscariello et al., 2009) and topic segmentation (Dredze et al., 2010) and classification.

To obtain such an abstract representation of linguistic content, several deep learning based models have been proposed to disentangle speaker specific information from the

linguistic content of a voice signal. In particular, unsupervised learning with autoencoders (Dunbar et al., 2019) has been applied to this task, with the hypothesis that a speech signal, $\mathbf{x}$, can be represented by an invertible low dimensional embedding vector, $\mathbf{V}$, and shown in Equation 2.13.

$$\begin{aligned} \mathbf{V} &= enc(\mathbf{x}) \\ \hat{\mathbf{x}} &= dec(\mathbf{V}) \end{aligned} \tag{2.13}$$

Where $enc()$ represents the encoder of an autoencoder and $dec()$ represents its decoder. The low dimensional vector representation, often termed as the *latent embedding*, is assumed to fully represent speaker specific information like prosody and timbre, as well as speaker independent linguistic information, as described in Section 1.1.3. Constraints like vector quantization (Wu & Lee, 2020; Wu et al., 2020) (VQ), instance normalization (Chou et al., 2019) (IN), dimensional restrictions (Qian et al., 2019), domain confusion (Ganin et al., 2016; Nachmani & Wolf, 2019) and adversarial training (Chou et al., 2018) are applied to the latent embedding to allow for disentanglement of the speaker specific components of the signal and the speaker independent components of the signal, as shown in Figure 2.21.

Vector quatization involves discretisation of the latent space of an autoencoder. Initially proposed as a generative model (Oord et al., 2017) for text, images and speech, the vector quantised variational autoencoder (VQ-VAE) consists of an encoder, a latent space and a decoder. The latent space is used as a lookup table for a set of one-hot vectors, known as the posterior categorical distribution, which quantizes the output of the encoder. This quantization is done giving a value of 1 to the category, the corresponding embedding of which in the latent space has the minimum distance to the output of the encoder. This embedding is fed to the decoder, which regenerates the input. Such an architecture can be seen as an autoencder with a non-linearity, represented by $q()$, that maps the continuous latent space to a discrete one-hot vector.

The vector quantization concept was adapted for the acoustic unit discovery in the

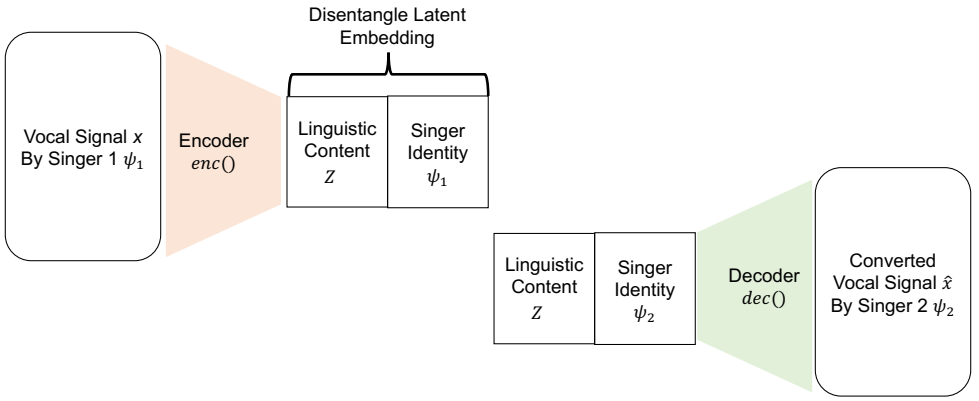

Figure 2.21: Voice Conversion via auntoencoders

ZeroSpeech 2020 challenge (van Niekerk et al., 2020) and subsequently for voice conversion (Wu & Lee, 2020; Wu et al., 2020). The VQVC and VQVC + models are two such Voice Conversion algorithms using VQ that assume that the discrete representation for each time step of a given input acoustic features from a voice signal represents the linguistic information in the signal. The model also assumes that speaker specific information is encoded in the difference between the continuous encoder output and the linguistic information. VQVC uses instance normalization (Ulyanov et al., 2017) for One-Shot Voice Conversion. VQVC+ (Wu et al., 2020) expands on this idea by using a U-Net architecture along with discretization along each layer of the encoder side. This operation is shown in Equation 2.14, with the encoder, $e_{vqvc+}()$ and the decoder $d_{vqvc+}()$ for a voice signal $\mathbf{x}$ with mel-spectogram features, $\mathbf{X_{mel}}$ over T time frames.

$$
\begin{aligned}
\mathbf{V}_{vqvc+} &= e_{vqvc+}(\mathbf{X}_{\mathbf{mel}}) \\
\mathbf{Z}_{vqvc+} &= q(\mathbf{V}_{vqvc+}) \\
s_x^{vqvc+} &= \mathbb{E}[\|\mathbf{V}_{vqvc+} - \mathbf{Z}_{vqvc+}\|] \\
S_{vqvc+} &= \underbrace{\{s_x^{vqvc+}, s_{vqvc+}, s_{vqvc+}.....s_{vqvc+}\}}_{\text{T times}} \\
\hat{\mathbf{X}}_{\mathbf{mel}}^{vqvc+} &= d_{vqvc+}(\psi^{vqvc+} + \mathbf{Z}_{vqvc+})
\end{aligned}
\tag{2.14}
$$

The network is trained using a reconstruction loss, $\mathcal{L}_{recon}^{vqvc+}$ and a latent loss, $\mathcal{L}_{latent}^{vqvc+}$, the weighted sum of which gives the final loss used for training, $\mathcal{L}_{final}^{vqvc+}$. This loss is shown in Equation 2.15.

$$
\begin{aligned}
\mathcal{L}_{recon}^{vqvc+} &= \mathbb{E}[\|\hat{\mathbf{X}}_{\mathbf{mel}}^{vqvc+} - \mathbf{X}_{\mathbf{mel}}\|_1^1] \\
\mathcal{L}_{latent}^{vqvc+} &= \mathbb{E}[\|IN(V) - \mathbf{Z}_{vqvc+}\|_2^2] \\
\mathcal{L}_{final}^{vqvc+} &= \mathcal{L}_{recon}^{vqvc+} + \lambda_{vqvc+}\mathcal{L}_{latent}^{vqvc+}
\end{aligned}
\tag{2.15}
$$

Where $IN()$ represents the Instance Normalization layer and $lambda_{vqvc+}$ represents the weight given to the latent loss. The mel-spectrogram features are used to generate the waveform of the voice signal using a WaveNet vocoder [8] (Shen et al., 2018). (van den Oord et al., 2016a).

The AutoVC (Qian et al., 2019) model is another effective methodology for zero-shot voice conversion. One-shot and zero-shot conversion refers to a conversion from a voice sample where both the source and target speakers may be from outside the training dataset. The model also utilises an autoencoder framework with an encoder, a bottleneck and a decoder. As shown in in Figure 2.22, the input to the system as well as the target is the mel-spectrogram representation of the voice signal, $\mathbf{X}_{\mathbf{mel}}$, pertaining to samples of a voice signal. These features, along with a speaker identity representation vector, ψ are passed through a *content encoder*, e_{autovc}, which consists of a series

[8] Vocoders are discussed in Section 2.4.1

of convolutional layers followed by batch normalization and a BLSTM. This encoder produces a latent embedding, $\mathbf{V}_{autovc}$.

A bottleneck size restriction is imposed on the latent embedding of the autoencoder, by using downsampling by a factor of 32, leading to a **content embedding**, $\mathbf{Z}^{autovc}$. This constraint limits the capacity of the embedding to represent information from which the decoder can reconstruct the input features. The content embedding is upsampled by repetition to the the size of the input, resulting in a vector, $\tilde{\mathbf{V}}_{autovc}$, which is passed along with the singer identity vector to the decoder, which is trained to reproduce the mel-spectrogram, $\hat{\mathbf{X}}^{autovc}_{\mathbf{mel}}$, as shown in Equation 2.16. This is inverted to the waveform using a WaveNet vocoder (van den Oord et al., 2016a), to produce the output signal, $\hat{\mathbf{x}}$. The researchers who proposed the model prove that by imposing a such a limitation on the bottleneck forming the latent space of the autoencoder constraints it to learn only the speaker independent linguistic information, if the speaker identity is provided to the decoder of the autoencoder.

$$\begin{aligned} \mathbf{V}_{autovc} &= e_{autovc}(\mathbf{X}_{\mathbf{mel}}) \\ \mathbf{Z}_{autovc} &= downsample(\mathbf{V}_{autovc}) \\ \tilde{\mathbf{V}}_{autovc} &= upsample(\mathbf{Z}_{autovc}) \\ \hat{\mathbf{X}}^{autovc}_{\mathbf{mel}} &= d_{autovc}(\psi + \mathbf{Z}_{autovc}) \end{aligned} \tag{2.16}$$

Where $downsample()$ and $upsample()$ represent the downsampling and upsampling operations respectively. For training the model, a reconstruction loss, $\mathcal{L}^{autovc}_{recon}$ and a content loss $\mathcal{L}^{autovc}_{content}$, resulting in the final loss, $\mathcal{L}^{autovc}_{final}$ [9]. The content loss maintains cyclical consistency between the input signal, $\mathbf{x}$ and the output $\hat{\mathbf{x}}$, and is calculated by passing the output mel-spectrogram through the content encoder.

[9] A post-net is also used in the AutoVC model, but is omitted here for simplicity

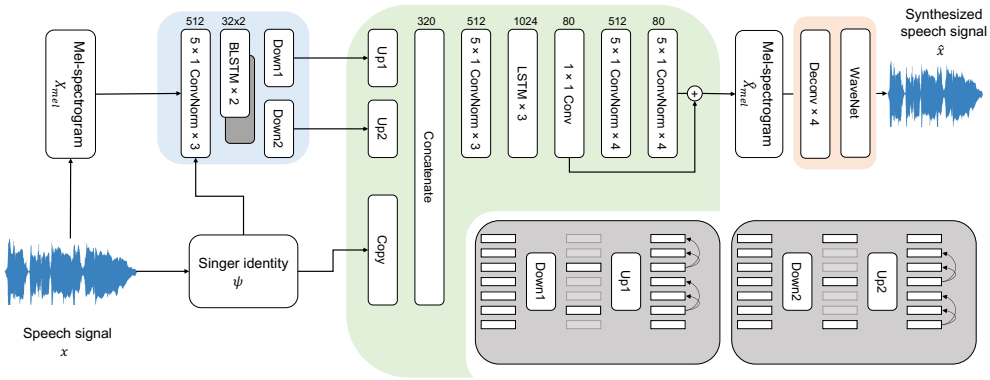

Figure 2.22: The AutoVC architecture for Zero Shot Voice Converstion (Qian et al., 2019)

$$\begin{aligned} \mathcal{L}_{recon}^{autovc} &= \mathbb{E}[\|\hat{\mathbf{X}}_{\mathbf{mel}}^{autovc} - \mathbf{X}_{\mathbf{mel}}\|^2] \\ \mathcal{L}_{content}^{autovc} &= \mathbb{E}[\|\mathbf{Z}_{autovc} - downsample(e_{autovc}(\hat{\mathbf{X}}_{\mathbf{mel}}^{autovc}))\|] \\ \mathcal{L}_{final}^{autovc} &= \mathcal{L}_{recon}^{autovc} + \lambda_{autovc}\mathcal{L}_{content}^{autovc} \end{aligned} \tag{2.17}$$

For inference, the vector passed to the decoder is one that represents the identity of the target speaker. Such vectors representing speaker identity are commonly known as *speaker embeddings* and are discussed in Section 2.5.3. The AutoVC architecture has been tried with one-hot speaker representations as well as speaker embeddings directly derived from audio, which allow it to perform zero-shot voice conversion. These speaker embeddings are discussed in Section 2.5.3.

The AutoVC architecture has been extended to disentangle linguistic content, timbre, pitch and rhythm from a speech signal, using a triple information bottleneck (Qian et al., 2020). This includes a random shuffling of the input time series so that it only retains linguistic content and not rhythmic content. Other deep learning methodologies proposed for Voice Conversion include the use of GANs (Kameoka et al., 2018b), Instance Normalization (Chou et al., 2019), sequence-to-sequence networks (Huang et al., 2020; Hwang et al., 2020) and Normalizing Flow (Serrà et al., 2019).

As such, these methodologies provide us a means to represent linguistic information of

a voice signal in an abstract and language independent manner. Additionally, such as representation does not require labelled data, which is hard to obtain. However, these techniques have been proposed for the speech signal, which, as discussed in Section 1.1.3, differs from the singing voice signal. In Chapter 7, we study how such methodologies can be adapted for the case of the singing voice.

2.5.2 Melody

The melody of a musical piece is *musicological* concept, defined in different contexts (Poliner et al., 2007; Typke, 2007; Ryynänen & Klapuri, 2008). One commonly accepted definition is related to the pitch sequence perceived by a listener, as *"the melody is the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of music, and that a listener would recognise as being the essence of that music when heard in comparison"* (Poliner et al., 2007). The perceived pitch, as discussed in Section 1.1.3, is related to the **fundamental frequency** (f0) of the signal. Research has focused on extracting the f0 of monophonic signal.Since the f0 is a measure of the rate of repetitions within a signal, the *auto-correlation* function is the most intuitive method for its computation. Auto-correlation is a measure of the similarities between observations of a signal across time, calculated by the taking the correlation of a signal with a delayed version of itself. Auto-correlation in both the time (Dubnowski et al., 1976) and frequency domains (Rahman & Shimamura, 2010) has been proposed for f0 estimation (Lahat et al., 1987). Like auto-correlation, the *cepstrum* (Noll, 1967) or the inverse Fourier transform of the log-scale magnitude component of the spectrum of a signal, also provides information about the periodicity of a signal and is used for f0 estimation.

The auto-correlation function however, is suspect to peak amplitude changes and calculating the f0 in such a manner is prone to octave errors. The Yin (De Cheveigné & Kawahara, 2002) algorithm was proposed to alleviate this problem by using the cumulative mean normalized difference function with the auto-correlation function to

estimate the f0. The pYin (Mauch & Dixon, 2014) algorithm expands on this idea by estimating a probability distribution over a range of f0 candidates, followed by HMM based pitch tracking to pick out the most probable f0 contour. HMMs can also then be used for melody note annotation (Mauch et al., 2015). Other algorithms proposed for monophonic f0 estimation include the average magnitude difference function (AMDF) (Ross et al., 1974), the normalized cross-correlation function (NCCF) (Talkin & Kleijn, 1995), DIO (Morise et al., 2009) and SWIPE (Camacho & Harris, 2008) which matches the input signal with templates of a sawtooth waveform.

While such methodologies work for monophonic signals, they are not applicable to polyphonic signals. Indeed, the definition of the melody for a polyphonic context is different as there might be several melodic or harmonic instruments playing simultaneously in a musical piece. The concept of *predominant melody* has been proposed as the principle melody that a listener would recognize when listening to a piece of music (Poliner et al., 2007). In particular, it has been compared to the monophonic pitch sequence of the predominant melodic source present in a musical mixture. Methodologies using spectral peaks as f0 candidates have been proposed for predominant melody estimation (Ryynänen & Klapuri, 2008; Goto, 2004; Salamon & Gómez, 2012; Salamon et al., 2013b; Paiva et al., 2006; Dressler & Fraunhofer, 2009; Klapuri, 2006). The f0 contour of the predominant melody is then estimated from these candidates by using various heuristics including probabilities weighted on the presence of harmonics (Goto, 2004; Klapuri, 2006).

One such algorithm is called **Melodia** (Salamon & Gómez, 2012; Salamon et al., 2013b). In this algorithm, shown in Figure 2.23, the polyphonic mixture signal is first passed through an equal-loudness filter and an STFT transformation, followed by frequency and amplitude correction. Harmonic weighting is used to compute to salience function, from which the pitch contour is extracted using peak detection and contour characterization. Finally, an f0 contour is estimated using iterative melodic peak selection. Other methodologies for estimating the predominant pitch in a poly-

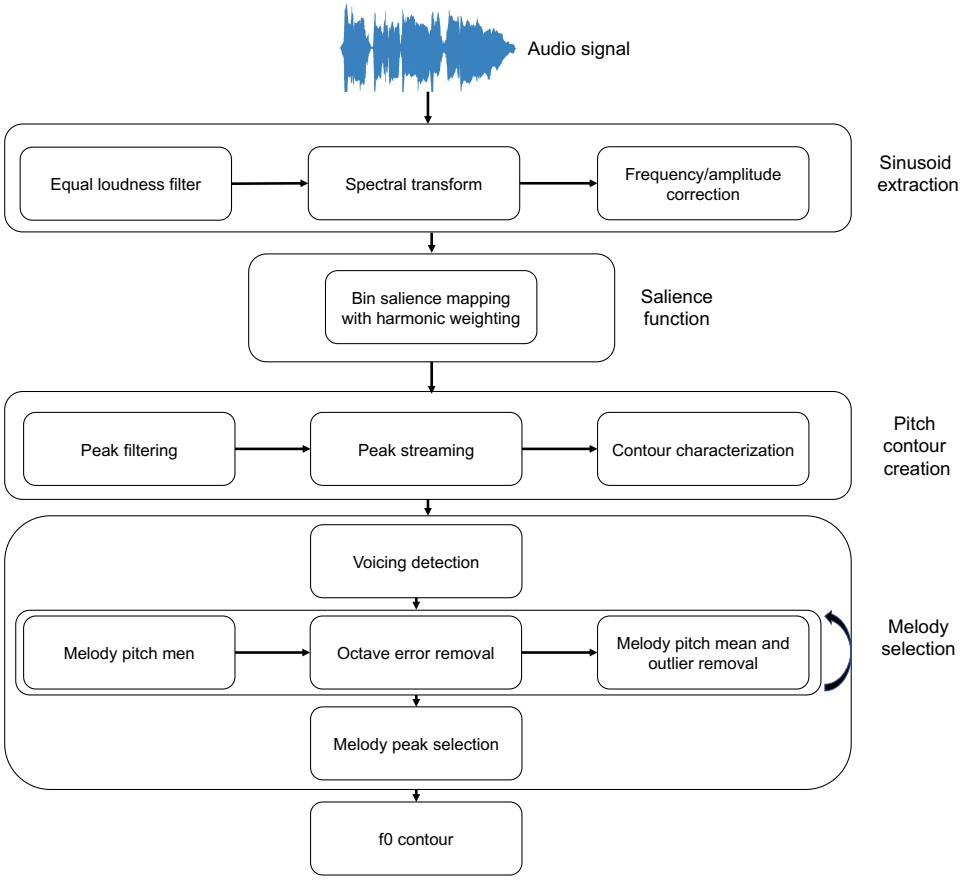

Figure 2.23: The framework for Melodia algorithm (Salamon & Gómez, 2012)

phonic musical signal include the use of the source-filter model (Bosch et al., 2016), Support vector machines (SVM) for classification of pitch contours (Ellis & Poliner, 2006; Poliner & Ellis, 2005) as well as hand-crafted features (Bittner et al., 2015).

Over the last few years, deep learning based algorithms have been applied to the task of f0 estimation for monophonic signals. The **convolutional representation for pitch estimation** (Kim et al., 2018a) (CREPE) is one such algorithm, taking the waveform of a musical signal as input. The architecture for this model consists of a series of convolutions, the end result of which is a 360-dimensional vector. The bins of this vector represent the probability distribution of f0 values defined in **cents**. The 360 bins cover a range between 32.70 Hz and 1975.5 Hz with a resolution of 20 cents per bin.

Binary cross entropy is used to train the network with a ground truth target distributed over the bins as a one-hot vector with Gaussian blurring with a standard deviation of 25 cents The **SPICE** (Gfeller et al., 2020) is another deep learning based algorithm that makes innovative use of data without annotations by applying a determined amount of pitch shift to the Constant-Q Transform (CQT) (Velasco et al., 2011) of the input signal. The difference between the f0 estimate for the original signal and the pitch-shifted signal is used to train the network.

Deep Learning has also been applied to predominant melody estimation for polyphonic signals, particularly for estimating the vocal melody in a polyphonic mixture (Rigaud & Radenen, 2016; Kum et al., 2016; Su, 2018; Dong et al., 2019; Jansson et al., 2019). Methodologies proposed include the use of a *residual convolutional network* (ResNet) (Doras et al., 2019) which uses skip connections between consecutive convolutional layers to allow for deeper propagation of information. Joint estimation of the f0 contour corresponding to the the vocals as well as source separation has been proposed (Jansson et al., 2019; Gao et al., 2021). The harmonic constant-Q transform (HCQT) has been proposed for f0 estimation, involving multiple CQTs over different frequency ranges to provide higher resolution. Deep salience representations (Bittner et al., 2017) have been proposed for representing the output of neural networks for this task in a categorical format and have been used for multiple pitch estimation (Cuesta et al., 2020) as well as vocal melody extraction using a U-Net architecture (Doras et al., 2019).

Aside from predominant pitch estimation, the estimation of the f0 of multiple melodic instruments playing simultaneously in polyphony has also been addressed (Cañadas Quesada et al., 2010; Arora & Behera, 2015; Bittner et al., 2017). Such algorithms try to discern the various pitch contours that might be present in a signal at a given time.

2.5.3 Singer identity

Speaker identification is a field of research that aims to obtain a low dimension representation of speaker specific characteristics from a voice signal. Such a represent-

ation should allow distinction between various speakers regardless of the content of the voice signal. Joint factor analysis (JFA), representing the signal using a set of low-dimensional total variability factors has been proposed for this task. The low dimension representation obtained using such an approach is commonly known as an **i-vector** (Dehak et al., 2010). Deep neural networks have also been applied to this task, extracting representations known as **d-vectors** (Variani et al., 2014).

Such low dimensional embeddings are not only useful for speaker identification, but have also been applied for zero-shot voice conversion, most notably in the AutoVC (Qian et al., 2019) model. The AutoVC model uses a 256 unit represent ion of speaker identity (Wan et al., 2018) which is derived using a stack of 2 LSTM layers trained using a **generalized end-to-end** (GE2E) loss. The idea behind the loss is to build a similarity matrix to define the similarities between various utterances by a single speaker. The VQVC (Wu & Lee, 2020) and VQVC+ (Wu et al., 2020) models for voice conversion model such speaker identity representation vectors as the difference between the latent representation of the speech signal and the quantized representation of the linguistic content.

Deep Learning methodologies have also been applied to derive vectors for representing singer identity, both from monophonic a capella singing and polyphonic musical mixtures (Lee & Nam, 2019). The model is trained using a margin-based hinge rank loss with cosine similarity (Frome et al., 2013) using an anchor, a positive example and multiple negative examples.

2.6 Summary

Source separation has been applied to the voice signal, particularly in the context of the singing voice mixed with instrumental accompaniment. Over the last decade, deep learning based algorithms have shown high potential to separate the voice signal from other instruments. However, most of these algorithms are based on the estimation of

time-frequency (TF) masks, which are applied to the mixture to filter out the voice signal. Such algorithms assume that the musical mixture is a linear sum of the sources.

We have also studied voice synthesis algorithms, which have also been revolutionized by the advent of deep learning. We note that singing voice synthesis algorithms typically take three components of the voice as input; the singer independent linguistic and melodic content and the singer identity, which is used to generate the timbre for the voice signal. We also note that low dimensional representations, known as vocoder parameters, of the voice signal are often used for voice synthesis and analysis.

Research in the MIR field has been focused on estimating perceptually relevant features from a musical signal. For contemporary popular music, such features include the predominant melody, the lyrics and the identity of the singer singing a song. Deep learning based algorithms have been proposed for all three of these tasks, which are pertinent to our thesis. Automatic lyrics transcription algorithms, which extract the linguistic content from a signal, generally estimate a sequence of phonemes present in the signal. This imposes language constraints on the algorithm. To overcome these constraints, we explore a more abstract representation of linguistic content, which is often used in low resource synthesis and voice conversion algorithms. Deep learning algorithms have been proposed to estimate such representation via autoencoders with restraints on the bottleneck. We propose to combine such representations with synthesis methodologies to synthesize the underlying vocal signal in a musical mixture. The framework for such a methodology is presented in Part II.

We also believe that TF mask based source separation methodologies can be applied to separate the individual parts within an SATB ensemble choir recording, that generally involves a linear sum of the individual soprano, alto, tenor and bass parts. Within the separated parts, there are often multiple singers simultaneously singing the same content in unison. We apply the methodology proposed in Part II to synthesize a prototypical single singing voice signal representative of the unison. The framework for separate the individual parts and synthesizing the prototypical signal is presented in

Part III.

Chapter 3

Datasets and evaluation strategies

Algorithms using data-driven deep learning based methodologies generally require large scale specialized datasets, typically with annotations. In this thesis we focus on contemporary popular music as well as choral singing. For contemporary popular music, we require datasets with clean unprocessed singing voice signals from multiple singers along with instrumental accompaniment and phonetic annotations. We initially tried recording such a dataset with sufficient and balanced coverage of linguistic and musical features. However, due to unprecedented circumstances encountered during the process, we could only record 9 songs, which is not sufficient for data-driven algorithms. Instead, we decided to leverage publicly available datasets which met our requirements.

In this Chapter, we look at some datasets for the singing voice, particularly in the context of contemporary popular music in Section 3.1 and choral singing in the SATB format in Section 3.2. We also look at evaluation strategies for voice synthesis and source separation algorithms in Section 3.3.

3.1 Contemporary popular music

One of the first publicly available datasets with separate singing voice and instrumental accompaniment stems was the **Music Audio Signal Separation** (MASS) data-

set (Vinyes, 2008), developed at the Music Technology Group (MTG) in the Universitat Pompeu Fabra (UPF). It only contained 2.5 minutes of data and later evolved into the **QUASI** dataset (Araki et al., 2012), with 11 tracks by professional sound engineers and was used in the 2010 and 2011 SiSEC evaluation campaigns.

Larger scale datasets were later developed including the **MIR-1K dataset** (Hsu & Jang, 2009) and the **ccMixter dataset** (Liutkus et al., 2015), which has 50 stereo tracks. The **iKala dataset** (Chan et al., 2015) is also one such dataset, which we use in this thesis. It contains 252 tracks, each of 30 seconds in duration. Clean unprocessed vocals and musical accompaniment are present for the songs in this dataset, along with manually annotated MIDI-note pitch annotations for each of the vocal tracks. The majority of the songs in the corpus are in the Chinese Mandarin language, although there are some English language songs as well. Both male and female singers are present in the dataset, which we use for training and evaluating one of our first systems for estimating synthesis parameters from musical mixtures, presented in Chapter 5.

The **DSD100 dataset** (Liutkus et al., 2017) contains 100 tracks with vocals, bass drums and others stems from Mixing Secret Free Multitrack Download Library of the Cambridge Music Technology group. While it is one of the largest and commonly used datasets for evaluating source separation algorithms, it was unfeasible for our study as the raw vocal tracks are not available. This problem is solved by the **MedleyDB dataset** (Bittner et al., 2014). This dataset contains 109 songs with raw recordings and processed stems for each of the individual instruments including the vocals. Vocals are present in 59 of the songs. We use this dataset for evaluation of our system for synthesizing the vocal signal from linguistic and melodic information extracted from a musical mixture, presented in Chapter 7. The **MUSDB18** (Rafii et al., 2017) dataset combines the MedleyDB and DSD100 datasets to present a collection of 150 full-length tracks with individual stems.

An English language dataset with phonetic annotations used in this thesis is the **National University of Singapore sung and spoken lyrics corpus** (Duan et al., 2013)

(NUS corpus), which consists of 48 popular English songs both sung and spoken by 12 non-professional singers, who were non-Native English speakers. Phonetic annotations are present for this dataset, with 25 474 phoneme annotations for a set of 20 songs. There are 6 male and 6 female singers, each of whom sing and speak the lyrics of 4 distinct songs from the set of 20 songs. We use this dataset for training and evaluating a multi-singer singing voice synthesizer trained with the Wasserstein-GAN methodology, presented in Chapter 6.

Other datasets pertaining to the singing voice in the context of contemporary popular music include the DAMP (Smith, 2013), DALI (Meseguer-Brocal et al., 2019), IRMAS (Bosch et al., 2012) and JVS-Music (Tamaru et al., 2020) datasets a summary of which is presented in Table 3.1.

3.2 Choral singing

Recording the individual singers of a choir ensemble in a realistic setting is a challenging task. The musical arrangement of a choir requires multiple singers to sing in synchronization. This involves interaction between the individual singers in the choir (Ternström, 2002), which needs to be captured in a realistic recording of a choir. Recording individual singers within the full choir setting can lead to signal leakage within the recordings. In the last few years, some attempts have been made to record individual singers within the choral setting, by using special directional microphones and singer isolation configurations.

The Choral Singing Dataset (CSD) (Cuesta et al., 2018) is one such recently published dataset. Initially proposed for the purpose of analyzing unison singing, the dataset contains a total of 7 min of audio recordings of 16 individual singers singing 3 songs in the SATB format. Each of the parts; soprano, alto, tenor and bass, was recorded in isolation, with 4 singers per part. Individual singers within the parts were recorded with dynamic handheld microphones, to minimize signal leakage between the singers

Name	Language	Annotations	Instrumentation
AITSS (Dai et al., 2015)	English	Fundamental frequency	No
DALI (Meseguer-Brocal et al., 2019)	English	Word and MIDI note	Yes
DAMP (Smith, 2013)	Multi-Lingual	Lyrics	No
Hansen's Dataset (Hansen & Fraunhofer, 2012)	English	Word	No
IRMAS (Bosch et al., 2012)	English	Instrument	Yes
Jamendo Corpus (Stoller et al., 2019)	English	Word	No
Jingju (Beijing Opera) corpus (Caro Repetto & Serra, 2014)	Chinese Mandarin	Syllable	No
JVS-MuSiC (Tamaru et al., 2020)	Japanese	Phoneme	No
Mauch's Dataset (Mauch et al., 2011)	English	Word	No
MTG-QBH (Salamon et al., 2013a)	English	Song	No
NIT-SONG070-F0013[a]	Japanese	Phoneme and MIDI note	No
PJS (Koguchi et al., 2020)	Japanese	Phoneme	No
RWC Music Database (Goto et al., 2004)	Japanese and English	None	Yes
Singing Voice Audio Dataset (Black et al., 2014)	Chinese	None	Yes
Tohoku Kiritan (Moris)	Japanese	Phoneme	No
TONAS (Mora et al., 2010)	Spanish	Musical notation	No
Tunebot (Huq et al., 2010)	Multi-Lingual	None	No
UltraStar[b]	English	Singer traits	No
Umbert's Dataset (Umbert et al., 2013)	Vowel only singing	Phoneme	No
VocalSet (Wilkins et al., 2018)	Vowel only singing	Singing technique	No

Table 3.1: Datasets with singing voice for contemporary popular music.

[a]Recorded by the Nagoya Institute of Technology (Nitech), available at http://hts.sp.nitech.ac.jp/archives/2.3/HTSdemo_NIT-SONG070-F001.tar.bz2
[b]https://usdb.eu/about

singing in unison. The 3 songs in the dataset are *Niño Dios*, in the Spanish language, *El Rossinyol*, in the Catalan language and *Locus Iste* in Latin. The dataset includes aligned annotations related to the f0 and MIDI notes as well as singer identity for each of the tracks.

The Dagstuhl ChoirSet (Rosenzweig et al., 2020) (DCS) is another dataset with 55 min of recordings pertaining to SATB choral singing. The dataset includes 2 songs; Locus Iste and *Tebe Poem*, a Bulgarian song. These songs were recorded with 13 singers, distributed into the soprano, alto, tenor and bass parts with a variable number of singers per part. All singers were recorded simultaneously, with different types of microphones including dynamic, handset and throat microphones. While the handset microphones provides recordings of each of the individual singers, the level of inter-singer leakage is higher than that in the CSD. We also note that a high amount of the 55 min of recording pertains to vocal exercises which cannot be used to distinguish parts in an SATB choir setting.

Other proprietary datasets pertaining to SATB choral singing include the ESMUC Choral Dataset (ECD), which is a proprietary dataset of the Escola Superior de Música de Catalunya (ESMUC) and the Bach Chorales Dataset (BCD). The ECD contains 20 min of recordings of 13 singers singing 3 songs in the German and Latin languages in the SATB format. The singers were recorded simultaneously with handheld dynamic microphones to reduce inter-singer leakage. However, it must be noted that the individual tracks in this dataset have significantly higher leakage than those in the CSD and DCS datasets. The BCD has 20 min of recordings of 4 singers performing 26 songs in individual isolated setups. This dataset has the least amount of inter-singer leakage and has been used for transcription experiments (Schramm & Benetos, 2017; McLeod et al., 2017).

3.3 Evaluation Strategies

Evaluation of synthesis and source separation methodologies is highly subjective and often depends on the end application that the methodology is designed for. A voice synthesis algorithm designed for a public announcement system, might need the synthesis to be clear and **intelligible** while transmitting the desired information to the public. On the other hand, some systems need the sound to sound more **natural** or indeed be representative of the prosodic and timbrel voice identity of a certain individual. Although hard to define objectively [10], a high **quality** voice synthesis is generally desired. Source separation algorithms might require the signal to be completely free of artifacts and interference from other sources while preserving the quality of the signal. However, different algorithms might have different levels of compromise between these criterion.

A common methodology used for quantifying such an evaluation for both synthesis and source separation algorithms is to have listening tests wherein various participants are explained certain criteria and listen to a few audio examples. The participants are asked to rate each of the examples on a scale that. generally goes from a low number to a higher number, indicative of the participants opinion of the audio example based on the criteria provided. An arithmetic mean is usually taken to represent the opinion of the set of the participants and is called the **mean opinion score** (MOS).

3.3.1 Voice synthesis evaluation

Voice quality assessment is an important task for voice transmission and standards and guidelines for such are usually set by the *International Telecommunication Union*. The MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) (Schoeffler et al., 2015) methodology is a variation of the MOS listening test. It is defined by ITU-R recommendation BS.1534-3 and provides guidelines for trained expert listeners to

[10] For some artistic applications, particularly in the case of music, certain degrees of change in what would normally be called quality could be desired.

evaluate the perceived quality of a speech speech signal. Such tests were traditionally carried out in controllable conditions, following standards like the ITU-T Recommendation P.800. Recently, crowdsourced evaluation (Naderi et al., 2020), allowing anonymous users to participate in evaluation processes has recently come to the fore. Standards such as the ITU-T Recommendation P.808 have been provided to allow for quality evaluation. High quality evaluations use around 100 participants sampled from a selected demographics (Naderi et al., 2020).

In practice however, such standards are hard to meet for evaluation of multiple systems. To get a statistically significant result from such listening test, a large number of participants is required. The demographic of participants for such subjective listening tests is very important and depends on the application intended. Also, the listening environment and other external conditions for the participants must be controlled and homogeneous to allow for fair and unbiased comparisons between systems. The Blizzard TTS challenge (Zhou et al., 2020; Wu et al., 2019; Karaiskos et al., 2008), which is a community based challenge to evaluate TTS systems stipulated a minimum of 10 participants per system for evaluation. Such participants included self-declared speech experts, volunteers recruited via social media and paid university students. The two branches of the evaluation campaign consisted of 6 and 7 sections, 17 and 9 samples, respectively.

Another form of subjective tests is the **AB testing**, which can be used to compare one synthesis system to another. The participants in this case are provided sets of examples of the two systems to be evaluated, without the knowledge of which one corresponds to which system. As shown in in Figure 3.1, the examples are generally labelled *A* and *B* and the participant is asked to choose one of them based on preference with respect to a provided criteria. A reference example to allow the participant to better judge the samples based on the criteria is often provided.

Given the difficulty in carrying out subjective listening tests, several perceptually motivated objective metrics have been proposed to judge the quality of a given speech

signal. Perceptually motivated measures, based on psycho-acoustics, have been proposed using models like the gammatone filter bank, one-third octave band filter bank, articulation index (AI) (French & Steinberg, 1947), speech-transmission index (STI) (Steeneken & Houtgast, 1980). Such measures include the perceptual speech quality measure (PSQM) (Beerends & Stemerdink, 1994), which was adapted as the (ITU-T) recommendation P.861 in 1996, along with measuring normalizing blocks (MNB) (Voran, 1999). Further metrics like the perceptual evaluation of audio quality (PEAQ) (Thiede et al., 2000) and the perceptual audio quality measure (PAQM) (Beerends & Stemerdink, 1992) were later adapted as ITU-R recommendation BS.1387 in 1999. Further extensions to the perceptual models for speech quality evaluation include the the bark spectral distortion (BSD) (Wang et al., 1992) and the perceptual analysis measurement system (PAMS) (Hollier et al., 1993). The perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) uses a combination of several perceptual models and is one of the most widely used evaluation metrics, following adaptation in the ITU-T recommendation P.862. The Perceptual Evaluation of Audio Quality (PEAQ) (Thiede et al., 2000) metric has been proposed to assess the objective quality of a distorted audio signal, given a high quality signal. This metric follows the listening test standards for ITU-R BS.1116. Another metric commonly used is the PEMO-Q (Huber & Kollmeier, 2006), which uses representations based on psycho-acoustically motivated cognitive aspects of the signal to calculate a perceptual similarity measure (PSM). The short-time objective intelligibility (STOI) (Taal et al., 2010) metric is a robust speech intelligibility measure that has been widely used. It has shown to be language invariant.

While these metrics require a ground truth reference to compare the signal to be evaluated against, some non-intrusive measures for speech audio quality have also been proposed. One such measure is the speech-to-reverberation modulation energy ratio (SRMR) (Falk et al., 2010) metric, which uses an auditory-inspired modulation spectrum representation with 23 gammatone filterbank channels to represent the signal and takes the ratio of the average modulation energy content in the first four bands to the last

four bands. Further improvements to this model have been proposed, including some specifically for cochlear implants (Santos et al., 2014; Santos & Falk, 2014). The non-intrusive codebook-based STOI (NIC-STOI) (Sorensen et al., 2017) uses codebooks of filter coefficients to model the spectral envelope of the speech and noise content in a signal. These are used to estimate a measure similar to the STOI.

Deep neural networks have also been used to directly evaluate the quality of a speech signal, the DNSMOS (Reddy et al., 2020) uses a CNN to map audio samples to ratings obtained using the MOS methodology following the ITU-T P.808 standard. The MOSNet (Lo et al., 2019) has been proposed to objectively evaluate Voice Conversion algorithms based on MOS ratings. Other objective metrics that have been proposed for assessing speech quality include Perceptual Objective Listening Quality Analysis (POLQA) (Beerends et al., 2013).

Objective measures for evaluating general audio quality using the distance between intermediate layers of neural networks trained for audio classification have also been proposed. The intuition between using this is that the intermediate layers of deep neural networks inherently learn representation of the audio data that are perceptually relevant. The discriminators used in training of GANs are also used for providing measures of audio and speech quality. Metrics such as the Inception Score (Salimans et al., 2016), Fréchet Audio Distance (FAD) (Kilgour et al., 2019) and the Kernel Inception Distance (KID) (Bińkowski et al., 2018) are often used as subjective measures of audio quality. For speech signals, the Fréchet Deep Speech Distance (Bińkowski et al., 2019) and Kernel Deep Speech Distance, have been proposed as metrics of audio quality.

While these metrics work well for estimating the perceived quality of a speech signal, the differences in speech and singing voice signals, listed in Section 1.1.4, make them incompatible for assessing the quality of a singing voice signal The mel-cepstral distance (MCD). has been used to provide an estimate of the quality of a singing voice signal (Blaauw & Bonada, 2016). Computation of the metric involves calculating the difference between the two signals, aligned in time, in the mel-cepstral domain. How-

ever, it has several limitations and cannot be considered an absolute criteria for evaluation.

3.3.2 Source separation evaluation

An ideal algorithm would be able to extract the exact signal that was used for the mixing procedure, without the presence of any of he other signals. However, most source separation algorithms introduce some form of artificial noise to the target signal and might also have some interference from the other signals in the mixture. These components can be quantified for the extracted signal, given the ground truth version of the individual signals present in the mixture and can be used for evaluation of source separation algorithms.

More importantly however, it is important to evaluate the perceived **quality** and **isolation** of the extracted signal. For singing voice or speech signal separation, **intelligibility** of the separated signal is also an important perceptual criteria. For these criteria, listening tests such as MUSHRA (Schoeffler et al., 2015) and MOS tests have been proposed. Such tests ask expert listeners to listen to different audio samples and rate them on a fixed scale based on various criteria.

However, such tests are quite subjective as the testing conditions and the exact stimuli provided for different models might not be exactly the same. Also, the conducting such listening tests at scale can be quite expensive, especially. since expert listeners are required to meet the requirements for the test. Therefore, it is also important to have objective evaluation metrics. One such set of metrics that is commonly used throughout the source separation community and indeed in the SiSEC campaign is the Blind Source Separation Evaluation (BSS Eval) (Vincent et al., 2006). The BSS eval performance metrics includes three sets of metrics; *bss_eval_sources*, pertaining to single-channel source signals, *bss_eval_images*, for multichannel spatial source signals and *bss_eval_mix* for mixing filters. As mention in Section 1.1.5, this thesis focuses on monoaural source separation for the voice signal and thus uses the *bss_eval_sources*

set of evaluation metrics.

For these evaluation metrics, a mixture of signals, $\mathbf{m}$ is assumed to be a linear sum of sources, $\mathbf{s}_i$; $\mathbf{m} = \sum_{i=1}^{K} \mathbf{s}_i$, where i is the index of the K sources. Additionally, some noise might be present in the signal. The idea behind the evaluation methodology is to decompose the error between the estimated source and the target source into three components, the filtering and spatial errors, the artifacts or the artificial noise present in the estimated signal and interference from the other sources present in the estimated signal. Within this evaluation methodology, the estimated source, $\tilde{\mathbf{s}}_i$, is first decomposed into four constituents as:

$$\tilde{\mathbf{s}}_i = \mathbf{s}_i^{target} + e_i^{interf} + e_i^{noise} + e_i^{artif} \tag{3.1}$$

$\mathbf{s}_i^{target}$ is a modified version of $\mathbf{s}_i$, which may contain certain allowed distortions, $\mathcal{F}$. e_i^{interf} represents interference coming from unwanted sources $(\mathbf{s}_{i'})_{i' \neq i}$ which might be mixed along with $\tilde{\mathbf{s}}_i$. e_i^{noise} represents noises such as forbidden distortions, not in the set $\mathcal{F}$ and e_i^{artif} represents burbling and other artifacts that might be introduced by the process of source separation.

Using these values, the following evaluation measures can be computed:

1. Source to Distortion Ratio (SDR):

$$SDR_i := 10log_{10} \frac{\|\mathbf{s}_i^{target}\|^2}{\|e_i^{interf} + e_i^{noise} + e_i^{artif}\|^2} \tag{3.2}$$

   This measure represents the overall performance of the source separation algorithm for the source indexed by i.

2. Source to Interferences Ratio (SIR):

$$SIR_i := 10log_{10} \frac{\|\mathbf{s}_i^{target}\|^2}{\|e_i^{interf}\|^2} \tag{3.3}$$

   This metric measures the amount of interference in the estimation of the source indexed by i from other sources present in the mixture.

3. Signal to Noise Ratio (SNR):

$$SNR_i := 10log_{10}\frac{\|\mathbf{s}_i^{target} + e_i^{interf}\|^2}{\|e_i^{noise}\|^2} \tag{3.4}$$

4. Sources to Artifacts Ratio (SAR):

$$SAR_i := 10log_{10}\frac{\|\mathbf{s}_i^{target} + e_i^{interf} + e_i^{noise}\|^2}{\|e_i^{artif}\|^2} \tag{3.5}$$

   This measure provides an estimate of artifacts in the estimated sourced, introduced by the source separation process.

A perceptually motivated adaptation of the BSS metrics, called the Perceptual Evaluation method for Audio Source Separation (PEASS) (Vincent, 2012; Emiya et al., 2011) has also been proposed recently. like the BSS metrics, this set of metrics extracts three distortion components from the signal to be evaluated. The saliance of these distortion components is computed by the PEMO-Q (Huber & Kollmeier, 2006) metric for audio quality assessment and these are then passsed through a neural network to compute three metrics; Target-related Perceptual Score (TPS), Artifacts related Perceptual Score (APS), Interference-related Perceptual Score (IPS), and Overall Perceptual Score (OPS). These metrics have shown higher correlation to the MOS obtained from MUSHRA tests. However, recent studies have shown that neither the BSS nor the PEASS set of metrics show strong correlation with the subjective rating provided by human listening tests and as such do not generalize well for different algorithms (Cano et al., 2016; Ward et al., 2018).

3.4 Summary

Our research focuses on source separation in two distinct domains, contemporary popular music and ensemble choral singing. As such we require two kinds of datasets for our study. The first kind is popular music datasets with the clean and processed singing voice stems, instrumental backing track and annotations pertaining to melodic and linguistic content, in the form of phonetic annotations. We note that the MedleyDB dataset

(Bittner et al., 2014) fulfills the first of these requirements, but does not have phonetic annotations. We also note that the number of tracks containing the singing voice in this dataset is quite low to efficiently train a deep learning based algorithm, but it can be used for evaluation. The iKala dataset (Chan et al., 2015) is another dataset which meets our requirement for singing voice without processing effects present along with the backing track. The NUS corpus (Duan et al., 2013) presents singing voice data with phonetic annotations but does not have an accompanying backing track. The amount of data present in this corpus is also limited.

We attempted by record a dataset to meet our requirements, but personal and global conditions imposed constraints on the author which prevented the fulfillment of this dataset. Fortunately, a proprietary dataset was provided to us for the purpose of training our models. This dataset consists of 12 hours of data, with 205 songs by 45 distinct male and female professional singers. The voice is presented without any effects and an instrumental backing track is included for each of the songs. The songs in this dataset are in the English and Japanese language.

For choral singing, we leverage some of the recently recorded datasets, including the Choral Singing Dataset (CSD) (Cuesta et al., 2018), the Dagstuhl ChoirSet (Rosenzweig et al., 2020), the ESMUC dataset and the Bach Chorales Dataset (BCD).

We also study evaluation methodologies for voice synthesis and source separation algorithms and note that while a number of objective metrics have been proposed for evaluation, subjective listening tests are the most practical methodology available for evaluation.

Throughout this thesis, we will use subjective tests following the AB format. For most of our experiments, $15-20$ participants from a limited demographic were asked to evaluate. We used the BeaqleJS (Kraft & Zölzer, 2014) JavaScript-based framework for subjective audio quality evaluations, with functionality as shown in Figure 3.1.

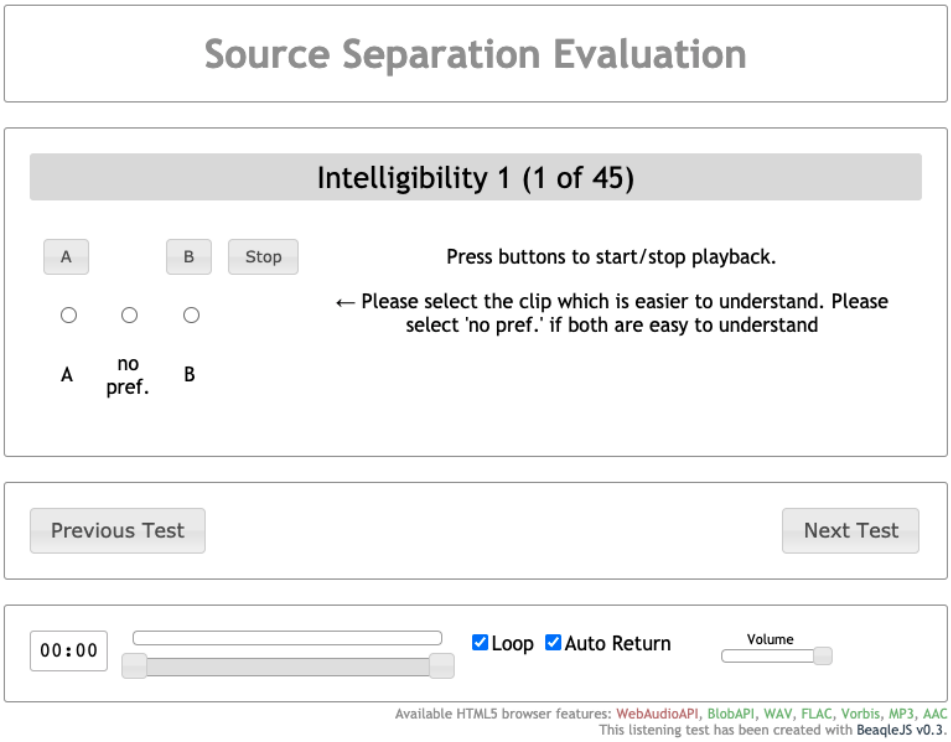

Figure 3.1: An example of an AB test using the BeaqleJS framework (Kraft & Zölzer, 2014).

Part II

Synthesis Applied To Source Separation

List of symbols

a A representation a general signal.

A Spectogram of the signal denoted by **a**.

c A representation of a general signal, different from **a**.

C Spectrogram of the signal denoted by **c**.

s The time domain waveform of an arbitrary source mixed in a musical mixture.

S Spectrogram of the source denoted by **s**.

x Time-domain waveform of voice signal, could be speech or singing.

X Spectrogram of the voice signal denoted by **x**.

$\mathbf{X}_{\mathbf{voc}}$ Compressed spectral envelope pertaining the voice signal denoted by **x**.

$\mathbf{X}_{\mathbf{mel}}$ Mel-scale spectrogram pertaining to the voice signal denoted by **x**.

y Time-domain waveform of voice signal with modulations added.

Y Spectrogram of the voice signal with modulations added, **y**.

$\hat{\mathbf{x}}$ Time-domain waveform of an output voice signal.

$\hat{\mathbf{X}}$ Spectrogram of the output voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{X}}_{\mathbf{voc}}$ Compressed spectral envelope pertaining to the voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{X}}_{\mathbf{mel}}$ Mel-scale spectrogram pertaining to the voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{y}}$ Time-domain waveform of an output voice signal, which has effects and modulations added.

$\hat{\mathbf{Y}}$ Spectrogram of the output voice signal with modulations added, $\hat{\mathbf{y}}$.

b Time-domain waveform of musical instrumental backing track.

B Spectogram of musical instrumental backing track

m The mixture signal formed by mixing **y** with **b**, the mix does not necessarily have to be a linear mixture.

M Spectogram of the mixture signal denoted by **m**.

enc The encoder network of an autoencoder.

dec The decoder network of an autoencoder.

V The latent embedding of an autoencoder.

gen The generator network of a GAN.

dis The discriminator network of a GAN.

Z The linguistic content of the voice signal, **x**.

η The melodic content of the voice signal, **x**.

ψ A representation of a singer or speaker, who is the source of **x**.

ω A soft-mask or Wiener filter used for source separation.

Chapter 4

Introduction

In Chapter 2, we observed that most musical source separation algorithms proposed for separating the singing voice from a musical mixture assume that the mixture is a linear sum of the individual sources. With this assumption, most algorithms use time-frequency (TF) masks to filter the singing voice from the mixture, typically applied to the magnitude component of the spectrogram of the mixture signal.

However, most contemporary popular music involves the application of processing effects like flanger, phasor, reverb or delay, which modify the voice signal. In addition, the singer might employ techniques like growling or screaming, which deviate from the typical voice production mechanism, to augment the content of the lyrics in a meaningful way. Contemporary music also utilizes a non-linear mixing process, along with mastering and post-production techniques, which violate the leaner sum assumption of source separation algorithms. Deep learning based music source separation algorithms, discussed in Section 2.3 have shown remarkable robustness to such vocal effects and mixing conditions. However, the TF mask based source separation algorithms have a limitation in that they can only filter out the processed form of the signal, which might be undesirable in some cases.

To overcome this limitation, we propose the framework for a system to synthesize a clean voice signal from a mixture signal, based on the underlying content. The pro-

posed methodology is based on a human listeners' process while trying to to replicate the singing voice signal in a musical song that he or she is listening to by parsing the linguistic and melodic content of the song and singing the same. We note that the non-linear processing and mixing applied to the voice signal does not alter its perceptual qualities and a human listener is able to discern the content despite these effects. Based on this, we propose a methodology to synthesize the clean singing voice signal from a musical mixture using the underlying perceptual content. In Chapter 5, we propose a methodology to extracted synthesis parameters directly from a musical mixture. In Chapter 6, we use a feedforward network optimized using adversarial training to map the linguistic and melodic content derived from a score, along with a representation of the singer identity to synthesis parameters. Chapter 7 presents a methodology to extract linguistic content and the singer identity from a mixture signal and use these to generate the synthesis parameters from which a clean singing voice signal can be generated.

We note that synthesis has previously been applied to the problem of separating the voice from a musical mixture. Sinusoidal models (Maher, 1989) for synthesis of the singing voice from a mixture after estimating the fundamental frequency of the signal and segregation using heuristics was amongst the first models proposed for singing voice separation. Similar approaches based on the fundamental frequency, combined with a frequency-locked loop algorithm and harmonically constrained trackers (Wang, 1994, 1995) have been proposed. Information from scores has been used for synthesis based separation approaches (Meron & Hirose, 1998), combined with sinusoidal models of speech information representation (Quatieri & McAulay, 1992) to separate the voice signal from a piano accompaniment. Other models include using a filter based on a harmonic model of the voice (Ben-Shalom & Dubnov, 2004), peak clustering and harmonic re-synthesis (Duan et al., 2008), synthesis using sinusoidal models of the voice (Mesaros et al., 2007; Fujihara et al., 2005, 2010). and spectral peak detection using with cross-correlation (Lagrange & Tzanetakis, 2007; Lagrange et al.,

2008) along with quadratic interpolation (Smith & Serra, 1987) and phase generation (Slaney et al., 1994). Source-filter models of the human voice have been used to estimate Wiener masks for singing voice separation as well (Durrieu et al., 2010). A more recent model for singing voice separation using synthesis (Rao et al., 2014) uses adaptive sinusoidal components activation with predominant f0 estimation followed by peak picking (Griffin & Lim, 1988). The singing voice signal is then synthesized using a harmonic sinusoidal model (HSM) with linear interpolation of amplitudes and cubic phase interpolation.

Chapter 5

Synthesis parameter estimation

Deep learning based methodologies have shown great potential to model the singing voice. Data-driven models have been shown to be robust enough to model TF masks for the singing voice even when the spectral structure of the singing voice is altered through effects like those discussed in Section 1.2.1. On the other hand, generative algorithms to synthesize the singing voice from an input context have also produced astounding results (Blaauw & Bonada, 2016, 2017; Blaauw et al., 2019). In this chapter, we combine the modeling ability of deep neural networks with the synthesis methodology used for generating voice signals. We note that while established evaluation metrics exist for TF mask based source separation, evaluation of singing voice synthesis is typically done through subjective listening tests.

We propose a methodology to estimate synthesis parameters from a musical mixture signal using a deep neural network as a function approximator. The methodology we propose is closest to one of the oldest models for singing voice separation, (Miller, 1973), which used a homomorphic vocoder (Oppenheim & Schafer, 1968) to synthesize the singing voice in a musical mixture signal after segmentation of parts based on heuristics and cepstral liftering to account for the accompaniment. The homomorphic vocoder used by Miller is one of several acoustic features that have been proposed to represent the synthesis parameters of the voice signal. Recently, a similar methodology

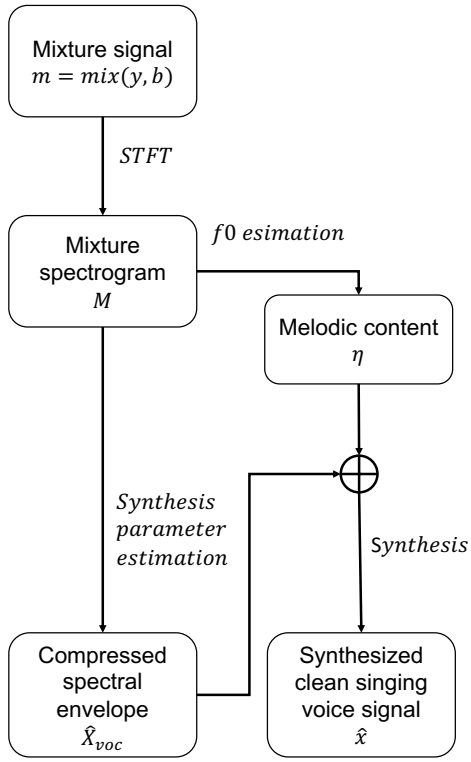

Figure 5.1: The framework for the proposed model. We use a non-autoregressive variant of the WaveNet (van den Oord et al., 2016a) architecture to estimate the compressed spectral envelope synthesis parameters as well as the f0

has recently been proposed for speech denoising (Maiti & Mandel, 2019).

The research presented in this chapter aims to answer the following questions:

- Is it possible to extract synthesis parameters pertaining to the singing voice from a polyphonic contemporary music mixture?

- How can the voice signal extracted using such a methodology be evaluated?

5.1 Synthesis parameters

For our case, we propose the use of the WORLD (Morise et al., 2016) vocoder, which was originally proposed for real-time synthesis of speech signals and has since been successfully adapted for singing voice synthesis (Blaauw & Bonada, 2017, 2016). The vocoding algorithm uses an f0 estimation known as *DIO* (Morise et al., 2009) to model the fundamental frequency of the voice signal. This information is used to estimate the *spectral envelope*, also known as the *harmonic* component of the signal, using the *CheapTrick* (Morise, 2015b) algorithm. Finally, the *aperiodic* component of the speech signal is estimated using the Definitive Decomposition Derived Dirt-Cheap (D4C) (Morise, 2016) algorithm. The harmonic and aperiodic components are distributed over 1024 bins for each time frame of the signal analyzed.

For synthesis, shown in Figure 2.17, the aperiodic element is used as the excitation signal that is convolved with the minimum phase response of the spectral envelope to estimate the vocal cord vibrations in the vocal signal.

As such, the vocoder allows for modelling the f0 independently, within limits, of the harmonic and aperiodic content of the speech signal which is desirable for singing voice synthesis. For estimation via neural networks, the dimensions of the harmonic and aperiodic components are usually reduced; the NPSS model uses 60 features for the harmonic component and 4 for the aperiodic component. We also use these 64 features in our methodology and refer to the combination as the **compressed spectral envelope**, $\mathbf{X_{voc}}$.

5.2 Parameter estimation

We use a temporal convolutional neural network (TCN) (Lea et al., 2016) based on the WaveNet (van den Oord et al., 2016a) architecture for synthesis parameter estimation. Our architecture is not autoregressive. Such an architecture has been used for speech de-noising (Rethage et al., 2018). We use dilated convolutions with a gated activation,

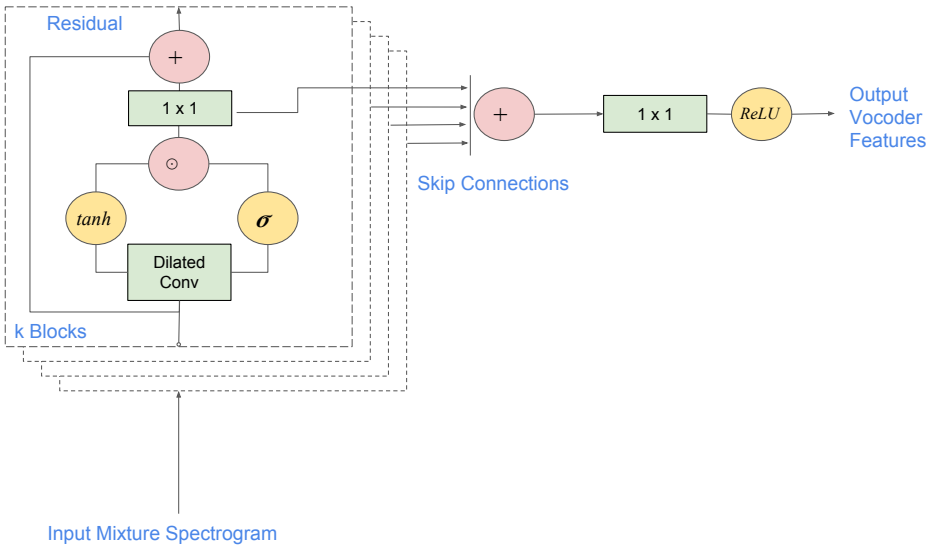

Figure 5.2: The convolutional block used in our model

as used by (Van Den Oord et al., 2016; van den Oord et al., 2016b). However, we do not enforce causality in the convolutional layers, but use zero-padding to ensure to ensure dimensional consistency in the time dimension between consecutive layers. The individual blocks of the architecture, as shown in Figure 5.2, also borrow from the WaveNet. The input to our model is the spectrogram of a linear mixture of an instrumental backing track and a clean singing voice signal and the output is the 64 dimension compressed spectral envelope, $\mathbf{X_{voc}}$, as described in Section 5.1. We treat the frequency bins of the spectrogram as different channels, like (Blaauw & Bonada, 2017), and thus each convolutional layers consist of 1-D convolutions across the time dimension.

As shown in Figure 5.2, we use k_{ss} blocks of convolutional layers with skip and residual connections and gated dilated convolutions. For training, we use N_{ss} consecutive frames of the mixture spectrogram as input to the network. This leads to an input of dimensions $N_{ss} \times D$, where D is the number of bins in the spectrogram. The first layer of the network are 1×1 convolutional layer, the output of which is $N_{ss} \times D_1$. This is followed by a series of gated stacks of 2×1 dilated convolutions (denoted by $*$), each

with D_1 units and a sigmoid (denoted by σ) non-linearity. A similar operation is carried out with a tanh non-linearity and an element-wise multiplication (denoted by $\odot$) to apply a gated non-linearity (van den Oord et al., 2016a; Dauphin et al., 2017). This operation is mathematically represented by Equation 5.1.

$$out = tanh(W_{\tau,k} * in) \odot \sigma(W_{\rho,k} * in) \tag{5.1}$$

Where W denotes a convolution filter and τ and ρ represent filter and gate layers, respectively. The input and output of each layer are represented by *in* and *out*, respectively. To increase the receptive filed of each block, the dilation factor is increased exponentially by 2 after each block. We apply two 1×1 convolutional layers, after the series of stacked convolutions. This ensures that the output of the network, $\hat{\mathbf{X}}_{\mathbf{voc}}$, has the same dimensions as the target compressed spectral envelope, $\mathbf{X}_{\mathbf{voc}}$. We use the MAE between the output of the network and the target as the loss function for the network, as shown in Equation 5.2 and optimize it using an ADAM optimizer (Kingma & Ba, 2014).

$$\mathcal{L}_{SS} = \mathbb{E}[\|\hat{\mathbf{X}}_{\mathbf{voc}} - \mathbf{X}_{\mathbf{voc}}\|] \tag{5.2}$$

5.3 Fundamental frequency estimation

We use a separate model for f0 estimation from the polyphonic mixture, which follows a similar architecture. We used a continuous representation of the f0, expressed in the logarithmic formula for MIDI notation, shown in Equation 5.3

$$\eta_{MIDI} = 12 \cdot log_2 \frac{\eta_{hertz} - 69}{440} \tag{5.3}$$

Where η_{hertz} is the f0 value in Hertz. The value was normalized to the range 0-1, using *min-max* normalization across the dataset. Heuristically, we found synthesis quality to

be better with such a representation than the discrete representation used in other deep learning based models (Bittner et al., 2017; Doras et al., 2019; Jansson et al., 2019; Kim et al., 2018a) [11] The f0 for the voiced frames was interpolated across the unvoiced frames, similar to the work done by (Blaauw & Bonada, 2016, 2017). We also used a separate network to predict a binary voiced/unvoiced feature for each frame of the output, which was optimized with a binary classification loss. Like (Blaauw & Bonada, 2016), we used a chain of networks wherein the output of the vocoder parameter estimation network was fed along with the mixture spectrogram to the f0 estimation network and the output of these two networks was fed into the the voiced/unvoiced prediction network.

5.4 Experiments

5.4.1 Baseline models

While Deep Learning based algorithms have exploded in the source separation field over the last few years, there were few open source implementations available at the time we first proposed our methodology. As such, we used a benchmark based on the NMF methodology, named the **Flexible Audio Source Separation Toolbox** (FASST) (Ozerov et al., 2012). In addition, we used a Deep Learning based methodology proposed by us, called the **DeepConvSep** (Chandna, 2016) algorithm as a Deep Learning based benchmark to compare our proposed methodology against. For evaluation, we term our proposed methodology as separation-via-synthesis, (**SS**).

We compare the f0 estimation model against the **Melodia** (Salamon et al., 2013b; Salamon & Gómez, 2012) algorithm for predominant melody estimation.

[11] Research on f0 estimation from polyphonic signals for the purpose of synthesis is being carried out by a masters' student in the Universitat Pompeu Fabra, under the supervision of the author of this thesis.

5.4.2 Datasets

We use the iKala dataset (Chan et al., 2015) for evaluation of our proposed methodology, as it contains vocal tracks without external effects like reverb or compression. We use a subset of 226 songs from the iKala dataset for training the proposed model and another subset of 15 for validation and 11 for testing.

For pre-processing the tracks, we compute a Short Time Fourier Transform (STFT), with an FFT-size of 1024, leading to $D = 513$ frequency bins. A hop time of 5 miliseconds is used for this calculation as well as for estimating the WORLD parameters that are used as the target. All features are normalized by using min-max Normalization, constraining the input and output features to the range 0 to 1.

5.4.3 Analysis and network hyperarameters

We trained the network for $50k$ iterations, using minibatch training with a batch size of 30 batches per iteration. Each batch contained $N_{ss} = 128$ consecutive time frames, randomly sampled from the training set. The networks had $k_{ss} = 5$ blocks of gated convolutions and $C = 128$ filter channels for each of the convolutional layers except the final layer. All input and output features were normalized to the range 0 to 1 using *min-max* normalization.

5.4.4 Evaluation methodology

As discussed in Section 3.3.1, evaluating a synthesized voice signal is not trivial. For our algorithm there are three aspects to evaluation; **intelligibility**, **separation from backing track** and **audio quality**. The **Source to Interferences Ratio** (SIR) metric from the BSS Eval set of metrics (Vincent et al., 2006) provides an objective estimate of the degree of interference from the backing track present in the output signal. Additionally, we use the **mel-cepstral distortion** (MCD) as a measure of the quality of the synthesized audio compared to the ground truth vocal track. Since the vocoder used for synthesis introduces degradation to the output quality, we use a version of vocal signal

re synthesized by the WORLD vocoder as an upper limit reference. This allows use to evaluate the estimation of the network.

Intelligibility of the synthesized signal is a subjective matter and depends on external factors such as familiarity of the listener with the song and the language used. As such, we use a comparative AB preference listening test, as described in Section 3.3.1 for evaluation of intelligibility as well as separation/interference and audio quality. In our experiments, we paired each of the three systems to be compared, leading to 3 pairs, with 5 samples per criteria, resulting in a total of 45 preference questions. The listener was presented with a clean vocal signal reference for questions pertaining to the audio quality criteria, whereas mixture audio was provided as reference for interference related questions. The participant was asked to choose the example which had less interference from the backing track in the later case. For the questions related to intelligibility, the listener was asked to choose the system which was more easily understandable. In this case, the reference audio might have caused a bias and was hence omitted from the question. We used 5 s samples from songs in the test set, not used for training the model. The online listening test, was presented in the Mandarin Chinese language.

For evaluation of the f0 model, we used the raw pitch accuracy (RPA), voicing false alarm (VFA) and overall accuracy (OA) (Bittner & Bosch, 2019) metrics from the `mir_eval` library (Raffel et al., 2014). We used the MIDI note annotations provided in the iKala dataset as reference.

5.4.5 Results

The results of the f0 evaluation are shown in Table 5.1. It can be seen that our proposed model slightly outperforms the knowledge based baseline (Salamon & Gómez, 2012; Salamon et al., 2013b) in terms of RPA and OA. The improved performance might be due to dataset bias, since our algorithm was tested on the same dataset that it was trained on. We also note that the VFA metric is lower for the Melodia algorithm than our pro-

posed network, this suggests that a higher number of frames were mis-classified by the proposed network for voiced/unvoiced classification. This is classification undesirable for synthesis in our proposed methodology. Further investigation into f0 estimation for the singing voice in polyphonic musical signals is currently being carried out by as Masters' student in the Universitat Pompeu Fabra under the supervision of the author of this thesis (Stillings, 2021).

Model	RPA	VFA	OA
SS	$85.5 \pm 5.0\%$	$30.5 \pm 17.8\%$	$82.2 \pm 8.2\%$
Melodia (Salamon et al., 2013b)	$80.0 \pm 12.9\%$	$27.6 \pm 16.9\%$	$74.3 \pm 15.90\%$

Table 5.1: The evaluation metrics for pitch accuracy comparing the proposed methodology, SS, with the Melodia algorithm (Salamon et al., 2013b) for predominant melody estimation. The values shown are the mean $\pm$ standard deviation.

The SIR metric for the three models compared is shown in Figure 5.3. It can be seen that the proposed model, termed SS outperforms the NMF based model, FASST, and the Deep Learning mask based model, DeepConvSep. This can be attributed to the synthesis methodology used, which only generates the vocal signal and not the accompanying track, which causes interference in the other two models.

The SDR mteric is shown in Figure 5.5, while the SAR is shown in Figure 5.6. We observe that the proposed model, SS, fall behind both the DeepConvSep and FASST algorithms in these objective metrics. We believe this is possibly due to variability in the synthesized version of the signal which, while perceptually similar to the ground truth signal, is not exactly the same signal. Mask based algorithms perform better on these metrics as they are estimations of the same sginal that was used for creating the mix. The MCD metric is shown in Figure 5.4 and shows similarity in the objective quality of the vocal signal extracted by all three models under consideration.

For the subjective listening test, we received responses from 16 participants, all of whom were native Chinese Mandarin language speakers. The results of this test are shown in Figure 5.7. We can observe that a clear preference is given to the pro-

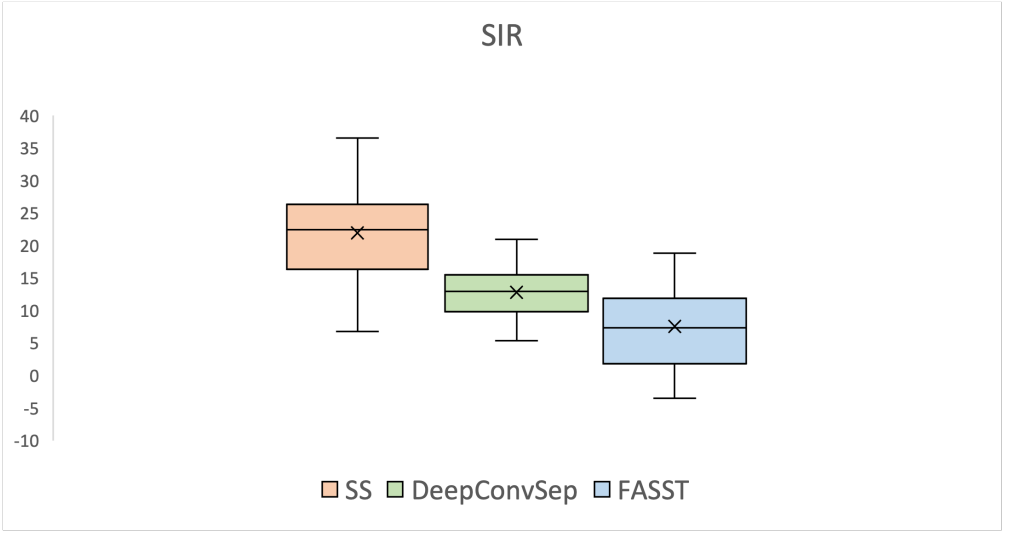

Figure 5.3: The SIR metric from the BSS Eval toolkit for the three systems to be compared. It can be observed that the proposed model, SS, achieves a higher score in this metric than the other two systems. This is expected since the use of voice specific vocoder features in our system prevents interference from other sources in the output.

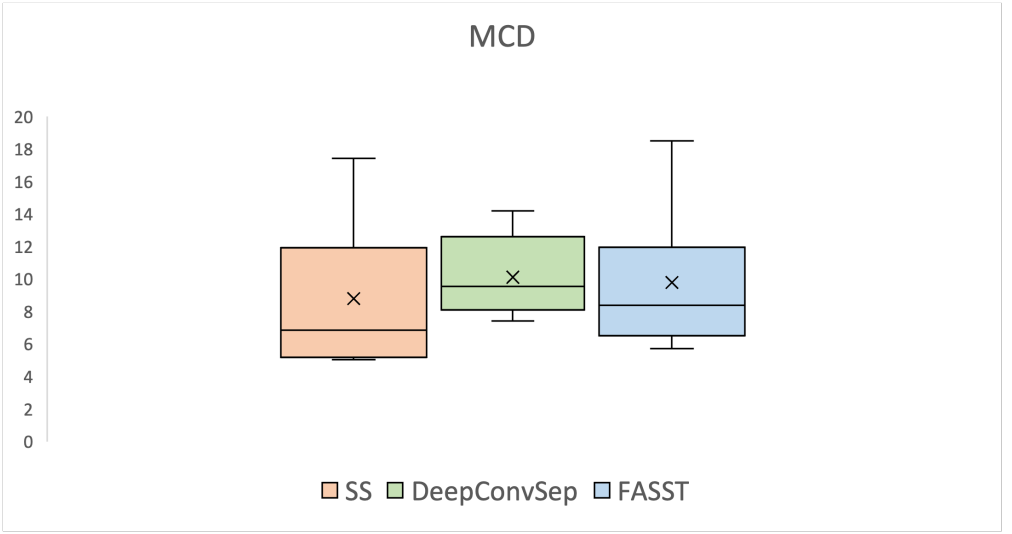

Figure 5.4: The Mel Cepstral Distortion (MCD), in dB, comparing the proposed model, SS with the separated singing voice signal from the DeepConvSep (Chandna et al., 2017) and the FASST (Ozerov et al., 2012) source separation algorithms.

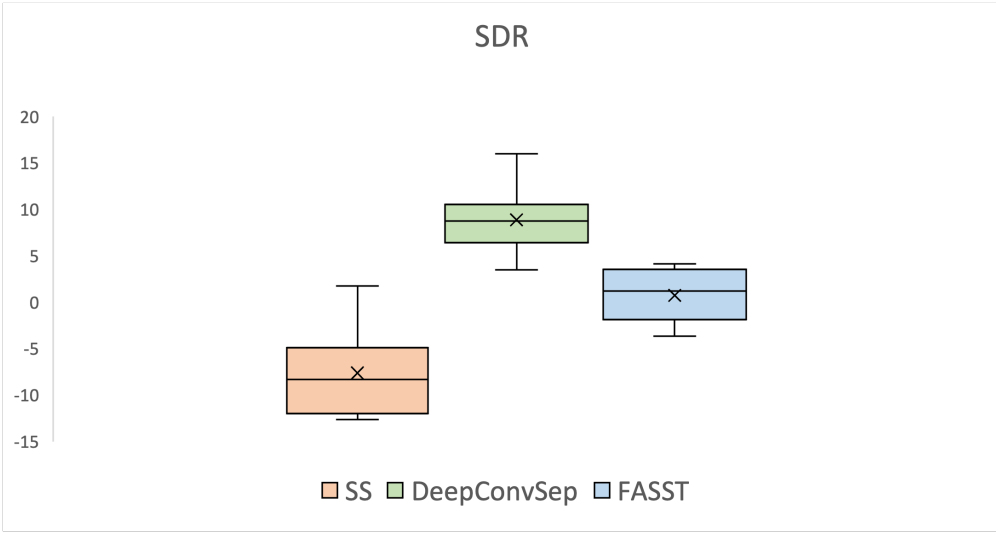

Figure 5.5: The SDR metric from the BSS Eval toolkit comparing the proposed model, SS with the separated singing voice signal from the DeepConvSep (Chandna et al., 2017) and the FASST (Ozerov et al., 2012) source separation algorithms.

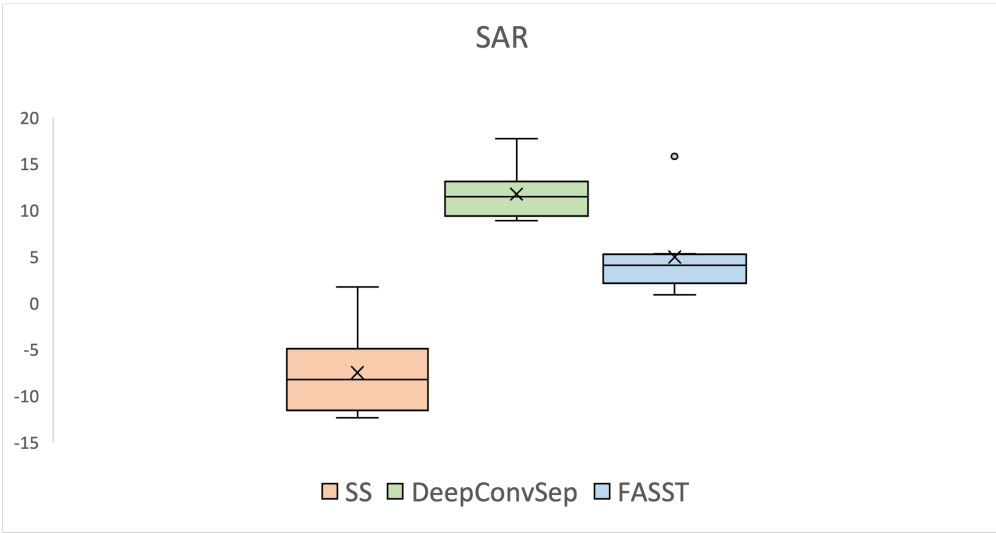

Figure 5.6: The SAR metric from the BSS Eval toolkit comparing the proposed model, SS with the separated singing voice signal from the DeepConvSep (Chandna et al., 2017) and the FASST (Ozerov et al., 2012) source separation algorithms.

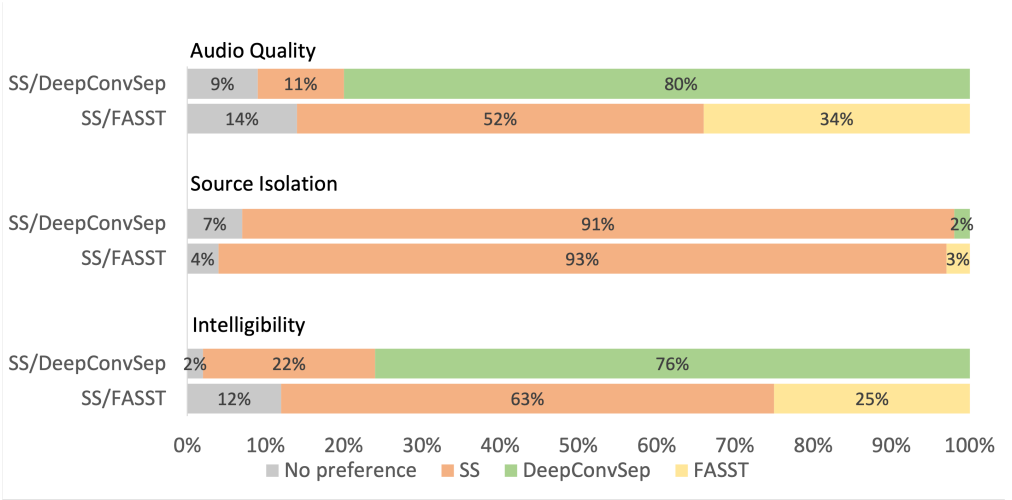

Figure 5.7: Results of the listening test comparing the proposed model, SS with the separated singing voice signal from the DeepConvSep (Chandna et al., 2017) and the FASST (Ozerov et al., 2012) source separation algorithms.

posed methodology over the NMF based methodology, particularly for the intelligibility and interference criterion, corroborating the objective results. DeepConvSep received higher preference in terms of intelligibility and audio quality over the proposed methodology, but SS was preferred over it in terms of interference from the accompaniment. The objective evaluation via the SDR and SAR metrics, shown in Figures 5.5 and 5.6, respectively, do not completely agree with these findings.

It can be seen that while DeepConvSep is preferred over our proposed model for the intelligibility and audio quality criterion, SS is perceived to perform better in terms of interference, by a majority of the people participating in the listening test.

5.5 Conclusions

We show that using deep neural networks as function approximators, we can directly estimate the synthesis parameters for the singing voice from a polyphonic popular music mixture, containing vocals. To this end, we use a non-autoregressive version of the WaveNet architecture, with skip and residual connections for effective information

propagation. The input to the proposed network is the magnitude component of the spectrogram of the musical mixture. The network is trained to predict the WORLD vocoder features, compressed via dimensionality reduction techniques and the MAE loss is used for this optimization. We use a separate network for prediction of the f0 of the signal, which is expressed in a continuous MIDI note format, as well as for prediction of the voiced and unvoiced nature of a frame within the signal. We note that this network performs comparably to one of the best knowledge based methodologies for predominant pitch estimation, while taking dataset bias into account. We are now in a position to answer the research questions presented in the introduction to this chapter:

- Is it possible to extract synthesis parameters pertaining to the singing voice from a polyphonic contemporary music mixture?

  We see that using deep neural networks, it is possible to map the magnitude component of the linear spectrogram of musical mixture signal to the corresponding synthesis parameters. While there is on going research on various synthesis parameters, in our case, we use the compressed spectral envelope of the WORLD vocoder features.

- How can the voice signal extracted using such a methodology be evaluated?

  We note that the objective evaluation metrics typically used for source separation (Vincent et al., 2006) do not agree with the subjective evaluation done via listening tests. We acknowledge that there are other metrics for evaluation of for both source separation and synthesis, as listed out in Section 3.3.2. However, the reliability of such metrics is still under debate (Cano et al., 2016). For the rest of this part of the thesis, we will use listening tests similar to those used in this chapter for subjective evaluation of the proposed synthesis methodologies.

The singing voice signal generated using the proposed methodology is completely free of interference from the backing track. This is one of the biggest problems observed

with the filtering masked based approach, commonly used for source separation. This is observed through both subjective and objective comparison with both a knowledge based source separation algorithm and a deep learning based source separation algorithm.

However, there is notable degradation in the quality of the synthesized signal, possibly due to the nature of the vocoder parameters used for synthesis. We note that the proposed methodology provides a framework wherein the individual components namely the architecture used for prediction, the f0 estimation and the vocoder parameters used for synthesis can be replaced with improved versions as research in the field evolves.

Chapter 6

Synthesis parameter generation

Singing voice synthesis (SVS) systems have become popular over the last decade, particularly with commercial applications like the Vocaloid (Kenmochi & Ohshita, 2007) and Melodyne. SVS systems take as input a score which provides linguistic, melodic and rhythmic information. From this information, an SVS system must generate a signal that follows the musical guidelines provided by the score. Additionally, the SVS system must also sound natural and in doing so, emulate the timbre and pitch inflections pertaining to a particular target singer. A singer following a score would generally not sing the exact same way twice, owing to natural timing and pitch deviations. As pitch deviations like overshoot, preparation and fine fluctuations, discussed in Section 1.1.3 are natural and at times involuntary. Vibrato is introduced voluntarily by a singer for artistic effect, but the phase of the vibrato might change amongst two different takes of a song. On the other hand, the timbre of the singer is generally consistent through multiple takes and across songs. As such, modelling pitch and timbre are two different tasks within the field of singing voice synthesis. Methodologies have been proposed to generate an expressive pitch contour emulating a specific singer singing a given score (Bonada & Blaauw, 2020). On the other hand, models like the NPSS (Blaauw & Bonada, 2016) take f0 curve and linguistic features as input to generate the singing voice signal modelling the timbre of a target singer and the linguistic content presented as the input. Such timbre models are pertinent to the topic covered in this thesis as we aim

to synthesize a signal that retains linguistic and melodic content of the singing voice signal in a polyphonic mixture whilst modelling the timbre of multiple singers. The melodic content is generally modelled as the f0 curve, which can be extracted from the polyphonic mixture, while the linguistic content is represented by a sequence of phonemes. In this chapter, we present a model for multi-singer singing voice synthesis that can generate a natural singing voice signal emulating the timbre of multiple singers. The framework for this model is shown in Figure 6.1

In this chapter, we present a methodology for multi-singer singing voice synthesis using a feedforward network. The linguistic content, the melody and the singer identity are provided to the network as input and it generates the compressed spectral envelope as the output. We train the network via an adversarial training methodology (Arjovsky et al., 2017). This chapter tries to answer the following research question:

- How can a feedforward neural network be used for singing voice synthesis given an input of linguistic content, singer identity and the f0 curve?

6.1 Generative networks for voice synthesis

Deep Learning based generative networks like the autoregressive WaveNet model (van den Oord et al., 2016a), variational autoencoders, normalizing flows and generative adversarial networks (GANs) have opened up new avenues for voice synthesis. Several Text-To-Speech (TTS) and Singing Voice Synthesis (SVS) methodologies, utilizing such generative models, have been proposed over the last few years. Most of these models use large volumes of data; the WaveNet (van den Oord et al., 2016a) model was trained using the English multi-speaker corpus from CSTR voice cloning toolkit (VCTK) dataset (Veaux et al., 2017) which consists of 44 hours of speech recordings from 109 speakers, while the Deep Voice models also used the LibriSpeech dataset (Panayotov et al., 2015), which has around 820 hours of data available. As noted in

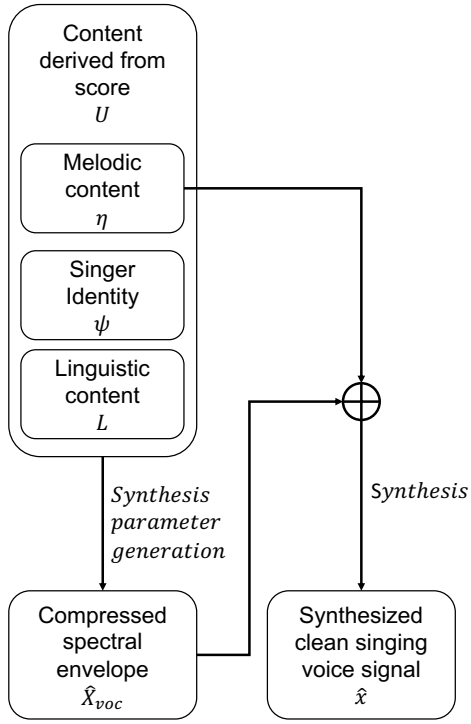

Figure 6.1: The framework for the singing voice synthesizer we propose. The proposed model is used for acoustic parameter generation of the compressed spectral envelope from the melodic content, linguistic content and the singer identity.

Chapter 3, such large volumes of data are not easy to obtain for the singing voice, especially with the linguistic annotations required for synthesis. The NUS corpus (Duan et al., 2013) that we use in this study has a total of 169 minutes of recordings, which is very little compared to the amount of data used for training TTS systems. To overcome the lack of data, we use data augmentation by using random sampling from the dataset and also use adversarial training in the form of the Wasserstein GAN to train a feedforward convolutional neural network to generate a singing voice signal given an f0 contour and linguistic annotations.

The Wasserstein GAN (WGAN) (Arjovsky et al., 2017) is an adaptation of the Gen-

erative Adversarial Network (GAN) (Goodfellow et al., 2014) framework for training neural networks to model data distributions to allow for generation. The GAN is a optimization methodology for training generative networks involving two neural networks, a generator, $gen()$ and a discriminator, $dis()$. As suggested by the name, the generator model takes as input either randomly sampled noise or certain constraints and conditions for the target output data and generates samples that try to match the target distribution under the constraints. In the case of the voice signal, such conditions could be the linguistic features and speaker identity. The second network, known as the discriminator, estimates the probability of an input sample either coming from the real target distribution or having been generated by the generator network. In other words, the discriminator tries to identify the generated samples from real samples while the generator tries to fool the discriminator and the two networks play an **adversarial min-max game**. This is represented by the loss function shown in Equation 6.1

$$\begin{aligned} \mathcal{L}_{GAN} = \min_{gen} \max_{dis} & \mathbb{E}_{\mathbf{a}\sim P_{\mathbf{a}}} \| log(dis(\mathbf{a})) \| \\ & + \mathbb{E}_{\mathbf{c}\sim P_{\mathbf{c}}} \| \log(1 - dis(gen(\mathbf{c}))) \| \end{aligned} \tag{6.1}$$

Where $\mathbf{a}$ is a sample from the real distribution and $\mathbf{c}$ is the input to the generator, which may be noise or conditioning as in the Conditional GAN (Mirza & Osindero, 2014) and is taken from a distribution of such inputs, $P_{\mathbf{c}}$. It has been shown that under optimal training conditions, when the two networks reach a *Nash equilibrium*, the loss function of the GAN represents the Jensen–Shannon Divergence between the generated data distribution and the real target data distribution. However, training a GAN can be quite difficult, as insufficient support for a lower dimension manifold of the data can lead to instability as well as other problems like vanishing gradient and mode collapse.

To alleviate such problems, the use of the Earth-Mover distance or the Wasserstein distance along with gradient clipping has been suggested as an effective alternative. The loss function for the WGAN is shown in equation 6.2. In this version of the GAN, the discriminator network is replaced by a network termed as critic, also represented by

$dis()$, which can be trained to optimality and does not saturate, converging to a linear function. Adversarial losses are often complemented with guidance losses like the MAE or the MSE, especially when used for conditional synthesis like in the WaveGAN (Donahue et al., 2019) and GANSynth (Engel et al., 2019) models.

$$\mathcal{L}_{WGAN} = \min_{gen} \max_{dec} \mathbb{E}_{\mathbf{a}\sim P_{\mathbf{a}}} \|dis(\mathbf{a})\| - \mathbb{E}_{\mathbf{c}\sim P_{\mathbf{c}}} \|dis(gen(\mathbf{c}))\| \tag{6.2}$$

Such methodologies have also been applied to TTS synthesis, particularly with recurrent architectures (Zhao et al., 2018; Yang et al., 2017; Ma et al., 2019). GANs have also been used as post-filters to overcome oversmoothing effects present in acoustic synthesis models (Kaneko et al., 2017b,a). Reasearch has also shown that the GAN methodology can be applied for singing voice synthesis (Hono et al., 2019), by modelling the inter-feature dependencies with each from of the output of the model.

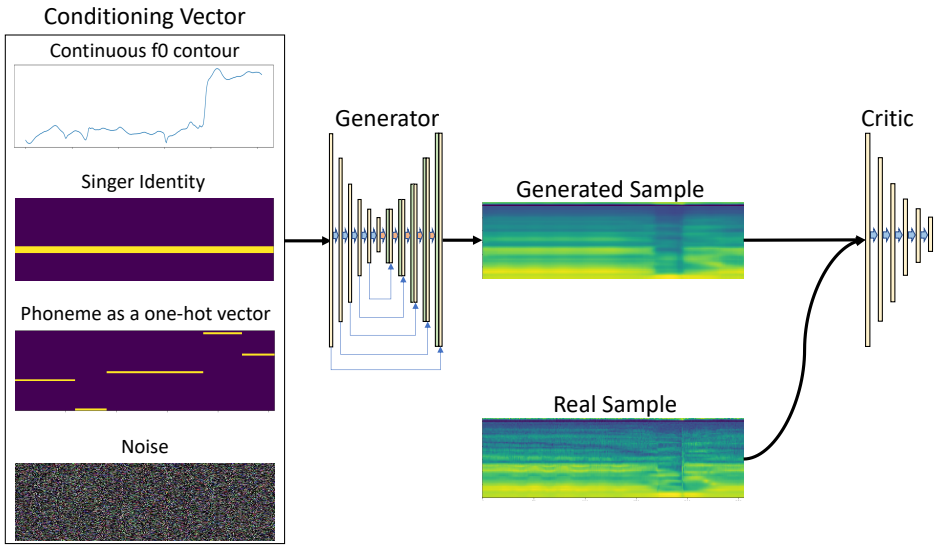

Figure 6.2: The conditioning vector for the generator and critic networks of our proposed model. A conditioning vector, consisting of frame-wise phoneme and $f0$ annotations along with speaker identity is passed to the generator. The critic is trained to distinguish between the generated sample and a real sample.

6.2 Proposed model for singing voice synthesis

We propose a convolutional architecture, similar to the DCGAN (Radford et al., 2016), using an encoder-decoder schema as shown in Figure 6.3. Both the encoder and the decoder in this architecture have $k_w gan$ layers with a filter size of 3. Skip connections between corresponding layers of the encoder and decoder are implemented by concatenating the encoder layer with the decoder layer (Ronneberger et al., 2015). Strided convolution is used for downsampling in the encoder (Radford et al., 2016) and the decoder uses linear interpolation followed by convolution for downsampling. This has been shown to reduce high frequency artifacts in the output that can be caused by transposed convolutions (Stoller et al., 2018). ReLU activation is used in all the layers of the network except the output layer, which uses a tanh activation that is commonly used in adversarial networks.

To model a voice signal, $\mathbf{x}$, the input conditioning to our system consists of frame-wise phoneme annotations, $\mathbf{Z}_{phone}$, represented as a one-hot vector and continuous fundamental frequency extracted by the spectral autocorrelation (SAC) algorithm, η_{f0} . This conditioning is similar to the one used in NPSS (Blaauw & Bonada, 2016). In addition, we condition the system on the singer identity, as a one-hot vector, ψ_{onehot}, broadcast throughout the time dimension. This approach is similar to that used in the WaveNet (van den Oord et al., 2016a). The three conditioning vectors are then passed through a 1×1 convolution and concatenated together along with noise sampled from a uniform distribution and passed to the generator as input. For simplicity, this concatenated input, shown in Figure 6.2, is represented by U.

For the target of the model, we use the WORLD vocoder (Morise et al., 2016) for acoustic modelling of the singing voice. We apply dimensionality reduction to the vocoder features, as described in Section 5.1, resulting in a 64 dimension compressed spectral envelope, referred to as $\mathbf{X_{voc}}$.

To optimize the network, we compliment the WGAN loss with a reconstruction MAE

loss, as shown in Equation 6.3. Such a loss is often used in conditional image generation models using the adversarial framework (Lee et al., 2019b).

$$\begin{aligned} \hat{\mathbf{X}}_{\mathbf{voc}} &= gen(U) \\ \mathcal{L}_{WGAN} = \min_{gen}\max_{dis} \mathbb{E}_{\mathbf{X}_{\mathbf{voc}}\sim P_{\mathbf{X}_{\mathbf{voc}}}} \|dis(\mathbf{X}_{\mathbf{voc}})\| - \mathbb{E}_{u\sim P_u}\|dis(\hat{\mathbf{X}}_{\mathbf{voc}})\| \\ \mathcal{L}_{recon} &= \min_{gen} \mathbb{E}_{u,\mathbf{X}_{\mathbf{voc}}} \|\hat{\mathbf{X}}_{\mathbf{voc}} - \mathbf{X}_{\mathbf{voc}}\| \\ \mathcal{L}_{total} &= \mathcal{L}_{WGAN} + \lambda_{recon}\mathcal{L}_{recon} \end{aligned} \tag{6.3}$$

Where λ_{recon} is the weight given to the reconstruction loss. Both the generator and networks are optimized following the Wasserstein GAN schema (Arjovsky et al., 2017). The critic for our system uses an architecture similar to the encoder part of the generator, but uses LeakyReLU activation instead of ReLU (Radford et al., 2016).

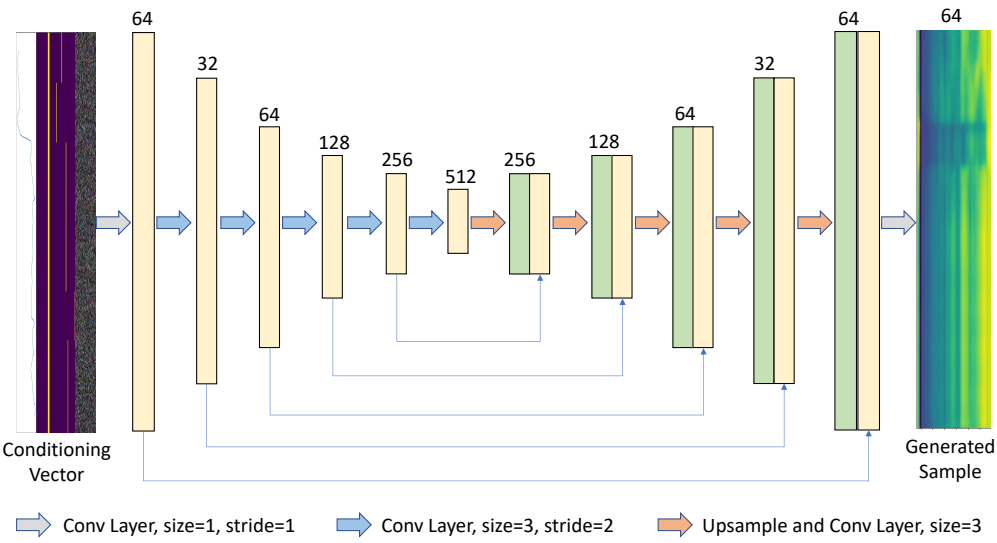

Figure 6.3: The architecture for the generator of the proposed network. The generator consists of an encoder and a decoder, based on the U-Net architecture (Ronneberger et al., 2015).

6.3 Experiments

6.3.1 Baseline models

We compare the proposed methodology for singing voice synthesis with the NPSS model (Blaauw & Bonada, 2016), described in Section 2.4.3. The input conditioning and the output synthesis parameters of the two systems are quite similar, thus allowing for a fair comparison.

6.3.2 Datasets

We use the NUS (Duan et al., 2013) dataset for training and evaluation of our proposed methodology. As mentioned in Section 3.1, this dataset contains of 48 popular English songs both sung and spoken by 12 non-native English speakers. Since the size of the dataset is quite small, as compared to datasets typically used for TTS and SVS synthesis, we use all but two songs, for training. The two songs held out for evaluation are Song 05 by a male singer, JLEE and Song 04 by a female singer, MCUR.

Along with the NPSS, we use a re-synthesis with the WORLD vocoder as the baseline as this is the upper limit of the performance of our system.

6.3.3 Analysis and network parameters

A hoptime of 5 ms was used for extracting the vocoder features and the conditioning. We used a block size of $N = 128$ for training the network.

A weight of $\lambda_{recon} = 0.0005$ was used for $\mathcal{L}_{recon}$ and the network was trained for 3000 epochs. We used the RMSProp optimizer for network optimization, with a learning rate of 0.0001.

This is the optimizer recommended by the researchers who proposed the Wasserstein GAN methodology (Arjovsky et al., 2017).

6.3.4 Evaluation methodology

As in Chapter 5, we use the mel-cepstral distortion (MCD) metric to give an indication of the quality of the synthesized singing voice signal. This metric is presented in table 6.1. We used a comparative online AB test for subjective evaluation. The participants in the listening test were presented two 5 s to 7 s phrases from the songs[12]. The participants were asked to select the preferred example in terms of Intelligibility and Audio Quality. We compared 3 pairs for this evaluation:

- WGANSing - Original song re-synthesized with WORLD vocoder.
- WGANSing - NPSS
- WGANSing, original singer - WGANSing, sample with different singer.

The participants were presented with 5 questions for each of the pairs, for each criteria. This lead to a total of 15 questions per criteria, with 30 questions overall. For the synthesis with a changed singer, we included samples with both singers of the same gender as the original singer and of a different gender, from within the dataset. To account for the natural differences in the ranges, the f0 input to the system was adjusted by an octave.

6.3.5 Results

We received responses from 27 participants from 10 nationalities for the listening test. Most of the participants were from native English speaking countries like England and the USA and the ages varied from 18 to 37. The results of the tests are shown in Figure 6.4.

[12] We found that WGANSing without the reconstruction loss as a guide did not produce very pleasant results and did not include this in the evaluation. However, examples for the same can be heard at https://pc2752.github.io/sing_synth_examples/

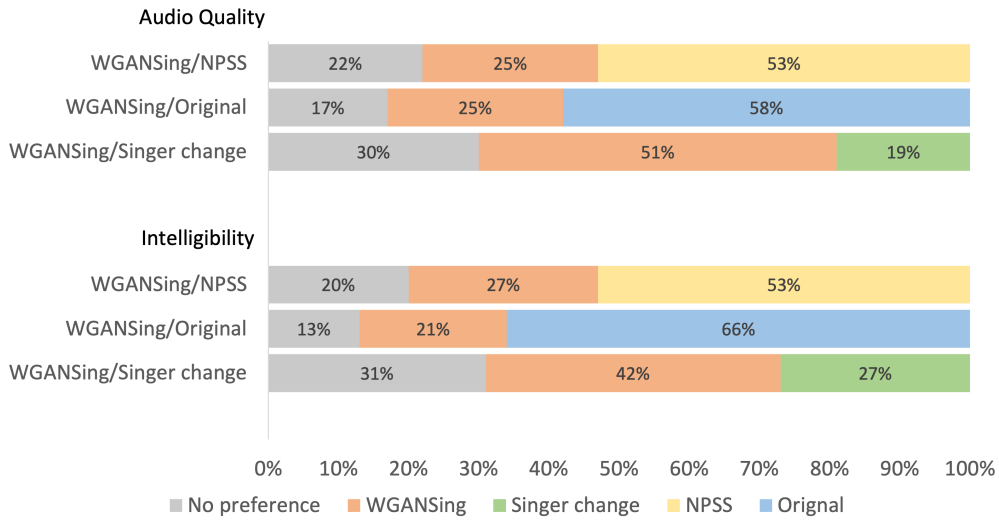

Figure 6.4: Results of the listening test comparing the proposed models, WGANSing with the NPSS (Blaauw & Bonada, 2016), the re-synthesized original and synthesis with the singer changed.

We observe from the results that while a preference is given to the NPSS, our proposed system is comparable to it in terms of both intelligibility and audio quality. Nearly half of the participants showed either no preference or preference towards our proposed methodology over NPSS. This result is in accordance with the objective evaluation shown in Table 6.1. It can be seen that both quality and intelligibility of the synthesis is compromised in comparison with the re-synthesized ground truth, which is an expected compromise in this case. We also observe a slight preference towards the synthesis of the songs with the original singer over the synthesis with a change in singer.

The subjective nature of the listening test as well as the diversity of the participants explains the variability in the observed results to an extent. However, we believe that while the quality is note quite state-of-the-art, it is still acceptable for a synthesis system trained with a small dataset.

Song	WGAN + $\mathcal{L}_{recon}$	WGAN	NPSS
Song 1 JLEE 05	5.36 dB	9.70 dB	5.62 dB
Song 2 MCUR 04	5.40 dB	9.63 dB	5.79 dB

Table 6.1: The MCD metric for the two songs used for validation of the model. The three models compared are the NPSS(Blaauw & Bonada, 2017) and the WGANsing model with and without the reconstruction loss.

6.4 Conclusions

We propose a system for feedforward voice synthesis of the singing voice signal for multiple singers. We used a U-Net based architecture to map the phoneme based annotation of the linguistic content $\mathbf{Z}_{phone}$, the f0 representation of the melodic content,η_{f0} and a one-hot representation of singer identity, ψ_{onehot} to the WORLD vocoder features, which were used for synthesis of the corresponding singing voice signal. The WORLD vocoder features were compressed using the dimensionality reduction techniques described in Chapter 5.

We use a U-Net based architecture, with connections between the corresponding layers of the encoder and decoder to generate the vocoder features given the contextual input. The network was trained on a small corpus of annotated singing data using the Wasserstein GAN methodology, which is an adversarial methodology that alleviates some of the issues faced while training GANs.

We used subjective and objective evaluation to compare the proposed system to a SOTA autoregressive singing voice synthesizer, the NPSS, and found the two systems to be comparable in performance. We also note that the use of a complementary MAE reconstruction loss improved the performance of the proposed system over just using the Wasserstein GAN loss. As such, the Wasserstein GAN loss can be seen as a complimentary loss to the reconstruction which guides the model for generation of the vocoder parameters used for synthesis. We are now in a position to answer the research question presented in the introduction to the chapter:

- How can a feedforward neural network be used for singing voice synthesis given an input of linguistic content, singer identity and the f0 curve?

  We believe that adversarial training is one of the possible methodologies for training a feedforward network for singing voice synthesis. We note that since the proposal of this methodology, several other deep learning based singing voice systems (Blaauw et al., 2019; Lee et al., 2019a; Hono et al., 2019; Ogawa & Morise, 2021) as well as music audio generation algorithms have been proposed (Dieleman et al., 2018; Zukowski & Carr, 2018; Engel et al., 2020b; Défossez et al., 2018). Several such models use an auxiliary loss along with the MAE or MSE reconstruction loss. We believe the Wasserstein GAN loss used in this chapter can be viewed as such an auxiliary loss, which is used to guide the model to model high frequency features of the compressed spectral envelope.

Chapter 7

Generation of synthesis parameters from content representations

The perceived content of a singing voice signal in a contemporary polyphonic mixture remains unchanged despite the use of spectral alterations and mixing process. To leverage this, we propose a system for synthesizing the singing voice signal, $\mathbf{x}$ from a polyphonic mixture, $\mathbf{m}$ by extracting the underlying linguistic $\mathbf{Z_x}$ and melodic content $\eta_\mathbf{x}$. We hypothesize that the linguistic and melodic content of a voice signal is retained even after it is processed through effects as described in Section 1.2.1 and is mixed using non-linear mixing processes with an instrumental accompaniment, ie $\mathbf{Z_x} = \mathbf{Z_y} = \mathbf{Z_m} = \mathbf{Z}$ and $P_X = P_{\underline{X}} = P_M = \eta$. We investigate how this content can be extracted from a polyphonic mixture. The melodic content can be extracted through f0 estimation[13]. Estimating the linguistic content for the purpose of synthesis however is a challenging task that we address in this chapter. Concretely, in this chapter, we will address the following research questions:

- Can the linguistic content of a singing voice signal be represented in a language independent manner from which a voice signal can be synthesized?

[13]Research on extracting f0 of a vocal signal from a polyphonic musical mixture is carried out by a masters' student, Logan Stillings, under the supervision of the author of this thesis and results of the finds are excluded from this thesis.

- Is it possible to extract such a representation of the linguistic content from a polyphonic contemporary music mixture?

- How can we derive a representation of the singer identity for the voice synthesis process?

Our initial experiments used explicit phonetic annotations with phonemes represented as one-hot vectors. We used deep learning based methodologies for this purpose including networks designed by us and other state-of-the-art methodologies (Demirel et al., 2020) for singing voice transcription to extract phoneme annotations from a polyphonic mixture. However, we found that accurately estimating the ground truth phonemes from a polyphonic mixture had some limitations; data-driven algorithms for such a task require a large volume of data to train a Deep Learning based methodology and the best performing polyphonic singing lyrics transcription methodologies do not have sufficiently high accuracy to allow for the methodology we propose. Additionally, the language of the data used for training imposes a constraint on the system used for synthesis. To overcome these limitations, we propose a system based on abstract representations of linguistic features often used in zero-resource synthesis (Glass, 2012; Jansen et al., 2013) and voice conversion (Mohammadi & Kain, 2017), as discussed in Section 2.5.1. We investigated the capability of voice conversion systems proposed for speech conversion to represent the linguistic content present in a singing voice signal and used one such state-of-the-art algorithm proposed for zero-shot voice conversion as a part our methodology. The framework we propose is shown in Figure 7.1. While we use the WORLD vocoder parameters for synthesis and for disentangling melodic information from the rest of the signal and AutoVC (Qian et al., 2019) voice conversion model for disentangling linguistic information, we believe the framework in itself is agnostic to these models.

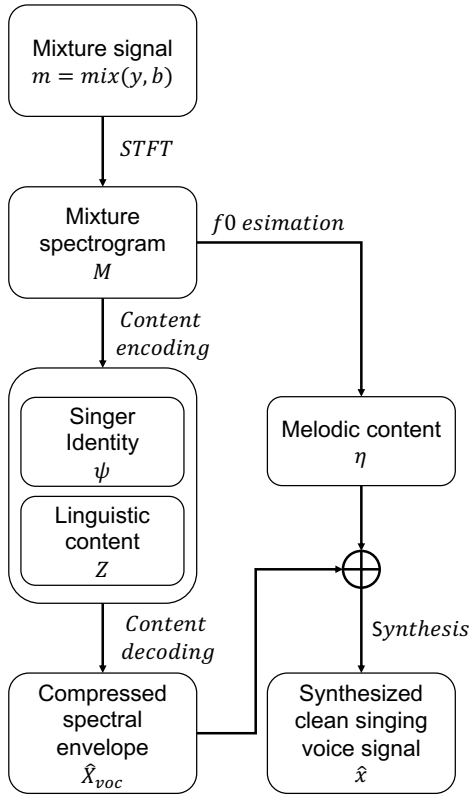

Figure 7.1: The proposed framework for synthesizing a clean singing voice signal from a mixture signal using the underlying perceptual content.

7.1 Representing linguistic content in a voice signal

A voice signal is generated by recording a person singing or speaking. In the case of speech, the signal possess certain characteristics such as timbre and prosody that allow the speaker to be identified as the *source* of the signal. Voice conversion is a field of audio processing that involves applying transformations to the voice signal so that the perceived source of the signal is changed to a target speaker (Mohammadi & Kain, 2017). It is important that the linguistic content of the signal is preserved through this transformation while the prosody and timbre of the signal is modified. These two characteristics can be represented by the fundamental frequency f0 and the spectral

envelope of the signal, respectively.

The first voice conversion algorithms proposed were trained to convert using *parallel* recordings, i.e. recordings from different speakers wherein the linguistic content was common amongst two pairs of recordings. *non-parallel* voice conversion algorithms were subsequently proposed, a common one using maximum likelihood constrained adaptation (Mouchtaris et al., 2004). As voice conversion requires the linguistic content to be retained, most algorithms proposed for this task require some estimation of linguistic features and most early algorithms required annotations of text transcriptions for conversion. Text independent voice conversion was first introduced in 2004 by (Ney et al., 2004), using a linear transformation of the spectral envelope of the signals. With the emergence of data modelling techniques using deep learning, several methodologies have been proposed separating the linguistic content of a voice signal from other properties of the signal, in a process commonly termed as **disentanglement**. Variational autoencoders have been used to learn a latent encoding to this end (Hsu et al., 2016; Huang et al., 2018), often including adversarial training (Hsu et al., 2017). Auxiliary classifiers have also been adapted for the voice conversion task (Kameoka et al., 2018a; Chou et al., 2018), as have Generative Adversarial Networks (GANS), with models like the StarGAN (Kameoka et al., 2018b; Kaneko et al., 2019b; Wang et al., 2020b) and the CycleGAN (Kaneko & Kameoka, 2017, 2018; Kaneko et al., 2019a; Fang et al., 2018).

Zero-shot and one-shot voice conversion has also been achieved using autoencoder inspired methodologies like the AutoVC (Qian et al., 2019) and the VQVC (Wu et al., 2020; Wu & Lee, 2020). The AutoVC model imposes an information bottleneck on the latent layer of an autoencoder while the VQVC and the VQVC+ model use a quantized latent code. As described in Section 2.5.1, both the AutoVC and the VQVC models use neural vocoders conditioned on mel-scale spectrogram representations of the speech signals. In doing so, both models effectively disentangle linguistic content and prosody from speaker identity in a speech signal.

While such methodologies work well for speech voice conversion, where prosody as well as timbre need to be converted, singing voice conversion requires the melodic information of the signal to be retained while modifying the timbre. To achieve this, the f0 needs to be disentangled from the spectral envelope. This can be achieved through a vocoder system like the WORLD (Morise et al., 2016) system used throughout this thesis.

7.2 Modifications to the AutoVC architecture

We adapt the AutoVC (Qian et al., 2019) architecture to take compressed spectral envelope, $\mathbf{X_{voc}}$, as defined in Section 5.1 as input. This modification is shown as shown in Figure 7.2. The encoder, $enc_{autovca}()$, encodes the linguistic information present in the input as $\mathbf{Z}_{autovca}$, while the decoder, $dec_{autovca}()$, takes this representation along with a singer representation, ψ to regenerate the compressed spectral envelope, $\hat{\mathbf{X}}_{\mathbf{voc}}^{autovca}$.

Heuristically, we found that changing the sampling rate of the information bottleneck to 16 instead of 32 led to better results when using the compressed spectral envelope as the input and target. The vocoder parameters pertaining to the compressed spectral envelope can directly be used for re-synthesis of the singing voice signal without the need for the WaveNet vocoder.

We tried various representations of the singer identity, including the one-hot vector encoding, the GE2E loss (Wan et al., 2018) based encoding that were used in the original AutoVC (Qian et al., 2019) and the Joint Embedding (JE) proposed for singer representation in both monophonic and polyphonic contexts (Lee & Nam, 2019). The methodology for deriving these speaker/singer representation embeddings is described in Section 2.5.3. In addition, we adapted the VQVC+ (Wu et al., 2020) model to take the compressed spectral envelope as input and target features.

We compared the adapted models to a SOTA singing voice conversion algorithm (Nachmani & Wolf, 2019) using subjective listening tests. The evaluation of the methodologies for

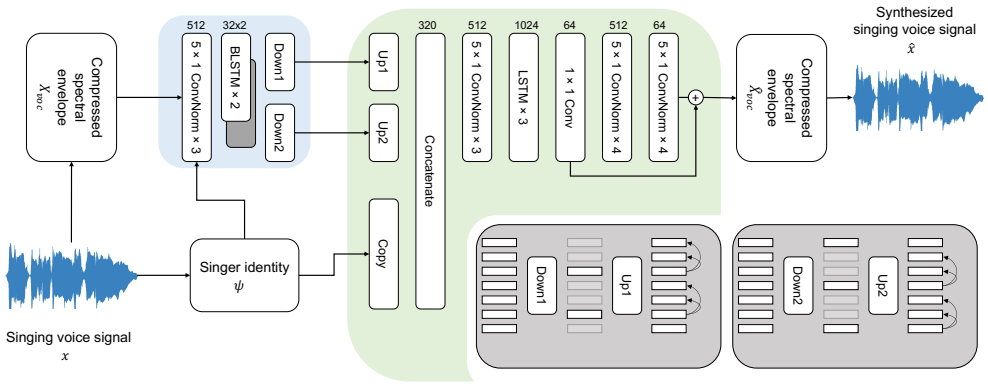

Figure 7.2: The proposed modifications to the AutoVC (Qian et al., 2019) architecture.

singing voice conversion is excluded from this thesis as it was carried out by a masters' student under the guidance of the author and is published separately (Apisov, 2020). We provide a brief summary of the results in Appendix C [14]. We observed that the adapted AutoVC models trained with GE2E and one-hot encoding singer representations out-performed the other models evaluated. As such, we decided to use the AutoVC model with one-hot singer representation for the rest of our research.

7.3 Deriving linguistic content from a polyphonic mixture

We see that the linguistic content in a clean singing voice signal can be represented by using abstract representations as those used in voice conversion. We propose a methodology to extract these representations from a polyphonic mixture signal by training a neural network using the adapted AutoVC models as a teacher network.

We use the AutoVC architecture, adapted for singing voice conversion, to train a *Singer Dependent Network* (SDN). The network has an encoder-decoder architecture similar to the adapted AutoVC, shown in Figure 7.2. The encoder part of this network, $enc_{ling}()$ takes as input the magnitude component of the spectrogram of the polyphonic mixture signal, $\mathbf{M}$ and is trained to replicate the linguistic content representation, $\mathbf{Z}_{autovca}$, as

[14] We note that a similar study using an architecture very similar to the one proposed was also conducted independelty by the research group at iZotope (Nercessian, 2020).

$\mathbf{Z}_{SDN}$. The decoder of the network, $d_{SDN}()$, generates the compressed spectral envelope, $\hat{\mathbf{X}}^{SDN}_{\mathbf{voc}}$, given the linguistic content $\mathbf{Z}_{SDN}$ and a one-hot representation of the singer identity, ψ_{onehot}, as shown in Equation 7.1. It should be noted that the encoder, $enc_{ling}()$, is not provided any representation of the singer identity and is thus singer independent, while the decoder $dec_{SDN}()$ reconstructs the signal in a singer dependent manner.

$$\begin{aligned} \mathbf{V}_{SDN} &= enc_{ling}(|\mathbf{M}|) \\ \mathbf{Z}_{SDN} &= downsample(\mathbf{V}_{SDN}) \\ \tilde{\mathbf{V}}_{SDN} = upsample(\mathbf{Z}_{SDN}) \hat{X}^{SDN}_{voc} &= d_{SDN}(\tilde{\mathbf{V}}_{SDN}, \psi_{onehot}) \end{aligned} \tag{7.1}$$

The network is trained to replicate the content embedding using a loss replication loss $\mathcal{L}^{SDN}_{replicate}$, reconstruct the vocoder features using a reconstruction loss, $\mathcal{L}^{SDN}_{recon}$ and maintain content consistency using a content loss, $\mathcal{L}^{SDN}_{content}$. The weighted sum of these losses results in the final loss of the network, $\mathcal{L}^{SDN}_{final}$, as shown in Equation 7.2. Used as such, the content loss acts as a complimentary loss to reconstruction loss, performing a function similar to the Wasserstein GAN loss in the Chapter 6

$$\begin{aligned} \mathcal{L}^{SDN}_{recon} &= \mathbb{E}[\|\hat{\mathbf{X}}^{SDN}_{\mathbf{voc}} - \mathbf{X}_{\mathbf{voc}}\|^2] \\ \mathcal{L}^{SDN}_{content} &= \mathbb{E}[\|\mathbf{Z}_{SDN} - downsample(enc_{ling}(\hat{\mathbf{X}}^{SDN}_{\mathbf{voc}}))\|] \\ \mathbf{Z}_{autovca} &= enc_{autovca}(\mathbf{X}_{\mathbf{voc}}) \\ \mathcal{L}^{SDN}_{replicate} &= \mathbb{E}[\|\mathbf{Z}_{SDN} - \mathbf{Z}_{autovca}\|] \\ \mathcal{L}^{SDN}_{final} &= \mathcal{L}^{SDN}_{recon} + \lambda_{SDN}\mathcal{L}^{SDN}_{content} + \mu_{SDN}\mathcal{L}^{SDN}_{replicate} \end{aligned} \tag{7.2}$$

Where λ_{SDN} and μ_{SDN} represent the weights given to the replication and content losses, respectively.

We then train a *Singer Independent Network* (SIN), which shares the linguistic encoder, $enc_{ling}()$ with the SDN. This leads to a shared linguistic content, $\mathbf{Z}_{SIN} = \mathbf{Z}_{SDN} = enc_{ling}(\mathbf{M})$. The distinction between the networks is the use of a *singer encoder* net-

work, $enc_{singer}()$, to learn the singer identity, ψ_{SIN}, directly from the mixture input. This network has the same architecture as the encoders of the SDN and the AutoVC models, thus ensuring that the embedding extracted is of the same size as the linguistic content embedding. The decoder of the network, $dec_{SIN}()$, takes as input the linguistic content, $\mathbf{Z}_{SIN}$ and the singer representation, ψ_{SIN}, to generate the corresponding compressed spectral envelope, $\hat{\mathbf{X}}^{SIN}_{\mathbf{voc}}$. This is shown in Equation 7.3. Since we provide the linguistic content, $\mathbf{Z}_{SIN}$, to the decoder of the autoencoder, we hypothesize that the bottleneck restriction will force the learned latent embedding to represent information pertaining to singer identity, ψ_{SIN}, given the input mixture spectrogram.

$$\begin{aligned} \psi_{SIN} &= downsample(enc_{singer}(|\mathbf{M}|)) \\ \mathbf{Z}_{SIN} &= \mathbf{Z}_{SDN} = enc_{ling}(|\mathbf{M}|) \\ \tilde{\mathbf{V}}_{SIN} &= upsample(\mathbf{Z}_{SIN}) \\ \hat{\mathbf{X}}^{SIN}_{\mathbf{voc}} &= dec_{SIN}(\tilde{\mathbf{V}}_{SIN}, upsample(\psi_{SIN})) \end{aligned} \tag{7.3}$$

The network is trained using a reconstruction loss, $\mathcal{L}^{SIN}_{recon}$ and a content consistency loss, $\mathcal{L}^{SIN}_{content}$, with a weight λ_{SIN}, as shown in Equation 7.4.

$$\begin{aligned} \mathcal{L}^{SIN}_{recon} &= \mathbb{E}[\|\hat{\mathbf{X}}^{SIN}_{\mathbf{voc}} - \mathbf{X}_{\mathbf{voc}}\|^2] \\ \mathcal{L}^{SIN}_{content} &= \mathbb{E}[\|\mathbf{Z}_{SIN} - downsample(dec_{SDN}(\hat{\mathbf{X}}^{SIN}_{\mathbf{voc}}))\|] \\ \mathcal{L}^{SIN}_{final} &= \mathcal{L}^{SIN}_{recon} + \lambda_{SIN}\mathcal{L}^{SIN}_{content} \end{aligned} \tag{7.4}$$

In addition, we also trained decoder networks which take GE2E (Wan et al., 2018) and Joint Embeddings specifically proposed for singer identification (Lee & Nam, 2019), ψ_{JE}, along with an adaptation of the VQVC+ model to derive the linguistic content for synthesis from a polyphonic mixture signal, using the same principle. We found that the proposed SIN methodology outperformed the adapted VQVC+ model as well as the GE2E (Wan et al., 2018) and Joint Embeddings for singer identification (Lee & Nam, 2019).

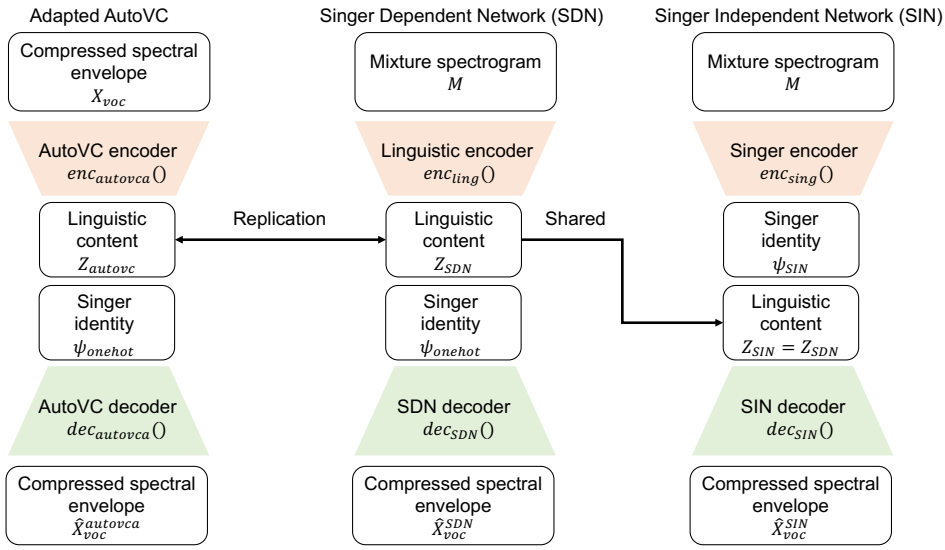

Figure 7.3: The framework for extracting linguistic content from a polyphonic mixture signal for synthesis of the singing voice.

For f0 estimation, we use the same model as proposed in Chapter 5.

7.4 Experiments

7.4.1 Baseline models

We compare our proposed model to the previously proposed methodology for source separation via synthesis, presented in Chapter 5, which we term as **SS**. We also compare our proposed model with the U-Net model (Jansson et al., 2017), which at the time was one of the best performing algorithms for source separation, particularly for the case of the singing voice.

7.4.2 Datasets

We use the proprietary dataset, described in Section 3.4, for training the model and the MedleyDB dataset (Bittner et al., 2014) for evaluation. The training set consists of 205 songs by 45 distinct male and female singers, with a total of around 12 hours of data. the songs are mostly pop songs in the English and Japanese languages. We use 90 %

of the proprietary dataset for training and 10 % for validation, which we use for early stopping during training of the model. We have access to the raw vocal track in this training set as well as annotation of the singers, which makes it ideal for our proposed model.

For testing, we use the the MedleyDB dataset, which contains 122 songs. The raw audio tracks and the mixing *stems* for which are present. We use the raw audio vocal tracks of 6 of the songs for computing the vocoder features, which are used to re-synthesise the singing track and are used as a reference during evaluation. To the best of our knowledge, there is no overlap amongst the singers in the training set and the singers present in MedleyDB. Therefore the use of this dataset for evaluation makes sense as we are using both songs and singers not seen by the model during training. For reconstructing the voice signal with the SDN network a singer of the same gender as the target singer, from the training dataset, was provided to the decoder of the SDN network, d_{SDN}.

7.4.3 Analysis and network parameters

We used a sampling rate of 32 kHz, with a Hanning window of size 1024 for the short time Fourier transform (STFT) of the mixture spectrogram. The vocoder features and the STFT were calculated with a hoptime of 5 ms. We use dimensionality reduction described in Chapter 5, leading to 64 synthesis parameters.

We use $\lambda_{SDN} = 1$ and $\mu_{SDN} = 1$, as in the original AutoVC (Qian et al., 2019).

7.4.4 Training

We use the Adam (Kingma & Ba, 2014) optimizer for training the various networks with batch size of 30. The training batch were randomly sampled from the tracks, with a length 640 ms. Variable gains were applied to the vocal and accompaniment tracks during the training to allow of data augmentation.

7.4.5 Evaluation

As in Chapter 5, we use mel-cepstral distortion as an objective measure of the quality of the synthesized voice signal and an AB preference test to subjectively evaluate the output on the criteria of **intelligibility**, **audio quality** and **isolation** from the backing track. The three models to be evaluated were grouped into pairs. As the listening tests are demanding to carry out, we chose to compare the SDN and SIN models with the U-Net model for the audio quality criteria, whereas the SDN, SIN and SS models were compared with the U-Net for source isolation. For intelligibility, we compared the SDN, SIN and the SS models.

7.4.6 Results

There were 25 participants, claiming proficiency in the English language, participated in our listening test, from various countries. 18 of the participants had previous musical training. The results of the listening test are shown in Figure 7.4.

We observe that all three proposed synthesis models outperform the mask based model, U-Net, on isolating the vocal signal from the mixture. This also shows the robustness of the models as the evaluation set has no overlap in terms of singers or songs with the training set.

We note that the SDN and SIN network perform better than the model previously proposed in Chapter 5, showing the effect of the added linguistic information extracted using the voice conversion algorithm. We can see that both the SDN is ranked higher on intelligibility than SS, thus showing that the content encoder is able to effectively extract the underlying linguistic features from the input mixture spectrogram. The SIN also outperforms the SS model showing that the network can even for singers not seen during training and for a mixed and processed vocal track. We observe that the SIN model outperforms the SDN model. This suggests that the singer encoder, e_{singer} learns more than just the singer identity from the input mixture spectrogram. However, for the purpose of source separation using synthesis, we believe this to be acceptable. Al-

Model	MCD (dB)
SS	7.39 ± 1.25
SDN	7.55 ± 0.63
SIN	6.45 ± 0.75
U-Net	6.58 ± 1.88

Table 7.1: The Mel Cepstral Distortion (MCD) metric in dB, comparing the proposed models, SDN and SIN with our previous model, SS, and the U-Net (Jansson et al., 2017)

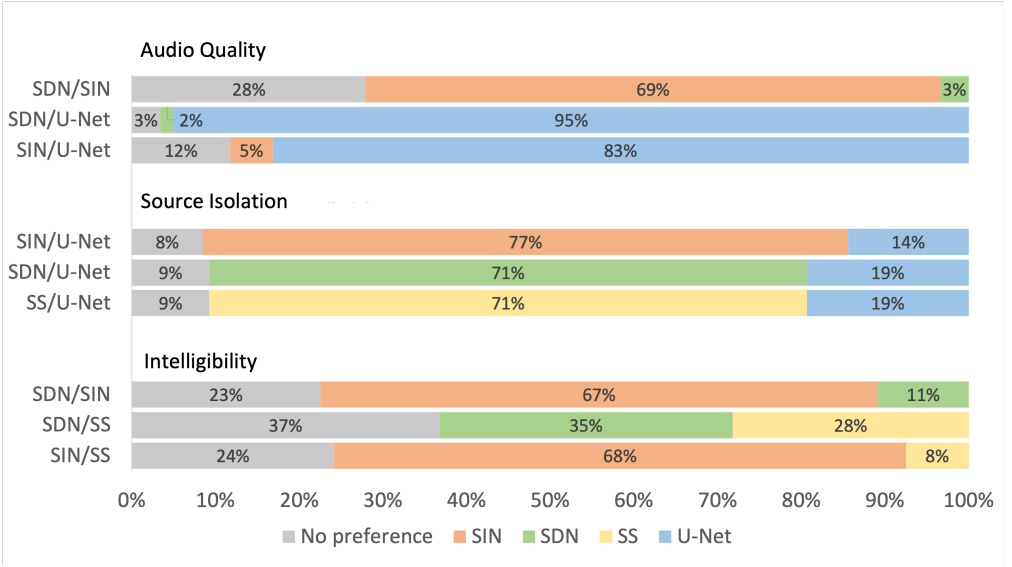

Figure 7.4: Results of the listening test comparing the proposed models, SDN and SIN with our previous model, SS, and the U-Net (Jansson et al., 2017).

though further tests on the singer representation need to be carried out in the future.

Audio quality for the proposed models still lags behind the mask based U-Net separation algorithm. This is partly due to the degradation introduced by the synthesis process, but can be improved by using more effective neural vocoder methodologies like the WaveNet vocoder (Shen et al., 2018).

7.5 Conclusions

We propose a methodology for singing voice separation using synthesis by extracting the underlying linguistic and melodic information from a polyphonic musical signal. We use an abstract representation of the linguistic content, based on voice conversion. This representation is used for voice conversion based on an autoencoder methodology. The speaker independent linguistic component of a voice signal is disentangled from the speaker dependent components like timbre and prosody, by imposing a bottlenecks constriction on the latent embedding of the autoencoder. While such models work well for voice conversion in the context of speech, they need to be modified to preserve the melody in the of the singing voice. To this end, we adapt voice conversion models for the singing voice using the WORLD vocoder (Morise et al., 2016) that also disentangles the effect of the f0 from the timbre of the voice. The evaluation of these models is published in a separate publication and in Appendix C.

We then use the modified voice conversion algorithm as a teacher network to train a network with a similar architecture to extract the linguistic content from a mixture spectrogram in a singer-independent manner, while synthesizing the voice signal in a singer-dependent manner and is termed SDN. We also propose a methodology based on the autoencoder bottleneck to directly extract the singer identity from the mixture spectrogram, as the speaker independent linguistic content is provided to the decoder. This model is capable of synthesizing the voice signal in a singer-independent manner is as termed SIN. We note that the network optimization in both networks is done via a reconstruction loss complimented by a content consistency loss. This loss functions similar to the Wasserstein GAN loss in Chapter 6, and follows the principle used in (Bińkowski et al., 2019), namely that intermediate layers of a deep neural network trained on audio inherently learn perceptually relevant features of the audio data. In this case, the perceptually relevant features pertain to the linguistic content of the singing voice signal, which is consistent in the input and output.

Through subjective evaluation, we observe that the proposed methodology improves over the results presented in Chapter 5. As the SDN network is able to reproduce the singing voice signal, we can confirm that the singer independent linguistic encoder of the network is able to encapsulate the linguistic information from the mixture spectrogram. However, we note that the singer identity representation learned by the SIN model might be encapsulating more information than just the singer identity, although this functions quite effectively for the purpose of source separation. We also note that there is still room for improvement, particularly concerning the synthesis of the final vocal signal. This can be improved by using more effective vocoder representations like those discussed in Section 2.4.2. For robustness to language and recording effects, we performed several tests with songs from various languages including Hindi, Catalan, Spanish, Latin and German along with popular English songs which have effects added. While we were not able to properly evaluate the output of the SDN ans SIN models using listening tests, we do provide examples on our supplementary website https://pc2752.github.io/sep_content/. We note that while the quality of the output can be improved, the synthesized signal even in languages not used for training is intelligible for native language listeners. We also note that the f0 estimation system needs further improvement, indeed for some songs with multiple vocal effects, the perceived pitch is indescribable (Nieto, 2013) and synthesizeing a clean vocal signal for such songs would require a f0 generation model for the synthesis methodology proposed. We are currently working on improving the f0 estimation methodology. The framework for our methodology, shown in Figure 7.1, has three main components, each of which can be replaced by newer components as research proceeds in the fields of voice representation via vocoders, linguistic representation and speaker/singer identity representation.

We are now. in a position to answer the research question presented in the introduction to this chapter:

- Can the linguistic content of a singing voice signal be represented in a language

independent manner from which a voice signal can be synthesized?

Zero resource synthesis and voice conversion are two fields of research which aim to disentangle the speaker independent linguistic content from the speaker dependent content of a speech signal. Deep learning based models from both these fields use unsupervised training with autoencoders. Constraints are applied to the latent embedding of the autoencoder to separate the linguistic content from the speaker dependent features of the signal. As the end goal of these models is to synthesize a voice signal, we believe the intermediate of the same can be used for singing voice synthesis.

- Is it possible to extract such a representation of the linguistic content from a polyphonic contemporary music mixture?

  We show that by adapting a voice conversion model for the singing voice, we can use it as a teacher network to train another network to extract the linguistic content from a mixture signal. In our first experiment, we use a one-hot encoding of the singer identity along with the extracted linguistic features to generate the compressed spectral envelope for synthesis.

- How can we derive a representation of the singer identity for the voice synthesis process?

  We test a number of methodologies for singer identity representation and we also propose our own methodology for the same. The proposed methodology is based on the assumptions used in both the VQVC (Wu et al., 2020; Wu & Lee, 2020) models and the AutoVC (Qian et al., 2019) model for voice conversion. While our proposed methodology works well for the task at hand, we believe that it extracts more information than just the singer identity from the input mixture signal.

Part III

Source Separation For Ensemble Singing

List of symbols

a A representation a general signal.

A Spectogram of the signal denoted by **a**.

c A representation of a general signal, different from **a**.

C Spectrogram of the signal denoted by **c**.

s The time domain waveform of an arbitrary source mixed in a musical mixture.

S Spectrogram of the source denoted by **s**.

x Time-domain waveform of voice signal, could be speech or singing.

X Spectrogram of the voice signal denoted by **x**.

$\mathbf{X}_{\mathbf{voc}}$ Compressed spectral envelope pertaining the voice signal denoted by **x**.

$\mathbf{X}_{\mathbf{mel}}$ Mel-scale spectrogram pertaining to the voice signal denoted by **x**.

y Time-domain waveform of voice signal with modulations added.

Y Spectrogram of the voice signal with modulations added, **y**.

$\hat{\mathbf{x}}$ Time-domain waveform of an output voice signal.

$\hat{\mathbf{X}}$ Spectrogram of the output voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{X}}_{\mathbf{voc}}$ Compressed spectral envelope pertaining to the voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{X}}_{\mathbf{mel}}$ Mel-scale spectrogram pertaining to the voice signal denoted by $\hat{\mathbf{x}}$.

$\hat{\mathbf{y}}$ Time-domain waveform of an output voice signal, which has effects and modulations added.

$\hat{\mathbf{Y}}$ Spectrogram of the output voice signal with modulations added, $\hat{\mathbf{y}}$.

b Time-domain waveform of musical instrumental backing track.

B Spectogram of musical instrumental backing track

m The mixture signal formed by mixing **y** with **b**, the mix does not necessarily have to be a linear mixture.

M Spectogram of the mixture signal denoted by **m**.

enc The encoder network of an autoencoder.

dec The decoder network of an autoencoder.

V The latent embedding of an autoencoder.

gen The generator network of a GAN.

dis The discriminator network of a GAN.

Z The linguistic content of the voice signal, **x**.

η The melodic content of the voice signal, **x**.

ψ A representation of a singer or speaker, who is the source of **x**.

ω A soft-mask or Wiener filter used for source separation.

Chapter 8

Introduction

A musical ensemble of singers singing simultaneously is commonly known as a choir. Choral music is a tradition that has been practiced throughout society from the medieval ages to modern times, involving diverse groups of singers of various capabilities and ranges. As such, choral singing is a social activity that can be performed in various arrangements with or without instrumental accompaniment. The earliest form of ensemble choir singing can be traced back to the Gregorian chants of the 4th century, which involved multiple singers singing the same content simultaneously, in **unison**. Evidence of polyphony, with more than one *part* or *divisi* can be found in the 14th century English sacred music manuscript known as the *Old Hall Manuscript*. The use of multiple parts and polyphony in ensemble singing composition continued through the Renaissance and Baroque periods with the works of composers like Claudio Monteverdi, Heinrich Schütz and Johann Sebastian Bach. Wolfgang Amadeus Mozart, Louis-Hector Berlioz and others like Johannes Brahms and Franz Peter Schubert continued to evolve compositions of ensemble choral singing through the Romance period. Such compositions are practiced widely even today in dedicated conservatories across the world.

Being a social activity, one of the most popular formats of choral singing makes use of the distinct male and female vocal range, with female singers capable of singing high

pitches arranged in parts known as **soprano** and **alto**, while male singers are consigned to **tenor** and **bass** parts. The soprano part is typically for singers comfortable in the 260 Hz to 880 Hz vocal range. For the Alto section, the associated range is 190 Hz-660 Hz. Singers comfortable in the lower ranges 145 Hz-440 Hz and 90 Hz-290 Hz are generally assigned the Tenor and Bass voices, respectively (Scirea & Brown, 2015). An SATB choir can have just four singers, one singing each of the parts, resulting in a **quartet** arrangement. It is also common to have multiple singers singing in **unison** in each of the parts, resulting in even more pronounced **choral** effect. Choirs typically perform in large chambers which add natural reverberation to to choral mixture.

Despite its cultural and historical significance, choral music has largely been understudied in the **Music Information Retrieval** field, largely due to a lack of datasets and sufficient technological advancements to study such complex musical arrangements. Recently however, efforts have been made to create datasets, including the Choral Singing Dataset (Cuesta et al., 2018), Dagstuhl ChoirSet (Rosenzweig et al., 2020) and the ESMUC Choral Dataset have been published and are discussed in Section 3. Along with data-driven deep learning methodologies, this opened the door to studies of choral singing; initial research has focused on separating the individual parts (Petermann et al., 2020; Gover & Depalle, 2019, 2020) in SATB arrangements and estimating the fundamental frequencies of the individual voices (Cuesta et al., 2020). We leverage and further this research particularly for the case of separating the individual voices using some of the recently proposed Deep Learning based source separation algorithms proposed for musical source separation and speech source separation and adapting them to the case of SATB choir music.

We denote the voice signal of a singer in the soprano parts as $\mathbf{x}_{So}^{j}$, where $j = 1, .., J$, with J being the number of singers singing in unison in the soprano voice. The signal for the unison of sopranos, $\mathbf{x}_{So}^{U}$, is a linear mixture of the individual singers, $\mathbf{x}_{So}^{U} = \sum_{j=1}^{J} \mathbf{x}_{So}^{j}$. We assume that the linguistic content for each of the individual signals is the same and is that of the unison signal, i.e.: $\mathbf{Z}_{So}^{j} = \mathbf{Z}_{So}^{U} \forall j$. Similarly, the individual voice

signals of the singers in the alto, tenor and bass voices are denoted as $\mathbf{x}_{Al}^{j}$, $\mathbf{x}_{Te}^{j}$ and $\mathbf{x}_{Ba}^{j}$, respectively. The unison signals for the respective parts are denoted by $\mathbf{x}_{Al}^{U}$, $\mathbf{x}_{Te}^{U}$ and $\mathbf{x}_{Ba}^{U}$.

As a choir singing group participates simultaneously, the choral mixture can generally be assumed to be a linear mixture, albeit with reverberation from the the surroundings, which are typically enclosed. In our study case, the datasets we used were recorded under special conditions with limited reverberation, particularly on the choral mixture. As such, we assume a linear mixture wherein the sum of the unison signal gives us the choral mixture signal; $\mathbf{m}_{chorus} = \mathbf{x}_{So}^{U} + \mathbf{x}_{Al}^{U} + \mathbf{x}_{Te}^{U} + \mathbf{x}_{Ba}^{U}$. As this follows the typical assumption of the source separation algorithms discussed in Section 2.3, we can use TF mask based source separation algorithms to separate the individual voices within the mixture. To this end, we adapt some of the state-of-the-art deep learning based model for music and source separation to the task of voice separation in SATB choirs. As research in this field is still in its nascent phase, we conduct some initial experiments in Chapter 9 that allow us to asses the type of models that can be used for choral voice separation as well as the quality of the data needed for further research.

Such a voice separation procedure estimates of the unison components in the mixture, $\mathbf{m}_{chorus} \longrightarrow \hat{\mathbf{x}}_{So}^{U}, \hat{\mathbf{x}}_{Al}^{U}, \hat{\mathbf{x}}_{Te}^{U}, \hat{\mathbf{x}}_{Ba}^{U}$. We further propose a methodology to synthesize a prototypical single voice representation of unison singing within an SATB choir recording to allow for audio manipulations and remixing. To this end, we use the model proposed in Chapter 7, to model the linguistic content $\mathbf{Z}_h$, where $hin\{So, Al, Te, Ba\}$ of the unison signals pertaining to the individual parts. We also model the melodic content, η_h and use these to synthesize a prototypical single voice signal representing the pitch and lyrical content of the unison signal. The result of this synthesis is a single voice that can easily be transformed, e.g., pitch shifted, for creative and educational applications. This research is presented in Chapter 10. We also analyze unison singing recordings from the CSD.

Chapter 9

Choral voice separation

Source separation in the audio domain has been extensively studied for contemporary popular musical mixtures and speech signals, as discussed in Section 2.3. In the case of contemporary music source separation, the sources typically considered are *vocals*, *drums*, *bass* and other instruments that are typically grouped together as *others*. These four sources, often termed as *stems*, are used in the Signal Separation Evaluation Campaign (SiSEC) to benchmark source separation algorithms. While the musical mixture in this case comprises of melodic instruments, like the bass, the voice , synthesizers, piano and guitar, each of these instruments typically have distinct harmonic structures which are exploited for source separation, particularly by knowledge driven algorithms (Virtanen, 2007; Févotte et al., 2009; Ozerov et al., 2012; Candès et al., 2011; Huang et al., 2012).

In the case of SATB choirs, we would like to separate the soprano, alto, tenor and bass voices from a mixture signal. In this case, the sources to be separated are all voice signals, the harmonic structures of which share similarities. Additionally, singers in choir generally try to blend their voices together while singing in a choir. Voiced parts are longer in choral singing as the composition makes use of the harmonic structure of the voice. This also leads to overlapping harmonics amongst the individual parts, making it a harder task than source separation for contemporary popular music or speech.

However, one distinguishing feature that can be exploited for source separation is the distinction in the vocal range of the individual voices of an SATB mixture.

Speech source separation is a task similar to separating the voices in an SATB mixture. However, scenarios in which speech separation is applied like conversations or the cocktail party problem, where multiple speakers might be speaking simultaneously, generally do not involve synchronization between the different speakers. This allows for temporal cues which can be exploited by source separation algorithms. In addition, the voices of the speakers in the mixture have distinct timbres and frequency ranges.

Data-driven deep learning based algorithms, while often considered to be *black boxes* have shown the capability to inherently model distinguishing features within data to outperform knowledge based algorithms in the musical and speech domain. Most of these algorithms assume a linear mixture of sources and filter the mixture signal to separate the individual sources. Since the choral mixture can under reasonable limits, be considered a linear mixture, we hypothesize that such algorithms can also be effectively be applied to voice separation for the case of SATB choirs. We adapt and evaluate some of the state-of-the-art source separation algorithms from both the music and speech domains for this task. As research in source separation for ensemble mixtures is still in its nascent stage, we propose some initial experiments to asses the deep learning based models that can be adapted for choral voice separation and the quality of data required. We also note that data, as presented in Section 3.2, is quite limited and under varying recording conditions. We also note that it is easier to record a quartet, or a choral ensemble of a single singer per part, than it is to record a full choir. In this study, we answer the following research questions:

- Can waveform based source separation algorithms work as well as spectrogram based models for choral voice separation?
- Are music source separation algorithms better suited to choral voice separation or should speech source separation algorithms be used?

- How can we curate data from varied datasets which has been recorded under different conditions?

- Can quartet based data with a single singer per part be used to train deep learning based algorithms for voice separation even with multiple singers per part in unison?

9.1 Related work

Source separation for synthetic choral data has been studied using score-informed Non-Negative Matrix Factorization (NMFs) and the Wave-U-Net architectures (Gover & Depalle, 2019, 2020). In this case, the researchers synthesized 371 pieces of choral compositions by Bach using a commercial MIDI synthesizer named *FluidSynth*. This allowed for synthesized choral mixes and stems aligned with score information. Following this, the Wave-U-Net (Stoller et al., 2018) architecture was adapted to accepted temporal conditioning. The conditioning was applied both at the input and output layers, as well as the downsampled bottleneck layer. It was shown that the Wave-U-Net architecture outperformed the NMF based baseline, even without the conditioning.

Voice separation in real world recordings of SATB choirs has been studied using transfer learning (Bugler et al., 2020), with a ChimeraNet model (Luo et al., 2017) pretrained on the MUSDB and Slakh (Manilow et al., 2019) datasets. This model was then fine tuned to separate the male and female voices in SATB recordings in the Daghstuhl dataset. Another model for voice separation for real world recordings of SATB choirs was proposed by us (Petermann et al., 2020), using a conditioned variant of the U-Net model (Meseguer-Brocal & Peeters, 2019). In this case, a feature-wise linear modulation (*FiLM*) layer (Perez et al., 2018b). This layer uses an affine transform across the model architecture, allowing for the application of linear transformations to intermediate feature maps. These specialized layers conserve the shape of the original intermediate feature input while modifying the underlying mapping of the filters

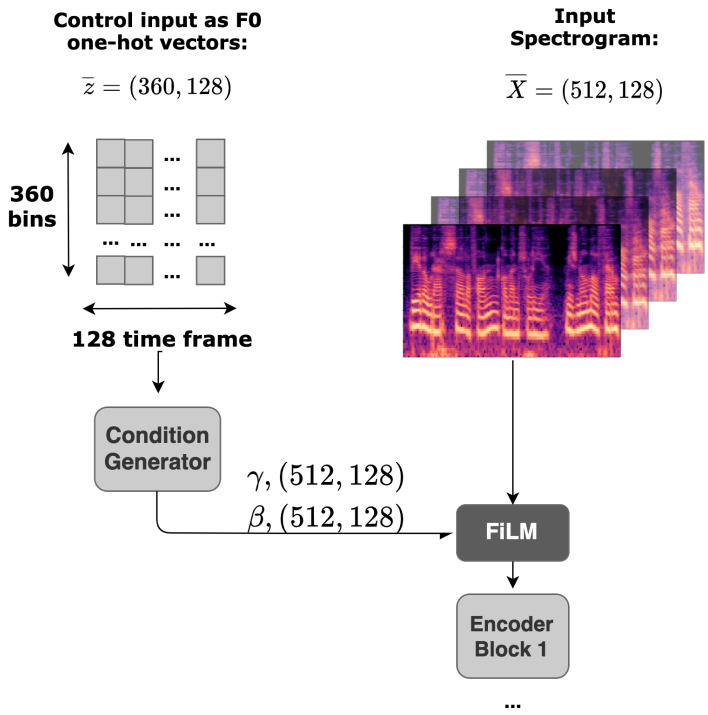

Figure 9.1: C-U-Net Control Mechanism adapted for voice separation in SATB choirs, using the oracle f0 as a condition for separating the voices (Petermann et al., 2020).

themselves This work was carried out by a masters' student in the Universitat Pompeu Fabra, under the supervision of the author of this thesis. A brief overview of this model is presented in Appendix D.

9.2 Source separation algorithms for voice separation

While several models have been proposed for music and speech source separation, we chose 4 to answer the research questions highlighted in this chapter. We compare the U-Net (Jansson et al., 2017) algorithm with its waveform equivalent, the Wave-U-Net (Stoller et al., 2018). While the original U-Net (Jansson et al., 2017) used separate networks for estimating the masks for the vocals and instrumental accompaniment stems, we used a single network with 4 output layers to predict the masks for each of the parts to be separated. We do this since we want to directly compare the waveform based

Wave-U-Net model with the spectrogram based U-Net model.

We also use the Open-Unmix (Stöter et al., 2019) model which is one of the best performing models for music source separation and compare it with the Conv-TasNet (Luo & Mesgarani, 2019) model, which was proposed for asynchronous speech source separation. While the Conv-TasNet has been adapted for music source separation (Samuel et al., 2020; Défossez et al., 2019), we specifically use the version proposed for speech source separation since we want to want to asses which domain is better suited to choral voice separation, which has elements of both music and speech sources. We use the single channel version of the Open-Unmix model, which uses single channel Wiener filters instead of multichannel Wiener filter (MWF) (Nugraha et al., 2016). The Open-Unmix model uses separate networks for each of the voices to be separated. The models used are summarized in Table 9.1.

Model	Input	Originally Proposed For
U-Net[15]	Spectrogram	Music Source Separation
Wave-U-Net[16]	Waveform	Music Source Separation
Open-Unmix[17]	Spectrogram	Music Source Separation
Conv-TasNet[18]	Waveform	Speech Separation

Table 9.1: The deep learning based source separation models we adapt for voice separation in SATB choirs, along with the input type and the context they were originally proposed for.

9.3 Experiments

9.3.1 Datasets and training

As observed in Section 3, there is a dearth of data available for ensemble singing, as compared to the data available for contemporary polyphonic music. The models we wish to adapt for voice separation in ensemble singing, were initially trained with large datasets. For our study, we used the Choral Singing Dataset (CSD), the Daghstuhl Dataset (DSD), the Bach Chorales Dataset (BCD) and the ECD for training and evaluation of the models. For consistency, we re-sample all data to 22.05 kHz and use different combinations of singers within the datasets to augment the data for training.

Since each of these datasets had distinct recording settings and varying amounts of data, they provide us an opportunity to effect of incrementally increasing the data required for training the various models. Additionally, different combinations of singers within songs can add variations in pitch timing and timbre, leading to slight differences in the input and target signals.

We define two variations of the data; *case*_1 includes all combinations of quartets within the datasets, limiting the number of singers per voice to 1 and *case*_2, in which we train the models with all possible combinations of singers for a song. This allows us to assess the impact of training the source separation models with quartets while evaluating the models on separation for full choirs, i.e. mixtures which might have unison signals of two or more singer present per part.

We train the models with the CSD dataset, using the nomenclature $modelname_C$ and with the CSD and BCD, named $modelname_{CB}$. Where $modelname \in \{UNet, WaveUNet, Unmix, CanvTasNet\}$. Since the BCD only contains quartets, this allows us to evaluate the impact of augmenting full choir data from the CSD with quartet data for separating the parts in mixtures which have unison singing present. On initial evaluation, we found the $Unmix_{CB}$ model to be the best performing of these models. We use this model to clean the individual parts of the ECD, which had significantly higher leakage amongst the individual parts than the CSD dataset. We do this by passing the stems of each of the singers of each of the parts through the corresponding model of the trained $Unmix_{CB}$ model. This allows us to filter out interference from parts of the choir that do not pertain to the target part. We use this dataset for evaluation of the trained models.

We use early stopping for training the U-Net, Wave-U-Net and Open-Unmix models, using one singer per part of one song from the CSD dataset for validation while training the $modelname_C$ models. For training the $modelname_{CB}$ models, we use one song from the BCD dataset as well as the validation set from the $modelname_C$ models. We use a patience of 50 epochs for the U-Net and Wave-U-Net models and a patience of

200 epochs for the Open-Unmix model. The Conv-TasNet model is trained for 2000 epochs, without early stopping.

9.3.2 Evaluation

We use the Source to Distortion Ratio (SDR), Source to Interferences Ratio (SIR) and Sources to Artifacts Ratio (SAR) metrics from the *bss_eval_sources* set of evaluation metrics (Vincent et al., 2006) for evaluation of the adapted models. While we calculated these metrics for all models, we only present the evaluation of the models trained with *case*_2. We found that there was a slight improvement in the models trained using *case*_2 over the models trained with *case*_1 data, showing that given additional quartet based data, we can further train the models to improve performance, even for separating the parts in a full choir with unison singing in the individual parts.

9.3.3 Results

The results of the evaluation of the first experiment, with the $modelname_C$ and $modelname_{CB}$ models evaluated using the *bss_eval_sources* set of metrics (Vincent et al., 2006) are shown below. Figure 9.2 shows the SDR metric for the four models, trained with *case*_2 data, while Figures 9.3 and 9.4 show the SAR and SIR metrics, respectively. The evaluation metrics were calculated over the entire cleaned version of the ECD, using full choirs, i.e. with 16 singers in unison per part.

We observe that the $modelname_{CB}$ models significantly outperformed the $modelname_C$ models, particularly in terms of SIR and SDR. This shows that the quartet data that was present in the BCD dataset was sufficient for training the models to separate the parts in a full choir containing unison. We also note that the $Unmix_{CB}$ has the best overall performance, which is expected since this model outperformed the others on the task of musical source separation. The performance of the *UNet* and *WaveUNet* models is comparable for each of the metrics calculated. This shows us that both waveform based and spectrogram based source separation models can effectively be used for separating

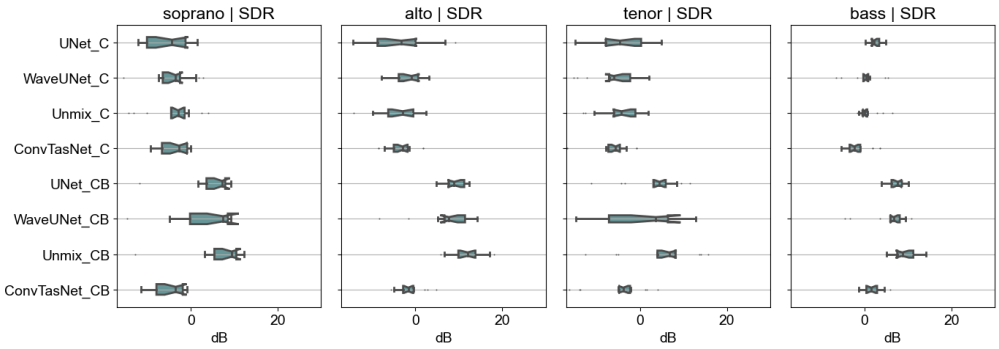

Figure 9.2: The SDR metric evaluated on the cleaned ECD for the four models trained with *case*_2 data.

the parts in an SATB mixture. Although we note that the *ConvTasNet* does not perform well on the SDR and SIR metrics, but outperforms the other models on the SAR metric. This result suggests that music source separation algorithms are better suited to voice separation for the case of SATB choirs than those proposed for asynchronous speech separation. We also observe that the SIR and SDR for the tenor parts of all models is significantly lower than that for the other parts. We believe this is in part due to the overlap between the vocal range for the alto and tenor parts. Since the f0 is a major distinguishing feature between the parts, we believe that the various models are confused between these parts. On further analysis, we found that there was indeed confusion between these parts in the separated voices, particularly with segments of the alto part being separated as the tenor part. This can also be observed in the SIR plot which has a much higher variance for the tenor part than it does for the alto part, especially for the $modelname_{CB}$ models.

9.4 Conclusions

We adapted and evaluated four open source deep learning based models for music and speech source separation for the case of voice separation for SATB choirs. The models are trained using data from four recently recorded datasets, which have different recording conditions and combinations of singers within each of the parts. The size of

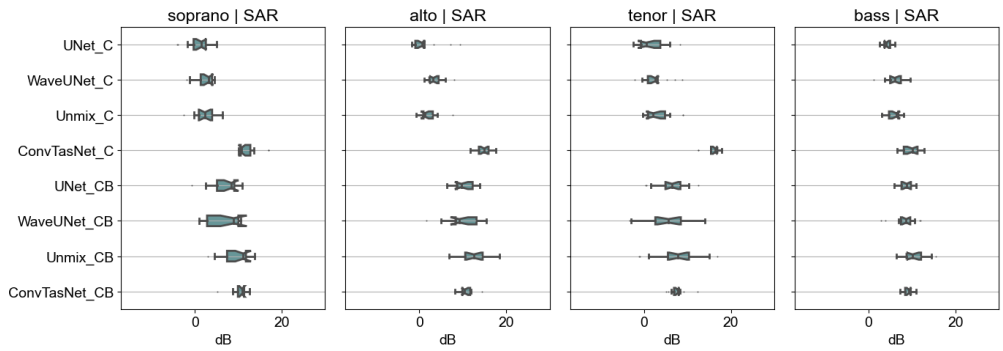

Figure 9.3: The SAR metric evaluated on the cleaned ECD for the four models trained with *case*_2 data.

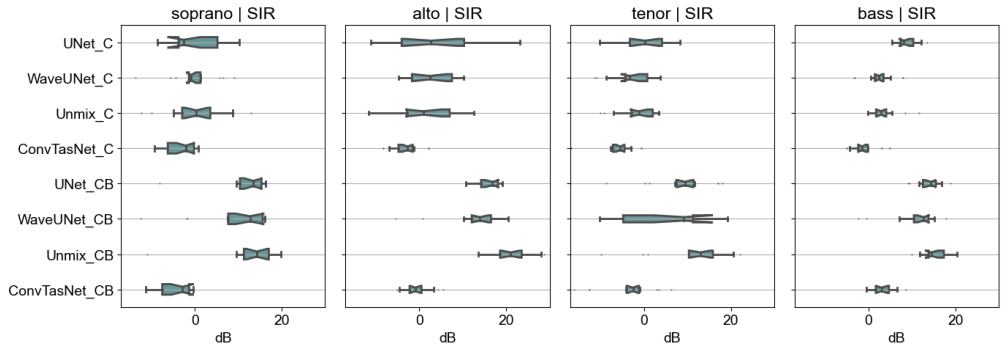

Figure 9.4: The SIR metric evaluated on the cleaned ECD for the four models trained with *case*_2 data.

the datasets used, even after consolidation, is relatively small compared to the size of the datasets used to train the models for music and source separation. However, allowing different combinations of singers during training allows us to augment the data. We are now in a position to answer the research questions presented in the introduction to this chapter:

- Can waveform based source separation algorithms work as well as spectrogram based models for choral voice separation?

  We observe that the performance of the waveform based Wave-U-Net (Stoller et al., 2018) is comparable to that of the spectrogram U-Net (Jansson et al., 2017) when both are adapted under similar conditions for the task for voice separation

in SATB choral mixtures.

- Are music source separation algorithms better suited to choral voice separation or should speech source separation algorithms be used?

  We observe that the music source separation algorithms, particularly the Open-Unmix (Stöter et al., 2019) model, outperform the model proposed. for asynchronous speech source separation, Conv-TasNet (Luo & Mesgarani, 2019). Although we note that an adapted version of the Conv-TasNet model outperforms the Open-Unmix model on the task of music source separation, the model proposed for speech source separation has been shown to under perform.

- How can we curate data from varied datasets which has been recorded under different conditions?

  We observe that by using a preliminary model trained on the dataset with least leakage, the CSD, we were able to reduce the inter-singing leakage in the individual tracks of the ECD and use the data for further training and evaluation of the models.

- Can quartet based data with a single singer per part be used to train deep learning based algorithms for voice separation even with multiple singers per part in unison?

  We observe that the best strategy to use is to use all possible combination of singers from within a song. From the first experiment, we can observe that quartet based data can be used to augment data for which multiple singers are available per part. Such data augmentation leads to significant improvement in the performance of the separation system when applied to a full choir with multiple singers per part. We also observe that using all possible combinations of singers per part leads to a slight improvement in results over restricting the number of singers per part to 1.

We believe this study to be an initial foray into the domain of voice separation for SATB choirs. We hope it provides a baseline for future work in this direction as algorithms for musical source separation evolve and more datasets for SATB choirs are made available.

Chapter 10

Analysis of unison singing

Choir singing is a group activity with multiple singers from diverse background coming together to sing. In Chapter 9, we discussed choir singing in harmony. In this chapter, we will cover multiple singers simultaneously singing the same melodic and linguistic content, leading to a an effect known as **unison**. We will denote the signals of the individual singers within the unison as $\mathbf{x}^j$ where $j = 1,2...K$, where K is the total number of singers in the unison. We also denote the unison signal as $\mathbf{x}^U = \sum_j = 1^K \mathbf{x}^j$. We note that while the unison signal is a linear sum of the individual signals, a filtering approach cannot be used to separate the individual signals since they are very similar in content. As such, we use the methodology proposed in Chapter 7 to model the linguistic content, $\mathbf{Z}^U$, of the signal. We also model the melodic content, η^U of the signal and use these to synthesize a prototypical single voice signal representative of the content of a unison input. We term the framework for this synthesis as Unison to Solo (UTS). We this model to study the perceptual pitch of the unison mixture.

We also analyze real world recordings of unison singing from the Choral Singing Dataset (CSD), studying the inter-singer timing and pitch deviations and use these to propose a methodology for generating a signal with the unison effect from an a capella input. We term this methodology as the Solo to Unison (UTS) model. In concrete, this chapter attempts to answer the following research questions:

- Is it possible to separate a single voice from within the unison singing signal?
- What are the perceptual qualities of a unison signal that distinguish it from a signal voice singing signal?

10.1 Related work

Unison singing has been studied in the past with the use of vowel based singing voice synthesizers (Ternström, 1991) with listening tests involving expert listeners. The studies show that even though multiple voices are present in a signal pertaining to unison singing, a listener only perceives a single pitch. Properties of the singing voice signal like pitch, timing, loudness and timbre can be defined as statistical distributions for ensemble unison singing. The researchers investigated **pitch dispersion** within the unison, defining it as the the bandwidth of the fundamental frequency and its harmonic partials across individual singers in a unison. This dispersion is related to small variations in the f0 that are considered too fast to be perceived as variations in pitch. Such variations are termed as **flutters**. The deviation of individual f0 contours over the mean of the individual contours in the unison is termed as **scatter** and a preference of 0 cents to 5 cents was shown by the participants in the listening test while 5 cents to 14 cents was shown to be the upper limit of tolerance of consonance. Studies of modelling scatter have also been conducted with real world choral recordings using small windows to compute the standard deviation between individual f0s in the unison signal (Cuesta et al., 2018). The pitch dispersion was found to be in the range of 20 cents to 30 cents for the recordings considered in the study.

Other studies of ensemble singing include an attempt to measure intonation quality within choral recordings (Weiss et al., 2019). The deviation of each individual f0 value was calculated from an ideal 12-tone equal temperament grid, allowing the analysis of the overall intonation of a full choir recording.

We further the analysis done by (Ternström, 1991) using the methodology proposed

in Chapter 7 for synthesizing a single singer representation of a unison singer from an SATB choir recording. This allows us to analyze the perceptual qualities of the unison via listening tests. Further, the methodology can also be applied for synthesizing a unison signal from an a capella singing voice recording.

10.2 Unison to solo

To synthesize a single singer representation of the content of a unison signal, we need to extract the linguistic and melodic content. We use the **Singer Independent Network** proposed in Chapter 7 to model the linguistic content from the unison signal.

To model the melodic content, we try to emulate the single perceived pitch of the unison, which was noted by Ternstöm (Ternström, 1991). Intuitively, this perceived pitch, denoted by η^U, must be a function of the fundamental frequencies of the individual signals that comprise it. We hypothesize that the simplest possible function, the mean, of the individual f0's can be used as a representation of the perceived pitch of the unison signal. The mean f0 value, $\tilde{\eta}^U$ is adjusted for timing differences, discussed in the next section, between the individual singers. To this end, we define the average to be zero (unvoiced frame) if and only if all individual values for that frame are zero. for all other cases, the average is calculated only accounting for the non-zero values.

To extract a representation of this pitch from a unison signal, we use a deep learning based monophonic pitch estimation algorithm, known as Convolutional Representation for Pitch Estimation (CREPE) (Kim et al., 2018b). Since the algorithm estimates a single pitch, it can be expected to estimate the single perceived pitch, $\hat{\eta}^U$, of the unison signal. We compare the extracted pitch with the theoretical perceived pitch in Section 10.5.

We use the SIN model presented in Chapter 7 to model the linguistic content of the signal and to generate the compressed spectral envelope corresponding to the prototypical representative signal for the unison signal, $\mathbf{x}^U$. As discussed in Chapter 8, we assume

that the linguistic content for each of the individual signals is the same and is that of the unison signal, i.e.: $\mathbf{Z}^j = \mathbf{Z}^U \forall j$. To this end, we use the linguistic encoder, $enc_{ling}()$ to extract $\mathbf{Z}^U$ from the input, $\mathbf{x}^U$. The singer encoder, $enc_{singer}()$, is used to derive a representation of the perceived singer identity from the unison signal, ψ^U. The linguistic content and the perceived singer identity are passed through the SIN decoder, $dec_{SIN}()$, to generate the compressed spectral envelope for the prototypical single voice signal pertaining to the unison, $\hat{\mathbf{X}}^S_{\mathbf{voc}}$

$$\begin{aligned} \psi^U &= downsample(enc_{singer}(|\mathbf{X}^U|)) \\ \mathbf{Z}_U &= downsample(enc_{ling}(|\mathbf{X}^U|)) \\ \hat{\mathbf{X}}^S_{\mathbf{voc}} &= dec_{SIN}(upsample(\mathbf{Z}^U), upsample(\psi^U)) \end{aligned} \tag{10.1}$$

Finally, we use the extracted pitch, $\hat{\eta}^U$ and the generated compressed spectral envelope, $\hat{\mathbf{X}}^U_{\mathbf{voc}}$ to synthesize the prototypical voice signal, $\hat{\mathbf{x}}^S$, as shown in Figure 10.1.

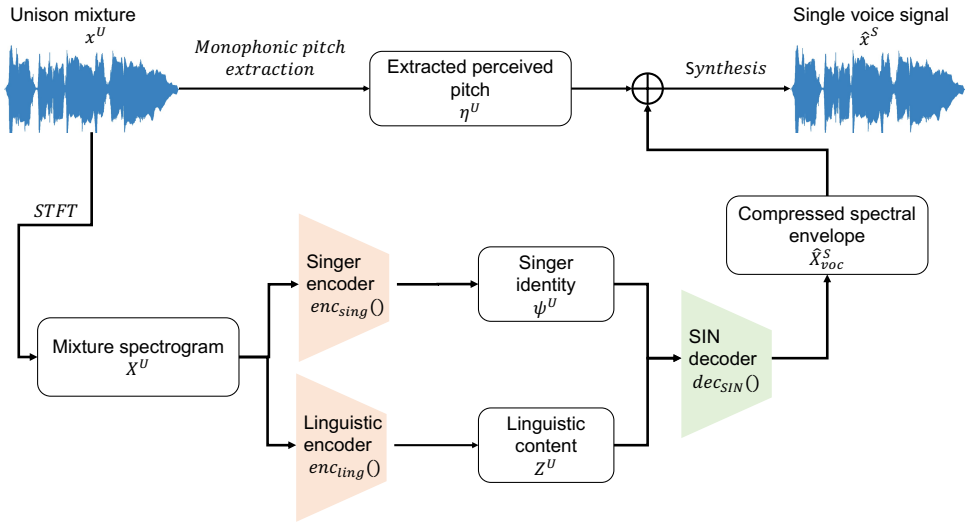

Figure 10.1: The framework for synthesizing a prototypical single voice signal from a unison mixture.

10.3 Timing and pitch deviations in unison singing

To analyze the pitch deviation within the unison, we build a statistical model for the individual contours in the unison, as suggested by (Ternström, 1991). In our model, the framewise f0 of an individual singer, η_i, can be represented as a distribution of values around the mean η_m^U with a deviation of η_{devi}, as shown in Equation 10.2

$$\eta_i = \tilde{\eta}^U + \eta_{devi} \tag{10.2}$$

This also allows us to define the η_{i+1} of a singer in terms of the η_i of another singer in the unison as shown in Equation 10.3.

$$\begin{aligned} \eta_{i+1} &= \tilde{\eta}^U + \eta_{devi+1} \\ \eta_{i+1} - \eta_i &= \eta_{devi+1} - \eta_{devi} \\ \eta_{i+1} &= \eta_i + \eta_{devi+1} - \eta_{devi} \\ \eta_{i+1} &= \eta_i + \Delta\eta \end{aligned} \tag{10.3}$$

We define $\Delta\eta$ as the inter-singer deviation, represented by Equation 10.4. For each pair of singers in the unison, we compute the frame-wise difference between the corresponding f0 contours in cents. For this calculation, only *voiced* frames were considered. We average these inter-singer deviations across time and songs, and obtain a single value for each part in soprano, alto, tenor and bass.

$$\Delta\eta_h = \frac{\sum_{i=1}^{n}\sum_{j=i+1}^{n} \eta_i - \eta_j}{\binom{n}{2}} \tag{10.4}$$

where the sub-index h indicates the choir section, $h \in [So, Al, Te, Ba]$, and n is the number of singers. In our use case, $n = 4$.

Pitch deviations across the singers in the unison mixtures calculated using this methodology are shown in Figure 10.4. We see that the calculated inter-singer deviation, $\Delta\eta_h$,

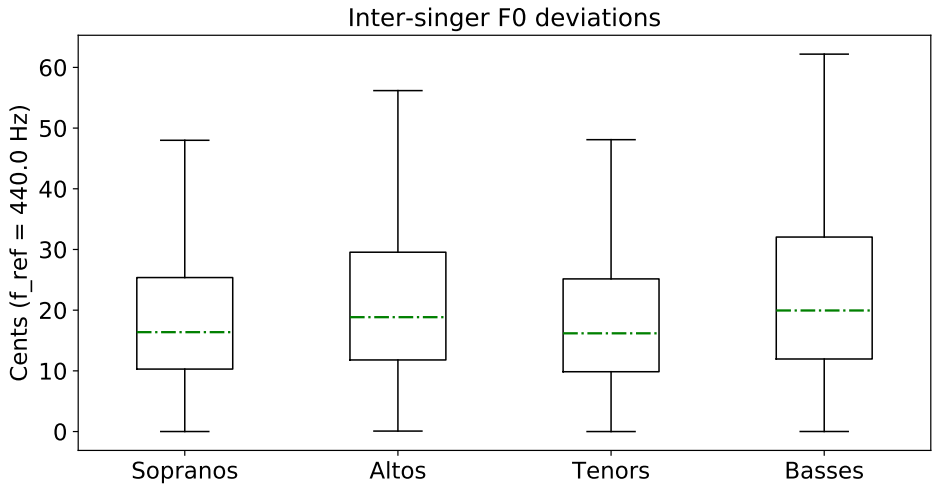

Figure 10.2: Inter-singer deviations in cents averaged across the whole dataset for each choir section. Deviations are calculated using Equation 10.4.

is in the range of 0 cents to 50 cents, with a mean of around 20 cents. This value is comparable to the pitch dispersion studied by (Cuesta et al., 2018). While the methodology for modelling is different, these results are in accordance with the reported per-section pitch dispersion: larger in the bass section, smaller in the sopranos, and very similar for altos and tenors.

To measure timing deviations, we focus on the transitions from *voiced* frames to *unvoiced* frames across the individual signals within the unison. The voiced frames are the frames where a positive f0 is annotated for the individual singer whereas unvoiced frames are those which have a 0 value for the f0 annotated. These transition regions occur at the start and end of phrases where some singers might have start or stop singing earlier than others and might not be completely synchronized. We measure the length of all the transition regions in every unison from the CSD, and average the time value across choir sections.

Timing deviations calculated in this manner are shown in Table 10.1 We observe an average timing deviation of 0.1 s between the voices in the unison for all parts of the choir.

Section	Average Timing Deviation ± Standard Deviation
Soprano	0.134 ± 0.039 sec
Alto	0.093 ± 0.0024 sec
Tenor	0.100 ± 0.021 sec
Bass	0.124 ± 0.021 sec

Table 10.1: Timing deviations averaged across the CSD. These values measure the time span in which all singers in the unison transition from voiced to unvoiced, and vice-versa, averaged across all transitions in each song.

10.4 Solo to unison

For synthesizing a unison signal from an a capella singing voice input, we create voice clones of the signal with pitch and timing deviations as well as variations in timbre using the Singer Dependent Network (SDN) described in Chapter 7. To this end, we use the linguistic encoder, $enc_{ling}()$, to extract the linguistic content $\mathbf{Z}^{Si}$ from the input signal $\mathbf{x}^{Si}$. Timbre changes are emulated by using Voice Conversion using the SDN model, wherein singers of the male gender from the training set are used for the Tenor and Bass parts. For the Soprano and Alto voices, we use Female singers from the training set. We create 4 clones of each voice with singer identity ψ_i', where $i \in 1,2,3,4$. These are passed through the SDN decoder, $dec_{SIN}()$, to generate the compressed spectral envelope $\hat{\mathbf{X}}^{U}_{\mathbf{voc}i}$, as shown in Equation 10.5.

$$\begin{aligned} \mathbf{Z}_{Si} &= downsample(enc_{ling}(|\mathbf{X}^{Si}|)) \\ \hat{\mathbf{X}}^{U}_{\mathbf{voc}i} &= dec_{SDN}(upsample(\mathbf{Z}^{Si}), upsample(\psi_i')) \end{aligned} \tag{10.5}$$

To model inter-singer deviations, $\Delta\eta$, we add noise sampled from a normal distribution with a mean of 0 and a standard deviation represented by the parameter *std* to the f0 of the input a cappella signal, η^S as $\hat{\eta}_i^U = \eta^S + \Delta\eta_i$.

We shift the voiced portions of the input signal between two successive blocks of si-

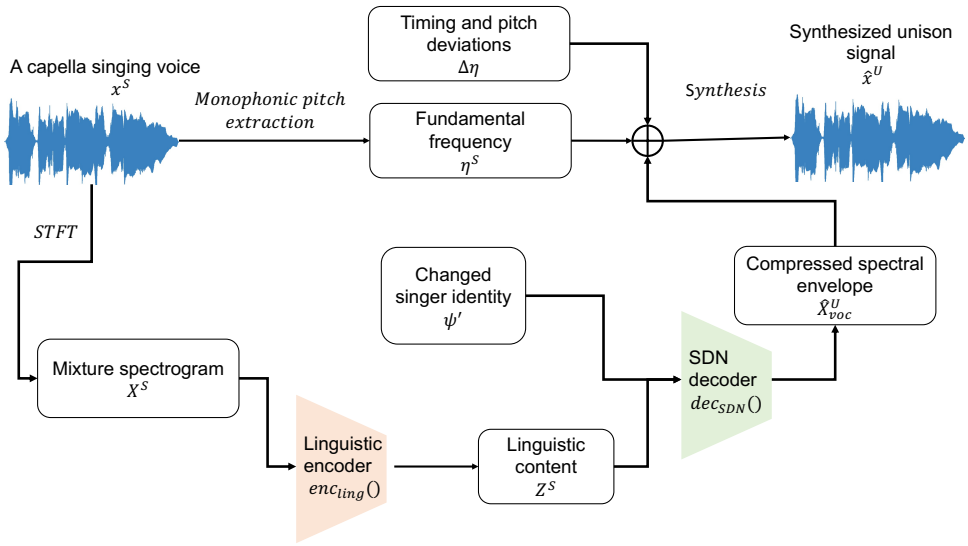

Figure 10.3: The framework for synthesizing a prototypical unison signal from an a capella input.

lence of more than 80 ms by a variable amount for each of the clones. The shift is randomly sampled from a normal distribution of mean 0 and standard deviation ts. The framework for this methodology is shown in Figure 10.3.

The individually synthesized signals, $\hat{\mathbf{x}}_i^U$ are summed to generate the output unison signal, $\hat{\mathbf{x}}^U = \sum_{i=1}^{4} \hat{\mathbf{x}}_i^U$

10.5 Experiments

We use the CSD for evaluation of the proposed methodologies for UTS and STU. We used the proprietary dataset, presented in Section 7.4.2, for training the SDN and SIN models used in the STU and UTS models. To the best of our knowledge, there was no overlap in the singers used for training and those involved in CSD.

For the UTS model, we evaluate the accuracy of the monophonic pitch extraction system, the remseblance of the perceived singer of the synthesized prototypical signal to the individual singers in the unison. These evaluations are presented in Section 10.5.1 and Section 10.5.2, respectively. To evaluate other perceptual aspects of the STU and

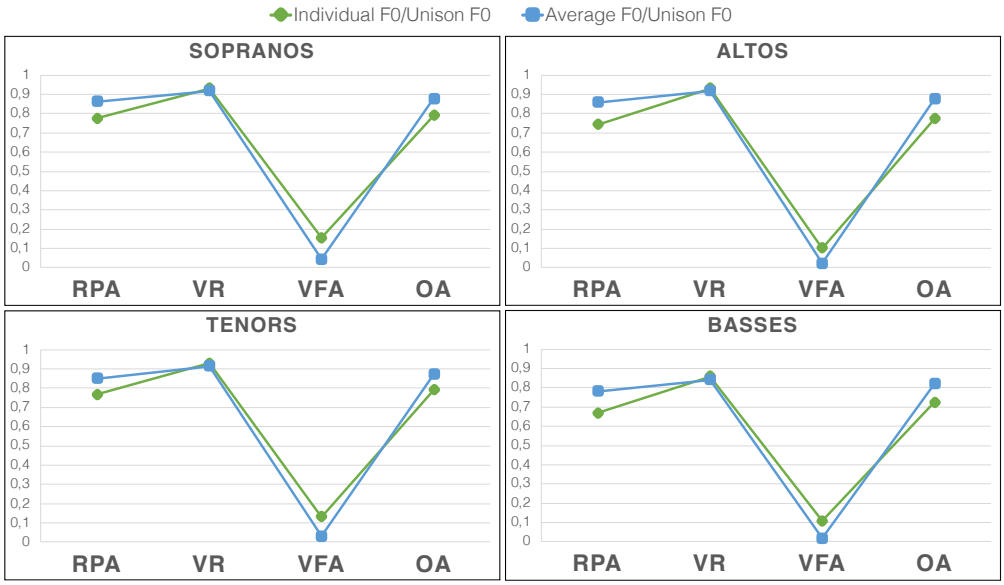

Figure 10.4: Resemblance of the estimated unison $\hat{\eta}^U$ estimation to each individual η_i contour (green) and the average (blue) using pitch evaluation metrics averaged across each choir section.

UTS models like adherence to melody, intelligibility and quality, we used subjective listening tests, presented in Section 10.5.3.

10.5.1 Pitch accuracy

We evaluate the accuracy of the monophonic pitch estimation system, compared to the theoretical mean f0 of the unison signal. We also measure the resemblance of the estimated f0, $\hat{\eta}^U$, to f0 annotations of each of the tracks. This is done using standard evaluation metrics for melody extraction including *Raw Pitch Accuracy (RPA), Overall Accuracy (OA), Voicing Recall (VR)* and *Voicing False Alarm (VFA)* between the extracted f0, $\hat{\eta}^U$, the average the mean, $\tilde{\eta}^U$, and each individual singer curve, η_i. We use the `mir_eval` library (Raffel et al., 2014) for this evaluation, with a pitch tolerance of 30 cents. The results of this analysis are shown in Figure 10.4.

We observe that all sections follow the same pattern with similar metric values, and the unison f0 estimated by CREPE, $\hat{\eta}^U$, is closer to the average $\tilde{\eta}^U$,than to the individual contours. In addition, all metrics improve when we compare the average f0 curve to

the extracted f0 contour from the unison: RPA, VR and OA are higher in the blue plots, while VFA is lower. We can thus use the pitch estimated by CREPE, $\hat{\eta}^U$, as a representative of the mean single pitch contour perceived in a unison mixture.

10.5.2 Singer analysis

We evaluate the resemblance of the perceived singer of the synthesized single voice prototypical signal, ψ^U, to the individual singers within the unison, ψ_i. To this end, we use the GE2E speaker representation embeddings (Wan et al., 2018), described in Section 2.5.3. For visualization, we use the t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten & Hinton, 2008), which is a commonly used dimensionality reduction methodology used for visualization of high-dimensional data. We reduce the dimensions of the embedding from 256 to 2, to allow for a 2-D visualization. This visualization is shown in Figure 10.5. Within the figure, the dimension-reduced speaker embedding extracted from each individual singer from all songs within the CSD dataset for each of the parts of the SATB choir are shown along with the dimension-reduced embeddings for the synthesized prototype signals for each of the songs.

It can be seen that reduced embeddings from the various parts form clusters along the two dimensions plotted. The rembeddings extracted from the synthesized prototpye signals fall within the clusters in the plot. However, we note that the speaker embedding extracted as such might not completely represent the timbre of the singers as it is also influenced by the f0 of the singing voice. The shown clusters might also be influenced by the f0, which consistently falls within the same range for a given part in the choir, regardless of the singer in question.

10.5.3 Subjective evaluation

We use a subjective MOS based listening test to evaluate the final synthesized prototypical voice signal based on three criteria:

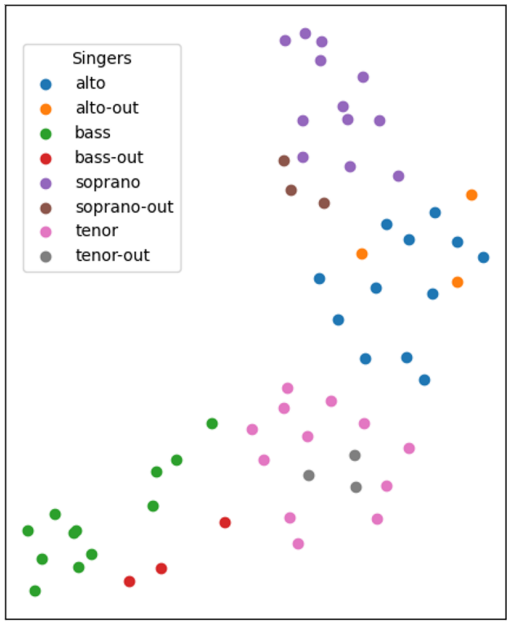

Figure 10.5: The t-SNE plots for the reduced dimension spekar embeddings. The different parts of the SATB choir are shown with different colours, as indicated by the key. The speaker embedding extracted from the synthesized. protopytical signal is labelled as $h - out$, where $h \in [soprano, alto, tenor, bass]$.

- **Adherence to melody and lyrics**: We wanted to see the similarity of the perceived pitch contour of the output to that of a ground truth unison mixture. We provide a ground truth unison reference sample made by summing the corresponding four individual singers of a part to form a unison mixture. We term this reference as *REFU*. The participants were asked to rate test samples which included the single voice prototype of the unison as output by the UTS system, referred to as *UTS*. We also evaluated the output of STU with a pitch deviation with parameter *std* set to 50 cents, the acceptable limit of pitch deviations, as shown by our analysis and suggested by (Ternström, 1991). Four singers were used for generating this test case, with parameter *ns* set to 4, and it is referred to as *STU_PS*. We also evaluated the output of the UTS system with both pitch and timing deviations with parameter ts set to 40 ms. While our analysis suggests that higher values of ts could have been used, we found that increasing the value

beyond 40 ms leads to a unacceptable level of degradation in output quality. We refer to this test case as *STU_PTS*. In addition, we provided a lower anchor of a sample of the same length from another vocal part.

- **Perception of unison**: We aim to study the perception of unison in this study, using a ground truth unison reference sample, *REFU*. Participants in the listening test were asked to rate outputs from the STU system based on their similarity to the reference in terms of the perception of unison. In addition to the *STU_PTS* and *STU_PS* cases with pitch, timing and timbre variance, we also tested the case for just timing and singer variation, referred to as *STU_TS* and a case with just pitch and timing deviations, referred to as *STU_PT*. Timbral changes were not done for the voice clones used for creating the STU_PT samples. We provided a lower anchor of an *a cappella* sample of a single singer singing the same example as the reference.

- **Audio Quality**: For the evaluation of audio quality, we set an upper limit of audio quality to the re-synthesis of a single voice recording with the WORLD vocoder. We term this as *REFS*. We also use a lower anchor with re-synthesis of a unison mixture with the WORLD vocoder, termed as *RESSYNTHU*. The examples provided to the participants were the same as those provided for the adherence to melody case, with the exception of the lower anchor.

The listening test consisted of 4 questions for each aspect, each pertaining to a part of the SATB choir. The participants were asked to rate the presented samples in the question on a continuous scale of $1 to 5$ with respect to a presented reference. The test samples and references provided pertained to the the same section of the song and were between $7 s to 10 s$ each.

There were 17 participants in our evaluation, of which 10 had prior musical training. To account for inter-participant variance in subjective evaluation, the opinion score for each question was normalized over ratings for the reference and the lower anchor before

Test Case	Adherence To Melody	Unison Perception	Audio Quality
UTS	3.6 ± 0.93		2.1 ± 0.65
STU_PS	3.3 ± 0.83	2.6 ± 0.85	2.8 ± 0.45
STU_PTS	2.9 ± 1.14	3.2 ± 0.96	3.1 ± 0.63
STU_TS		2.3 ± 1.11	
STU_PT		3.0 ± 1.23	

Table 10.2: Mean Opinion Score (MOS) $\pm$ Standard Deviation for the perceptual listening tests across the test cases provided. The models shown correspond to the Unison to Solo (UTS), the Solo to Unison with pitch, timing and singer variations, indicated by the addition of the letters P,T and S as suffixes to the abbreviation, respectively. The scores for each question were normalized by the responses to the upper and lower limits for the responses defined in Section 10.5.3.

calculating the mean opinion scores (MOS) and the standard deviations in opinion scores, presented in Table 10.2.

We can see that a higher preference was given for the UTS model over the STU model, with regards to the perceived adherence to melody. This shows that the synthesized prototypical with the adjusted mean f0 does indeed follow the perceived melody as hypothesized. For the STU model, we note that variations in both timing and pitch together are important for the perception of the unison effect, although timbre variations are not as influential. We note that there is room for improvement in terms of audio quality, this can be addressed by using alternative neural synthesis techniques, as discussed in Section 2.4.2 instead of the WORLD vocoder. The subjective nature of the perceptual aspects evaluated must be taken into account for the evaluation and the mean opinion scores are indicative of preferences rather than absolute measures of quantity.

10.6 Conclusions

We analyzed the unison mixtures within the Choral Singing Dataset, building on the work done by (Ternström, 1991) and (Cuesta et al., 2018). We observe that the devi-

ation between the f0 contours of the individual singers in the unison mixtures in the dataset is in the range of $0\,\text{cents} to 50\,\text{cents}$, while the timing deviation is around 0.1 s.

We use the mean of the pitch of the individual singers within a unison to represent the perceived itch of the unison signal. We compare this pitch to the pitch curve extracted by a deep learning based monophonic pitch extraction system. We observe that thee extracted pitch is closer to the mean of the pitches than it is to any of the individual pitches in the signal. The methodology proposed in Chapter 7 is used along with the pitch extracted by the monophonic extraction system to synthesize a single voice prototype signal representing the melodic and linguistic content of a unison mixture input. Through subjective listening tests, we support our hypothesis that the single perceived pitch can be represented by the mean of the individual pitches.

We also use the methodology to synthesize a unison mixture from a single voice input, introducing timing, pitch and timbre variations. Based on these systems, we were able to conduct a perceptual evaluation of the unison, further supporting the claim of (Ternström, 1991) that the a mixture of different voices singing in unison is perceived to have a single pitch. In addition, we found that pitch and timing deviations together are important for the perception of the unison, and that variations in either aspect alone is insufficient for such. However, timbre variations were not found to be as relevant.

We are now in a position to answer the research questions presented in the introduction of the chapter:

- Is it possible to separate a single voice from within the unison singing signal? We believe that the individual signals mixed together in a unison singing signal are too similar to be separated via TF mask based methodologies. However, we propose a novel methodology to synthesize a single voice signal representative of the perceived content of the unison signal.

- What are the perceptual qualities of a unison signal that distinguish it from a signal voice singing signal? Through our experiments, we can confirm with past

findings that the single perceived pitch of a unison signal is closely related to the mean of the pitches of the individual singers in the unison. We also. note that pitch and timing deviations contribute more to the perception of a unison than variance in the timbre of the singers involved.

Part IV

Applications, Discussion and Future Work

Chapter 11

Applications

We discuss some potential applications of the research and methodologies proposed in this thesis. The proposed methodology for singing voice extraction via synthesis can be applied for creative use as well as for transcription and pedagogical use. We also apply the analysis-synthesis approach followed in this thesis to other musical elements like single-shot percussive sounds and loops.

11.1 Processing and re-mixing

The methodology proposed in Part II has several practical applications, including lyrics transcription and artistic remxing. Since the methodology synthesizes a clean version of the vocals, without effects that might have been added during processing, it can be easier for automatic lyrics transcription (ALT) algorithms (Mesaros, 2013; Kruspe & Fraunhofer, 2016; Gupta et al., 2018a,b; Demirel et al., 2020; Tsai et al., 2018) to extract interpretable linguistic representations like phonemes from the synthesized vocals than it would be from processed vocals. This can be particularly useful for commercial applications since most contemporary popular music contains vocals with processed effects, as discussed in Section 1.2.1.

Other applications could be for remixing the synthesized clean vocal signal. Since the proposed methodology synthesizes a clean vocal signal, we believe it can also be

applied in conjunction with TF mask based source separation methods. This would further mimic the human perception pipeline wherein the incoming audio is first filtered and then processed. Such a process. would eliminate artifacts that are carried. over from the TF mask filtering and generate a clean vocal signal. The synthesized voice signal can then be processed through tools like *Melodyne* [19] to allow for remixing and creative applications.

The artificial intelligence for music production research group, *DADABOTS*[20], has used this methodology for creating a clean voice mixture of the Opeth song, Ghost of Perdition, available on the streaming channel, YouTube: https://www.youtube.com/watch?v=XC8XDuT0RBE. The original song uses growling vocals, discussed in Section 1.2.1, while the remixed song uses a clean synthesized version of the same vocals, extracted through the proposed methodology. We note that growling based vocals do not have an inherent f0 and a single tone f0 corresponding to the note *middle D* or 293.665 Hz. We believe that generative modelling can be used for generating the f0 curve for such signals and plan to investigate this in the future.

11.2 Choral transcription and practice tool

We propose the framework for transcription and remixing of a full choir, with unison singing. Shown in Figure 11.1, our proposed framework allows for modulations of the individual parts of the choir, which can be used for practice, along with pitch estimation for each of the parts. To this end, we combine the research presented in Chapters 9 and 10. We propose to use the Open-Unmix (Stöter et al., 2019) model trained on all four datasets for separating the individual parts from an input. On a full choral recording, this would result in 4 parts with possible unison singing within each of the parts. We use the CREPE (Kim et al., 2018a) model to extract a representation of the single perceived pitch of the unison, as explained in Chapter 10 and use the STU model

[19] https://www.celemony.com/en/melodyne/what-is-melodyne
[20] https://dadabots.com/

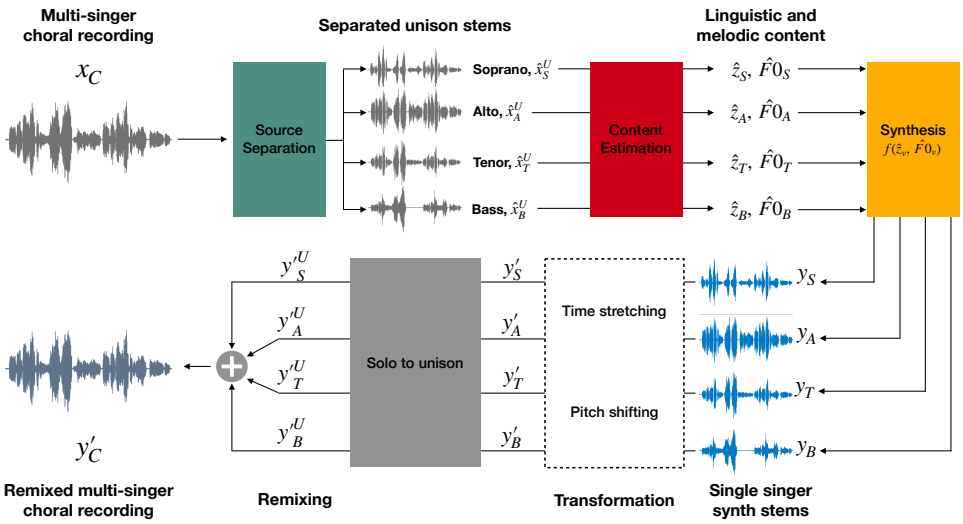

Figure 11.1: The proposed framework for the choir practice tool.

presented in the same chapter to synthesize a single voice signal representative of the content present in the unison. This signal can be modified, time-stretched or pitch-shifted according to the needs to the user and remixed using the STU methodology. We present the code for this application on a GitHub repository, https://github.com/MTG/SingingChoralSepAnalyzeSynthRemix, as well as a GoogleCollab notebook at https://colab.research.google.com/drive/1VnB2gtIDZIy31sZIt0PMeokcvIgaRafT?usp=sharing.

11.3 Application to other musical elements

We apply the analysis and synthesis framework to propose generative models for single shot percussive sounds (Ramires et al., 2020) and loops (Chandna et al., 2021). Percussive sounds are an important inharmonic component of modern music and are generally created by striking a hard surface. We present a feedforward model using the Wave-U-Net (Stoller et al., 2018) model to map intuitive control parameters to the waveform of percussive sounds. The control parameters are based on a set of timbrel features. These features were identified by user queries, used used for searching for

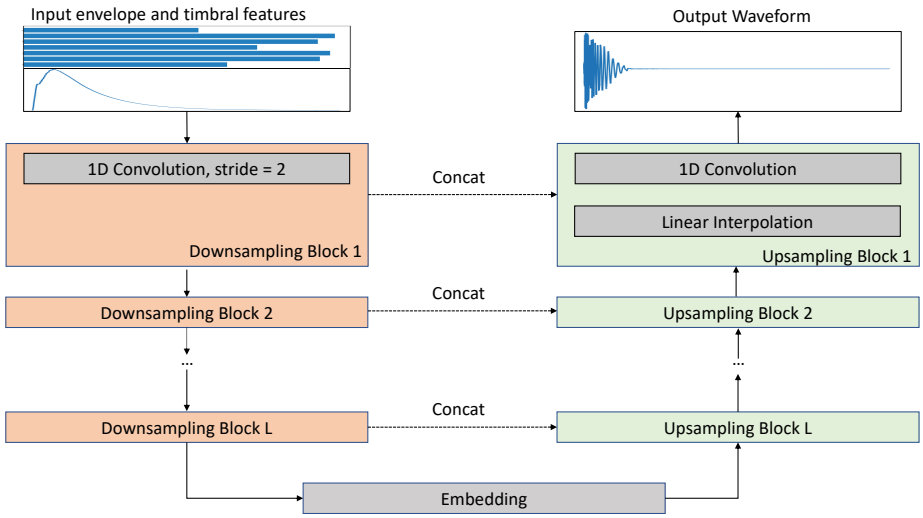

Figure 11.2: The architecture used for percussive synthesis.

sounds in large audio databases (Pearce et al., 2017). These features include hardness, depth, brightness, roughness, boominess, warmth and sharpness. We map these features, along with the envelope of the sound to the waveform of the sound. For this study, we curated a set of 10000 percussive one-shot sounds collected from Freesound (**?**). The source code for our model is available online[21], as are sound examples[22], showcasing the robustness of the models. We used a Wave-U-Net based architecture to map the input features directly to the waveform of the percussive sound, as shown in Figure 11.2.

Loops are seamlessly repeatable musical segments that are typically used in modern music production, that have lowered the barrier to entry into music making. We present a model similar to the one-shot percussive sound generator to generate loops using the same timbrel features along with Harmonic Pitch Class Profiles (HPCP) (Gómez Gutiérrez, 2006) and rhythm features extracted using an Automatic Drum Transcription algorithm (Southall et al., 2017). We curated a set of 8838 loops from

[21] https://github.com/pc2752/percussive_synth
[22] https://pc2752.github.io/percussive_synth/

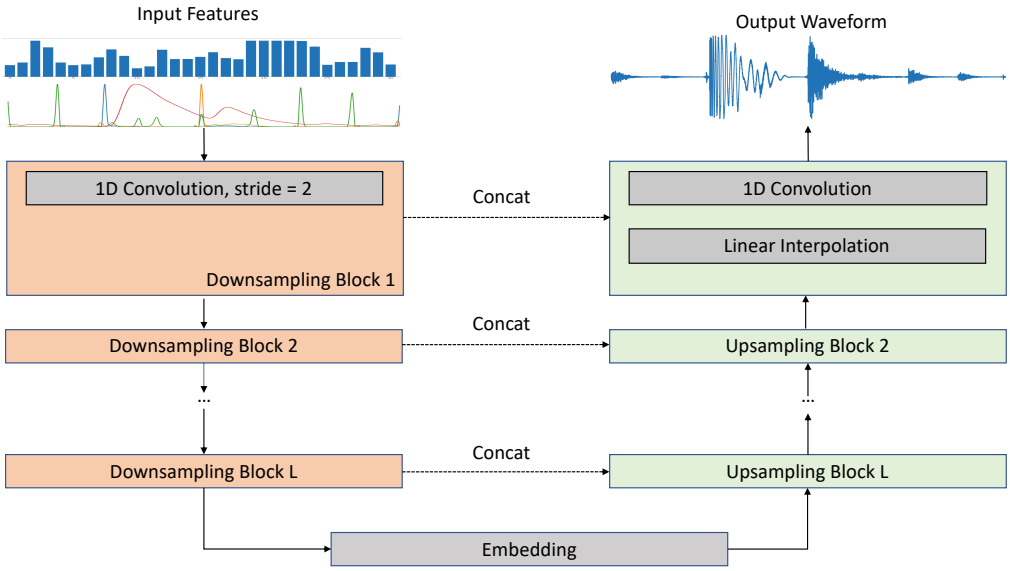

Figure 11.3: The architecture used for loop synthesis.

a community database, Looperman[23], for this task. The source code for this application is available at github.com/aframires/drum-loop-synthesis. The architecture for this model is shown in Figure 11.3

[23] looperman.com

Chapter 12

Conclusions

This thesis presents research on processing the singing voice signal in monaural polyphonic music signals in two contexts; contemporary popular music that contains processed vocals with instrumental accompaniment and ensemble choral singing that contains multiple voices singing in harmony and unison. Our research exploits recent advancements in data-driven deep learning based methodologies that provide a powerful tool for data modelling.

In Part II, we observe that deep learning unsupervised learning via autoencoders can be used to learn abstract yet useful representations of linguistic content from a voice signal. Such representations are often used for voice conversion, which involves converting the perceived speaker of a voice signal while retaining the perceived linguistic content. The melodic content of a singing voice signal is often represented by the fundamental frequency f0 curve, which encapsulates both the pitch and rhythm of the melody. We propose the framework for a methodology to extract these features from a musical mixture and use these, along with a representation of the singer identity to synthesize the clean singing voice signal. We use a one-hot representation of the singer identity and also train a network to extract the singer identity directly from the mixture signal. We note that the presented methodology is quite robust, even for examples outside the training set and with extreme vocal effects like growls, which completely

change the harmonic structure of the voice signal. However, we also note that the synthesis quality can be improved further.

Part III of the thesis explores part separation for one of the most popular formats of choral singing, the soprano, alto, tenor and bass (SATB) format. This format has four parts, each of which might have multiple singers singing in unison, while the parts themselves are arranged in harmony. The four parts in harmony are usually a linear mixture and can thus be separated using mask based filtering algorithms such as those commonly used for musical and source separation.

We curate data, combing multiple recently recorded datasets and cleaning them. We note that even after this consolidation, the data we have is quite limited when compared to the data used for musical or speech source separation. We then adapt four state-of-the-art deep learning based algorithms to separate the individual parts from an SATB mixture. We note that waveform based models perform as well as spectrogram based source separation models and that quartet based data is effective for data augmentation for training the models to separate the parts, even when they contain multiple voices in unison. We also observe that algorithms proposed for the task of music source separation are better suited to part separation for SATB choirs than algorithms proposed for asynchronous speech source separation.

While the unison is a linear mix of individual voices, the singers in the unison are singing the same content simultaneously and are hard to distinguish. Past research has shown that a single pitch is perceived from this unison signal. As such, it is impossible to separate the individual singers within the unison signal using filter based methods. However, since the perceived content of the unison signal is consistent, we propose a framework to synthesize a prototypical single voice signal from the unison that can encapsulate the linguistic and melodic content of the signal. We use our previously proposed methodology to this end. We also extend the methodology to synthesize a unison signal from an a capella singing voice input, based on the analysis of real world choral singing examples. We note that both timing and pitch deviations are necessary

to generate the perception of unison.

Given the presented research, we are now in a position to answer the research questions presented in Section 1.3.

- Can a singing voice signal be synthesized from a musical mixture by using language independent representations of the perceived content of the signal?

  Part II presents a framework for singing voice synthesis given a polyphonic mixture signal containing a single voice that can be processed with effects or a unison of multiple voices singing simultaneously. This framework is inspired by the pipeline a human listener would follow while trying to replicate the singing voice signal in a song. To this end, we note that the singing voice signal comprises of three elements; the singer-independent linguistic and melodic content and the singer-specific timbre. We also note that low-dimensional invertible representations called vocoder parameters are commonly used for analysis and synthesis of the singing voice signal. In this thesis, we use the WORLD (Morise et al., 2016) vocoder, with the dimensions of the harmonic and aperiodic components reduced to 60 and 4, respectively. Such compressed features have been shown to be effective for synthesis of a singing voice signal (Blaauw & Bonada, 2016) and are referred to as the compressed spectral envelope throughout this thesis.

  - Is it possible to extract synthesis parameters pertaining to the singing voice from a polyphonic contemporary music mixture?

    In Chapter 5, we propose a deep learning based methodology to estimate synthesis parameters from the magnitude component of the spectrogram of a musical mixture. We use a non-autoregressive WaveNet based temporal convolutional neural network to estimate the compressed spectral envelope from a polyphonic mixture of popular music containing vocals. We use the magnitude component of the spectrogram of the mixture signal as input to the network, which is optimized to minimize the mean absolute error

(MAE) loss between the target compressed spectral envelope and the output of the network. The estimated spectral envelope is fed, along with the magnitude component of the spectrogram of the mixture, to another non-autoregressive TCN, to estimate the f0 of the singing voice present in the mixture signal. The f0 is representative of the melodic content of the voice signal and the target is expressed in a logarithmic format. The output of these two networks is used to compute the voiced or unvoiced nature of each frame corresponding to the voice signal. The estimated compressed spectral envelope, f0 and voiced/unvoiced parts are used to synthesize the clean singing voice signal present in the mixture.

We compare the proposed algorithm to an NMF based algorithm as well as a deep learning based algorithm for TF mask estimation proposed by us earlier, through both objective and subjective evaluation. We observe that the proposed algorithm surpasses both the NMF and the deep learning based algorithms in terms of isolation of the voice signal from the source. Although we note that the quality of the synthesized voice signal leaves room for improvement, the presented methodology should that neural networks can be used to estimate voice synthesis parameters from a musical mixture.

- How can the voice signal extracted using such a methodology be evaluated?

  The signal generated by the proposed methodology is synthesized and is dependent on the accuracy of the parameter and f0 algorithms. While the perceptual qualities of the synthesized voice signal are quite similar to that of the signal used in the mixing process, the signals themselves are quite different. On objective evaluation of our proposed methodology, we found that typically used objective metrics for source separation (Vincent et al., 2006), particularly the SDR and SAR metrics, did not agree with the subjective evaluation we carried out via listening tests. As discussed in Sec-

tion 3.3.2, a number of perceptually motivated evaluation metrics have been proposed for both source separation and voice synthesis. However, the correlation of these metrics with subjective evaluation is still a matter of debate (Cano et al., 2016). Also, some of the metrics, particularly for evaluation of voice synthesis were proposed during the later stages of our research and are tailored to speech signals rather than the singing voice signal. We believe that such metrics can be adapted for the evaluation of our proposed methodology in the future, but for the context of this thesis, we continue with the use of subjective evaluation via listening tests.

- How can a feedforward neural network be used for singing voice synthesis given an input of linguistic content, singer identity and the f0 curve?

  In Chapter 6, we propose a feedforward U-Net based architecture (Ronneberger et al., 2015) to generate the compressed spectral envelope from an input consisting of the linguistic information, in the form of phonetic annotations, melodic information, in the form of the f0 curve and the identity of the singer as a one-hot vector representation. Randomly sampled noise is concatenated along with this input and the network is optimized via the Wasserstein GAN (Arjovsky et al., 2017) adversarial training methodology. We use the NUS corpus (Duan et al., 2013) for training the model, which is a relatively small dataset compared to those typically used for speech synthesis.

  We compare the proposed methodology to a state-of-the-art singing voice synthesizer, the NPSS (Blaauw & Bonada, 2016), that uses an autoregressive framework to generate vocoder features similar to the ones we use for our model. Through subjective and objective evaluation, we find that the methodology we propose is comparable to the autoregressive singing voice synthesizer. We note that since we proposed the singing voice synthesis system, a number of methodologies for singing voice synthesis based on

feedforward neural networks have been proposed (Hono et al., 2019; Lee et al., 2019a; Blaauw et al., 2019). Some of these methodologies use seq2seq modelling to generate the singing voice signal end to end, without the need for aligned phonetic annotations.

- Can the linguistic content of a singing voice signal be represented in a language independent manner from which a voice signal can be synthesized?

  While synthesis of a singing voice signal is possible given a phonetic representation of linguistic content, extracting such a representation from a polyphonic music mixture is a challenging task. Deep learning based methodologies proposed for the extraction require a large amount of annotated training data (Demirel et al., 2020). Furthermore, the use of such a representation imposes a language constraint on the system. Exploring zero resource synthesis and voice conversion, we find that voice synthesis is possible using an abstract representation of the linguistic content in a voice signal. Autoencoder based models are used in both fields to extract meaningful representations from the voice representations by imposing constraints like domain confusion, bottleneck restriction and vector quantization on the latent embedding of the autoencoder. Such constraints force the encoder of the autoencoder to disentangle speaker specific information like timbre from speaker independent information like prosody and the linguistic content.

  We adapt one such autoencoder based voice conversion methodology, the AutoVC (Qian et al., 2019), for singing voice conversion. The AutoVC model uses a bottleneck constraint to force the latent embedding of the autoencoder to learn linguistic and prosodic content from the mel-spectrogram of a speech signal. We adapt the model for singing voice conversion by using the compressed spectral envelope as input. Doing so allows us to disentangle the f0 from the timbre of the voice signal. This allows the system

to retain the melody of the signal while converting the timbre. By doing so, the adapted model is able to disentangle the singer independent linguistic content from the singer specific timbre and the melody. This provides us with an abstract representation of the linguistic content from which the compressed spectral envelope can be regenerated. This also averts the linguistic limitations that are imposed by the use of explicit phoneme annotations. Such abstract representations provide a useful medium for analysis and synthesis.

- Is it possible to extract such a representation of the linguistic content from a polyphonic contemporary music mixture?

  In Chapter 7, we use the AutoVC model, adapted for singing voice conversion, to train a network to extract the abstract linguistic content derived by the AutoVC from a mixture signal. The input to the model is the magnitude component of the spectrogram and it is trained to replicate the linguistic content derived by the adapted AutoVC model in a singer-independent manner. The decoder of this network uses the linguistic content and a one-hot representation of the singer identity to generate the compressed spectral envelope corresponding to the voice signal. Since the decoder network uses a representation of the singer identity, we term this model the singer dependent network (SDN).

- How can we derive a representation of the singer identity for the voice synthesis process?

  We further explore various methodologies for representation of the identity of the singer, proposing our own methodology for such. Our proposed methodology exploits the bottleneck constraint that is used by voice conversion systems for disentangling speaker independent content from the speaker specific content of a voice signal. Through subjective and objective evaluation, we observe that both the SDN and SIN networks improve over

the previously proposed model for extracting synthesis parameters from a mixture signal, particularly in terms of audio quality and intelligibility. This shows that the SDN is able to replicate the linguistic content derived by the adapted AutoVC model. The SIN surpasses the SDN in terms of audio quality, suggesting that the singer encoder used in this network derives more than just the singer identity from the mixture signal. However, we believe this to be acceptable for the use case in question. We also note that all three synthesis based models outperform one of the state-of-the-art TF mask based algorithms in terms of isolation of the voice signal from the instrumental accompaniment.

- What are the potential applications of such a methodology?

  We present a methodology to synthesize clean singing voice signals from contemporary popular music mixtures containing processed vocals. Doing so overcomes one of the biggest limitations of State-of-the-Art (SOTA) source separation algorithms that use TF masks to filter out the vocal signal from polyphonic mixtures. The proposed framework is agnostic to the specific algorithms used for the individual components, which draw from the fields of voice synthesis, polyphonic f0 estimation, singer identity representation and voice conversion. We believe that components within the framework can be easily replaced with newer, better components as research in the individual fields advances.

  As listed in Chapter 11, we believe the proposed methodology. can be used for artistic applications such as remixing, particularly when used in conjunction with TF mask based source separation algorithms. The methodology can also lead to improvements in automatic lyrics transcription (ALT) as a pre-processing step for removing hindering effects and artifacts from processed vocal signals to be transcribed. We also use the analysis-synthesis approach for synthesis of other musical elements like drums and

loops.

- Can the individual voices in an ensemble choral singing recording be separated, given limited training data?

  In Part III of the thesis, we adapt and evaluate some of the SOTA source separation algorithms proposed for musical and speech source separation to separate the individual parts of an SATB mixture. This research is presented in Chapter 9. We adapt two spectrogram based model, the U-Net (Jansson et al., 2017) and the Open-Unmix (Stöter et al., 2019) and two waveform based models, the Wave-U-Net (Stoller et al., 2018) and Conv-TasNet (Luo & Mesgarani, 2019) to separate the individual soprano, alto, tenor and bass parts from an SATB mixture. The first three models listed were originally proposed for musical source separation, i.e., to separate the vocals, drums, bass and other instruments stems from a contemporary popular music mixture. Conv-TasNet was originally proposed for separating asynchronous speech signals from a mixture of the same.

  In addition, we propose analyze unison singing recordings within the choral singing dataset (CSD) (Cuesta et al., 2019) and use the insight gained to propose a methodology for synthesizing a single voice signal representative of the content of the unison singing signal.

  - Can waveform based source separation algorithms work as well as spectrogram based models for choral part separation?

    Through objective evaluation of the adapted models, using the blind source separation evaluation (BSS Eval) (Vincent et al., 2006), we observe that the performance of the spectrogram based U-Net (Jansson et al., 2017) model is quite similar to that of its waveform based equivalent, the Wave-U-Net (Stoller et al., 2018). The hyperparameters used for training the model were quite similar and we used a single model to predict all four parts to be separated, allowing for a direct comparison between the two. As source

separation methodologies evolve towards end-to-end waveform based separation (Lluís et al., 2019), we believe that advanced models proposed we can effectively be used for part separation in choral SATB recordings, even with the dearth of data available for this domain.

- Are music source separation algorithms better suited to choral part separation or should speech source separation algorithms be used?

  Objective evaluation for the algorithms showed that the source separation algorithms proposed originally for music source separation performed better than the Conv-TasNet (Luo & Mesgarani, 2019), which was originally proposed for speech source separation. In particular, we note that the Open-Unmix (Stöter et al., 2019) model outperformed the other models. This was expected as the Open-Unmix model has shown superior performance to the other algorithms even for music source separation. We note that an adaptation of the Conv-TasNet model has been proposed for music source separation (Samuel et al., 2020; Défossez et al., 2019). However, in this study we use the version originally proposed for speech source separation.

- How can we curate data from varied datasets which has been recorded under different conditions?

  We note that the proposed models for source separation required large amounts of training data, while we had access to limited data in the form of the choral singing dataset (CSD) (Cuesta et al., 2019), Bach chorales dataset (BCD), ESMUC dataset and the Daghstuhl dataset (DSD) (Rosenzweig et al., 2020). Each of these datasets had been recorded in different settings and had different formats and leakage for the individual stems present. We note that the BCD had only single singer per part for the SATB songs, while the other datasets had multiple singers per part.

  We trained the models using two forms of data; *case*_1 wherein we used all possible combinations of quartets, i.e. limiting the number of singers to

1 per part, and *case*_2 where we removed the restriction on the number of singers per part and used all possible combinations for training. We also tested the models for data augmentation by first training with just the CSD dataset and then with the CSD and BCD datasets. On preliminary analysis, we found the Open-Unmix model trained on *case*_2 data to be the best performing of all the models and used this to clean the ESMUC dataset, which had a high amount of inter-singer leakage in the tracks. We used this cleaned ESMUC dataset for objective evaluation of the models.

- Can quartet based data with a single singer per part be used to train deep learning based algorithms for part separation even with multiple singers per part in unison?

  We evaluated the adapted models on the full choir songs from the ESMUC dataset, i.e. with unison singing present within the parts. Through objective evaluation, we note that there was a significant improvement in performance when the multiple combinations of singer from the CSD were augmented with quartet based single singer per part data from the BCD, indicating that additional quartet based data would be sufficient for improving the performance of the models in the future. This is significant, because recording quartets is an easier task than recording individual singers for the full choir. We believe that for future studies a combination of data from quartet based data and multiple singers per voice data can be used for training the source separation algorithms. We also note that pre-trained models can effectively be used for data cleaning, particularly for removing inter-singer leakage than might be encountered during the recording process.

- What are the perceptual qualities of a unison signal that distinguish it from a signal voice singing signal?

  In Chapter 10 we analyze unison signing. Unison singing involves multiple singers simultaneously singing the same linguistic and melodic con-

tent. Past research has shown that even though there are slight timing and pitch deviations amongst the singers in a unison, a listener listening to the unison will perceive a single pitch (Ternström, 1991). We analyze the unison signals present within the CSD and model the single perceived pitch by using the mean of the individual pitches. We use a deep learning based monophonic pitch extraction system to extract the perceived pitch of the unison signal and measure the accuracy of this extracted pitch to the theoretical pitch represented by the mean using the Raw Pitch Accuracy (RPA), Overall Accuracy (OA), Voicing Recall (VR) and Voicing False Alarm (VFA) metrics from the `mir_eval` library (Raffel et al., 2014). We observe that the extracted pitch is closer to the mean than to any of the individual pitches. We also measure the pitch and timing deviation between the singers in the unison and find an average timing deviation of 0.1 s and a pitch deviation in the range of 0 cents *to* 50 cents, with a mean of around 20 cents.

- Is it possible to separate a single voice from within the unison singing signal?

  We propose a framework to synthesize a prototypical single voice signal representative of the linguistic and melodic content of the unison signals, using the methodology proposed in Chapter 7 to extract the linguistic content and a single singer identity from the unison and the mean pitch as the perceived pitch. Using subjective analysis of the synthesized prototypical signal, we find that the mean is indeed representative of the perceived pitch of the unison. Through t-SNE visualisation, we observe that the singer identity extracted by our model falls within the cluster of singer identities for the target group.

  We also propose the framework for synthesize a unison signal from an a capella singing voice input by creating voice clones using voice conversion

with the same gender as the input. We also add timing and pitch deviations using the analysis described above. Using subjective evaluation we find that timing and pitch deviations together are necessary for creating the perception of unison whereas timbre variations do not change the perception much.

- How can choral source separation be useful in this context?

  Part separation as presented in Chapter 9 can be useful for remxining choir recordings for emphasis of a particular part. It can also be used for de-emphasizing a particular part for practice. Further, in Section 11.2, we explore the use of part separation for transcription of the individual parts of the choir. In conjunction with the methodology proposed in PartII, we propose a framework for analysis and modified synthesis of the individual parts of the choir. Modifications like time-stretching and pitch-shifting can be done on the re-synthesized signal allowing a student to practice parts within their capabilities.

12.1 Future work

In Part II of the thesis, we present the framework for synthesis of the singing voice signal from the underlying perceptual content present in a mixture. The framework consists of several components including; a singing voice f0 extraction system, a linguistic content representation and extraction component, representation of the singer identity and the vocoder parameters used for synthesis. We believe that improvements can be made to each of these individual components to improve on the quality and intelligibility of the synthesized signal generated. However, the framework proposed is agnostic to the exact methodology used in each component and each component can be supplemented by advancement models as research in the respective fields progresses. We also believe that a singing voice activity detection model can be used as a pre-processing step for the model, although on primary heuristic analysis, we have found

that the model outputs silence when a singing voice is not present in the input mixture signal.

TF mask based source separation algorithms can also be used in conjunction with the proposed methodology to allow for a more robust synthesis of the clean singing voice signal. We also plan to study the robustness of the proposed approach in various settings including different effects applied to the voice signal, varying mixing gains while creating the mixture and different languages. Further, we note that for many singing voice signals, particularly those mixed with effects such as those shown on our example website [24], there is no discernible f0. However, a human listener is still able to perceive a sense of a melody from the song. We believe that for such cases, generative models can be used to generate an f0 contour taking into account the harmonic features of the back track used.

In Part III of the thesis dissertation, we apply source separation algorithms for part separation in SATB choirs. We believe that these models can further be improved with more training data as further recordings of SATB choirs are produced. We also note that quartet data, which is easier to record than full choirs, is sufficient to improve system performance by data augmentation. We also observe that inter-singer leakage can be cleaned effectively by using one of the pre-trained models. This cleaned data can further be used for training the models.

As field music source separation algorithms continues to evolve towards end to end waveform based source separation, we believe that such models can effectively be adapted towards part separation for ensemble choir singing. As shown in our study, pre-trained models can be used for data cleaning of future datasets that might have either quartet based data or full choir data with inter singer leakage.

We also plan to explore further applications of the proposed framework for remixing of the re-synthesized prototypical signals, which can be modified and used for practice and training purposes.

[24] https://pc2752.github.io/sep_content/

Appendix A

Publications by the author

Full articles in peer-reviewed conferences

- **Chandna, P.**, Miron, M., Janer, J., & Gómez, E. (2017). Monoaural Audio Source Separation Using Deep Convolutional Neural Networks. In *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*

- **Chandna, P.**, Blaauw, M., Bonada, J., & Gómez, E. (2019). A vocoder based method for singing voice extraction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

- Cuesta H, **Chandna, P.**, & Gómez, E. (2019). A Framework for multi-f0 modeling in SATB choir recordings. In *16th Sound & Music Computing Conference (SMC)*

- **Chandna, P.**, Blaauw, M., Bonada, J., & Gómez, E. (2019). Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan. In *27th European Signal Processing Conference (EUSIPCO)*

- **Chandna, P.**, Blaauw, M., Bonada, J., & Gómez, E. (2019). Content based singing voice extraction from a musical mixture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

- Ramires, A., **Chandna, P.**, Favory, X., Gómez, E., & Serra, X (2020). Neural percussive synthesis parameterised by high-level timbral features. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

- **Chandna, P.**, Cuesta H, & Gómez, E. (2020). A deep learning based analysis-synthesis framework for unison singing. In *21st International Society for Music Information Retrieval Conference (ISMIR)*

- Petermann, D., **Chandna, P.**, Cuesta, H., Bonada, J., & Gómez. (2020). Deep learning based source separation applied to choir ensembles. In *21st International Society for Music Information Retrieval Conference (ISMIR)*

- **Chandna, P.**, Ramires, A., Serra, X, & Gómez, E. (2021). LoopNet: Musical Loop Synthesis Conditioned On Intuitive Musical Parameters. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*

Article submitted for review in peer-reviewed journals

- **Chandna, P.**, Cuesta H, Petermann, D., & Gómez, E (2021). A Deep-learning based Framework for Source Separation, Analysis, and Synthesis of Choral Ensembles. In *IEEE Transactions on Signal Processing*

Appendix B

Resources

The research presented in this thesis broadly follows the principles of research reproduciblity (Cannam et al., 2012). We provide open source code for each of the methodologies proposed in the thesis and trained models for the same. While we have tried to use public datasets and evaluation methodologies for our research, we note that some of the models have been trained on proprietary datasets that can be distributed freely. We also note that the listening tests used for evaluation of some of the algorithms are highly subjective and reflect the opinion of the participants, taking into account their biases and knowledge.

Code and Tools

The *Python* implementations of the methodologies presented in this thesis are made available through *GitHub* repositories, along with audio examples. The following list presents the various models presented during this thesis, along with the source code and sound examples for the same.

- **Separation-via-synthesis (SS)**: A vocoder based method for singing voice extraction, presented in Chapter 5.

  *Source code:*https://github.com/pc2752/ss_synthesis

*Audio examples:*https://ronggong.github.io/projects/pritish_mlslp_2018/demo.html

- **WGANSing**: A multi-voice singing voice synthesizer based on the wasserstein-gan, presented in Chapter 6.

  Source code: https://github.com/MTG/WGANSing

  Audio examples: https://pc2752.github.io/sing_synth_examples/

- **Singer dependent content based synthesis (SDN) and Singer independent content based synthesis (SIN)**: Models for content based singing voice extraction from a musical mixture, presented in Chapter 7.

  Source code: https://github.com/MTG/content_choral_separation

  Audio examples: https://pc2752.github.io/sep_content/

- **Solo to unison (STU) and unison to solo (UTS)**: Models for synthesizing a single voice signal from an unison input and vice versa, presented in Chapter 10.

  Source code: https://github.com/MTG/content_choral_separation

  Audio examples: https://pc2752.github.io/unison_analysis_synthesis_examples/

- **SATB processing**: A framework for source separation, moelodic estimation and re-synthesis of SATB recordings, presented in Chapter 11

  Source code: https://github.com/MTG/SingingChoralSepAnalyzeSynthRemix

  GoogleCollab notebook: https://tinyurl.com/43c2yv4s

Appendix C

Singing voice conversion subjective evaluation results

We compared the modified AutoVC (Qian et al., 2019) models explained in Section 7.2 against the Unsupervised Singing Voice Conversion (USVC) (Nachmani & Wolf, 2019) methodology proposed for singing voice conversion. This model uses an autoencoder based on the WaveNet (van den Oord et al., 2016a) architecture and imposes a domain confusion (Ganin et al., 2016) constraint pertaining to the singer identity on the latent embedding of the autoencoder. This allows the model to perform non-parallel singing voice conversion. However, the use of the WaveNet vocoder imposes some undesirable changes to the melody of the output signal.

The AutoVC models with GE2E and JE embeddings as well as the VQVC+ model can be used for Zero-Shot voice conversion, i.e. the source and target singers used for conversion do not necessarily have to be in the training set. The AutoVC model trained with one-hot vector encoding of singer representation and the USVC model do not fulfill this criteria and can only perform conversion within the singers used for training.

We used a proprietary dataset, described in Section 3.4, to train the Zero-Shot models and the NUS corpus (Duan et al., 2013) for evaluation. The USVC and the AutoVC

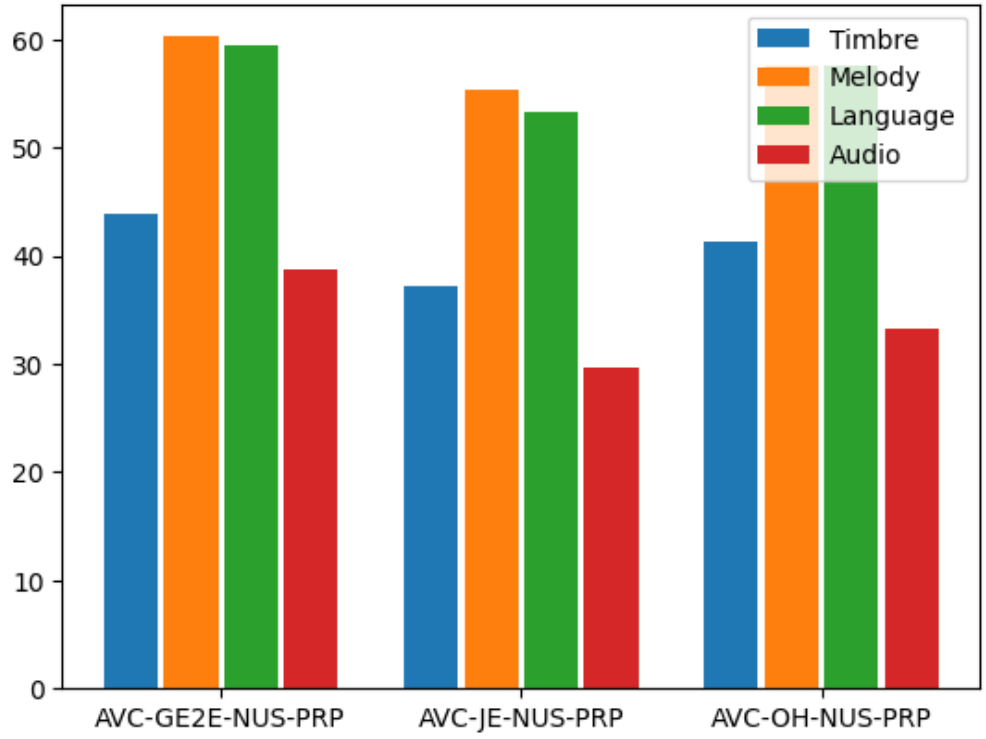

Figure C.1: MOS for subjective evaluation comparing the modified AutoVC (Qian et al., 2019) architecture using one-hot (OH) vector representations for singer identity with the same architecture using JE and GE2E embeddings for singer identity representation.

model with one-hot singer representation were trained and evaluated on the NUS dataset. We used a MOS based listening test to evaluate these models on four criteria including; the conversion of timbre between the source and target, the retention of melody, the intelligibility of the output and the overall audio quality of the output.

This work is published as the Masters' thesis of Pavlo Apisov (Apisov, 2020), carried out under the supervision of the author in 2020.

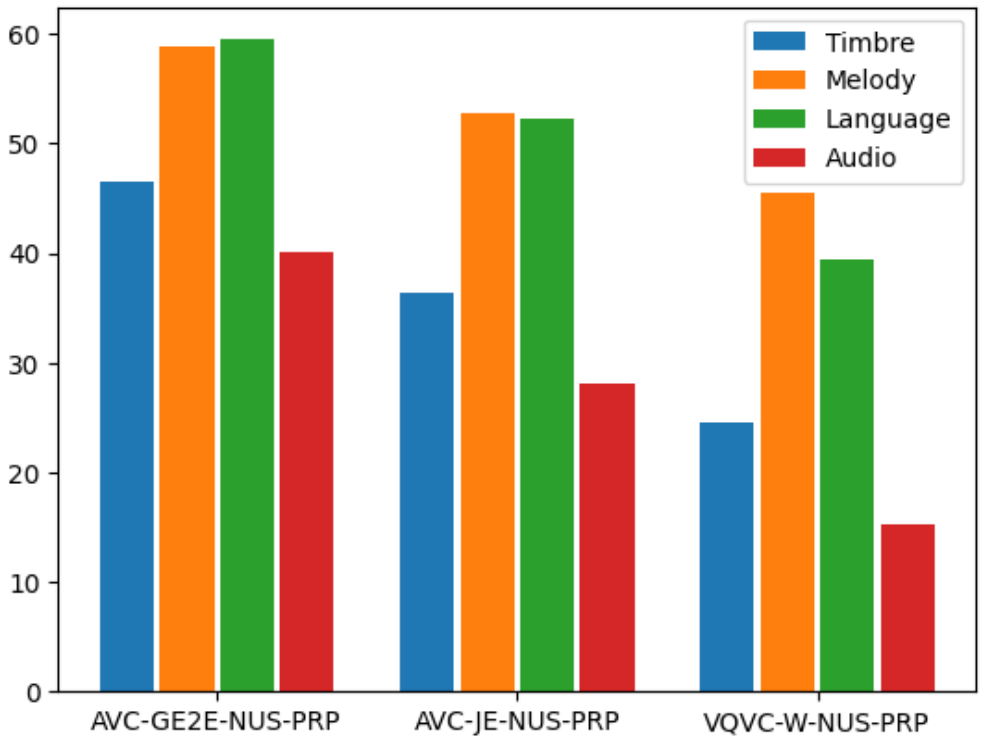

Figure C.2: MOS for subjective evaluation comparing the modified AutoVC architecture using JE and GE2E embeddings for singer identity representation with the modified VQVC+(Wu et al., 2020; Wu & Lee, 2020) model

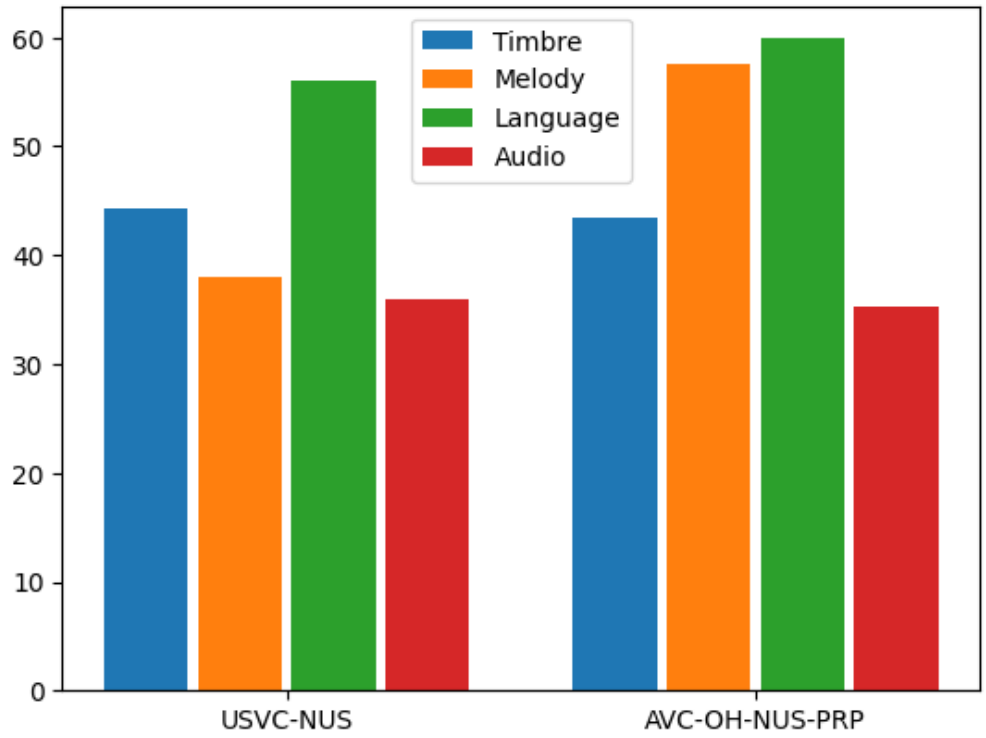

Figure C.3: MOS for subjective evaluation comparing the modified AutoVC architecture using one-hot (OH) vector representations for singer identity with the USVC model

Appendix D

Choral part separation conditioned on f0

We propose an adaptation of the conditioned U-Net model (Meseguer-Brocal & Peeters, 2019) for the task of part separation for SATB choirs. For evaluation of this model, we use the Oracle f0 values for each of the parts in an SATB choir, with the mean f0 of the individual singers used as the representative f0 in the case of unison mixture, as described in Section 10. We propose three variants of the model, as shown in Figure D.1 with local conditioning and global conditioning of the f0 using one-hot encoding. We train these models on the CSD and BCD datasets and evaluate them on a subset of the same that was withheld from training. The model is compared with the Wave-U-Net (Stoller et al., 2018) model, the U-Net (Jansson et al., 2017) model, the Open-Unmix (Stöter et al., 2019) and the original conditioned U-Net model (Meseguer-Brocal & Peeters, 2019).

The evaluation of these models with the BSS eval set of metrics is shown in Tables D.1, D.3 and D.1. Further evaluation with the PEASS set of metrics (Vincent, 2012; Emiya et al., 2011) is shown in Table D.4.

This work is published as the Masters' thesis of Darius Peterman (Pétermann, 2020), carried out under the supervision of the author in 2020.

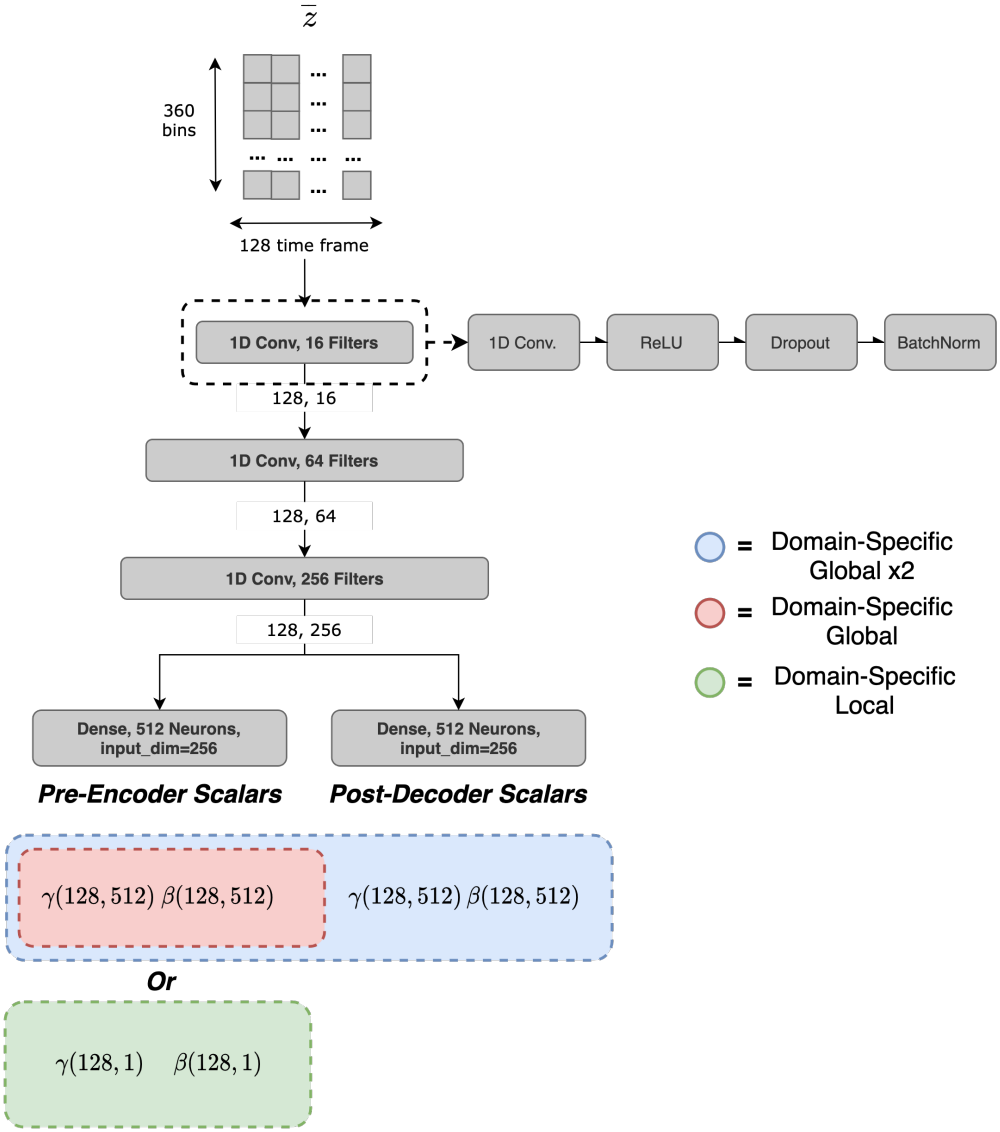

Figure D.1: The three variants of the control model architecture for three of our four proposed models. The convolution is performed across the frequency bins (treated as feature channels) for each time-step. At the output stage the dense layer(s) provides various numbers of scalars. *Local* conditioning embeds the target source's f0 into 2 scalars per time-step. *Global* conditioning codifies the f0 into a set of scalars for each frequency bin per input time-step. Lastly *Global x2* conditioning does it at both input and output levels.

Model	**Test Use-Case 1 - SDR (dB)**				
	Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	2.03±2.2	4.59±2.7	0.92±2.9	2.72±2.5	2.56±2.3
U-Net	3.78±2.1	5.15±3.7	2.29±2.7	3.22±1.5	3.61±2.5
C-U-Net D-A	3.57±2.0	2.05±2.1	-1.25±2.6	1.96±2.2	1.58±2.2
Open-Unmix	5.61±2.1	5.70±2.3	1.60±1.7	3.66±2.2	4.14±2.1
C-U-Net D-S L	3.70±1.3	6.99±1.9	3.82±1.6	3.74±1.7	4.56±1.6
C-U-Net D-S G	**5.76**±1.2	**7.67**±1.5	**5.39**±1.4	**4.07**±1.8	**5.73**±1.5
C-U-Net D-S G x2	3.46±1.4	5.30±1.6	1.81±1.7	1.56±1.5	3.03±1.6
C-U-Net D-S Enc	3.05±1.4	5.97±2.1	3.35±1.9	2.99±1.3	3.84±1.7

Model	**Test Use-Case 2 - SDR (dB)**				
	Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	3.30±1.6	4.73±0.8	2.09±2.0	1.24±1.4	2.84±1.5
U-Net	5.14±1.5	6.63±1.0	4.74±1.7	3.12±1.6	4.91±1.4
C-U-Net D-A	4.61±1.8	2.67±2.7	0.52±2.8	1.98±1.6	2.45±2.2
Open-Unmix	**6.67**±2.1	6.49±1.3	2.70±1.6	**3.49**±2.0	4.83±1.7
C-U-Net D-S L	4.34±0.9	**7.06**±1.2	4.77±1.6	3.48±1.5	4.91±1.3
C-U-Net D-S G	5.34±1.2	6.44±1.4	**4.93**±1.5	3.18±1.1	**4.97**±1.3
C-U-Net D-S G x2	4.58±1.4	5.51±1.1	3.45±2.1	2.62±1.2	4.04±1.4
C-U-Net D-S Enc	4.53±1.5	6.57±1.3	4.65±1.6	2.98±1.5	4.68±1.5

Table D.1: SDR (signal-to-distortion) results mean±std on the four SATB parts and their overall average for the four domain-agnostic architectures as well as for our four proposed domain-specific models. The top table depicts the results obtained from the first use-case test set while the bottom ones are from the second use-case test set.

Model	Test Use-Case 1 - SIR (dB)				
	Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	5.99±2.4	9.19±2.9	4.62±2.1	8.49±3.5	7.07±2.7
U-Net	10.28±2.4	10.77±4.1	6.70±3.2	9.45±2.0	9.30±2.9
C-U-Net D-A	10.09±2.6	7.81±1.6	3.32±2.2	7.61±2.4	7.21±2.4
Open-Unmix	12.36±2.7	13.19±2.9	6.43±2.0	**11.41**±2.6	10.85±2.6
C-U-Net D-S L	9.71±1.7	12.37±1.5	9.89±2.2	9.71±1.7	10.42±1.8
C-U-Net D-S G	**12.72**±1.8	**14.04**±1.5	**11.79**±2.1	9.78±2.1	**12.08**±1.7
C-U-Net D-S G x2	10.03±2.5	11.01±1.1	8.22±1.8	7.60±2.1	9.21±1.9
C-U-Net D-S Enc	9.41±1.9	11.95±1.5	9.86±1.8	9.74±2.1	10.24±1.8

Model	Test Use-Case 2 - SIR (dB)				
	Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	8.13±2.1	10.02±0.9	6.80±2.2	7.45±2.0	8.10±1.8
U-Net	12.41±1.8	13.11±1.2	10.26±1.1	8.50±2.3	11.07±1.6
C-U-Net D-A	11.99±1.8	9.08±2.8	5.65±3.1	7.60±2.0	8.58±2.4
Open-Unmix	**14.71**±2.8	**14.40**±1.5	7.95±2.4	**10.67**±1.8	**11.93**±2.1
C-U-Net D-S L	10.32±1.1	13.06±1.7	10.77±1.5	8.89±2.2	10.76±1.6
C-U-Net D-S G	12.08±1.5	13.50±2.5	**12.05**±1.3	8.91±2.1	11.63±1.8
C-U-Net D-S G x2	10.86±1.7	11.22±2.3	10.20±2.2	8.30±2.6	10.15±2.2
C-U-Net D-S Enc	11.15±1.6	13.43±2.0	11.62±1.3	9.19±2.7	11.28±1.9

Table D.2: SIR (signal-to-interference) results mean±std on the four SATB parts as well as their overall average for the four domain-agnostic architectures as well as for our four proposed domain-specific models. The top table depicts the results obtained from the first use-case test set while the bottom ones are from the second use-case test set.

Model	Test Use-Case 1 - SAR (dB)				
	Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	5.36±1.7	7.11±2.4	4.79±1.4	4.89±1.4	5.50±1.7
U-Net	5.35±1.8	7.13±3.0	5.32±1.8	4.94±1.1	5.69±1.9
C-U-Net D-A	5.19±1.6	4.41±2.8	2.65±1.2	4.12±2.0	4.09±1.9
Open-Unmix	7.00±1.8	6.84±2.0	4.43±1.2	4.83±1.9	5.75±1.7
C-U-Net D-S L	5.44±1.0	8.75±2.0	5.58±1.3	5.51±1.7	6.32±1.5
C-U-Net D-S G	**7.02**±1.1	**9.02**±1.6	**6.86**±1.5	**5.93**±1.6	**7.21**±1.4
C-U-Net D-S G x2	5.08±0.9	7.02±1.7	3.65±1.7	3.58±1.1	4.83±1.4
C-U-Net D-S Enc	4.74±1.0	7.55±2.2	4.94±1.9	4.54±0.9	5.43±1.5

Model	Test Use-Case 2 - SAR (dB)				
	Soprano	Alto	Tenor	Bass	Avg.
Wave-U-Net	5.75±1.0	6.73±1.0	4.79±1.5	3.23±0.9	5.13±1.1
U-Net	6.31±1.4	7.97±1.0	**6.59**±1.8	5.27±1.1	6.54±1.3
C-U-Net D-A	5.77±1.7	4.60±2.9	3.39±1.8	4.15±1.1	4.48±1.9
Open-Unmix	**7.60**±1.8	7.43±1.3	5.01±1.1	4.80±1.8	6.21±1.5
C-U-Net D-S L	6.02±0.9	**8.59**±1.2	6.45±1.7	**5.50**±1.1	**6.66**±1.2
C-U-Net D-S G	6.68±1.2	7.70±1.3	6.17±1.6	5.17±0.8	6.43±1.2
C-U-Net D-S G x2	6.12±1.1	7.45±1.4	4.97±2.0	4.79±0.5	5.83±1.3
C-U-Net D-S Enc	5.95±1.4	7.84±1.2	6.04±1.7	4.82±1.0	6.16±1.3

Table D.3: SAR (signal-to-artifacts) results mean±std on the four SATB parts as well as their overall average for the four domain-agnostic architectures as well as for our four proposed domain-specific models. The top table depicts the results obtained from the first use-case test set while the bottom ones are from the second use-case test set.

Model	PEASS Scores			
	OPS	TPS	APS	IPS
Wave-U-Net	**27.50**±6.00	51.14±14.24	0.16±0.16	**84.46**±2.62
U-Net	19.54±3.17	48.87±12.68	0.53±0.51	82.07±2.39
C-U-Net D-A	22.69±5.50	4.65±3.70	1.71±2.18	79.91±2.79
Open-Unmix	22.58±4.05	**52.42**±11.93	0.42±0.41	83.52±2.13
C-U-Net D-S L	17.67±2.66	11.07±6.68	**5.91**±4.86	80.23±2.21
C-U-Net D-S G	16.23±2.64	11.00±5.55	5.67±4.42	81.21±1.80
C-U-Net D-S G x2	18.70±3.05	8.53±5.38	4.47±3.87	81.34±1.75
C-U-Net D-S Enc	17.80±3.08	8.92±5.53	4.67±4.44	80.96±1.98

Table D.4: Overall TPS, IPS, OPS, and APS means±std across all parts and use-cases.

Bibliography

Afouras, T., Chung, J. S., & Zisserman, A. (2018). The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*. [Cited on page 51.]

Akagi, M. & Kitakaze, H. (2000). Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours. In *Sixth International Conference on Spoken Language Processing*. [Cited on page 17.]

Apisov, P. (2020). Zero-shot singing voice conversion. In *Masters' thesis, Universitat Pompeu Fabra*. [Cited on pages 138 and 212.]

Araki, S., Nesta, F., Vincent, E., Koldovskỳ, Z., Nolte, G., Ziehe, A., & Benichoux, A. (2012). The 2011 signal separation evaluation campaign (sisec2011):-audio source separation. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 414–422. Springer. [Cited on pages 52 and 86.]

Arik, S., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., & Zhou, Y. (2017a). Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *arXiv preprint arXiv:1705.08947*. [Cited on page 66.]

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., & Others (2017b). Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 195–204. JMLR. org. [Cited on page 65.]

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. [Cited on pages 46, 122, 123, 127, 128, and 197.]

Arora, V. & Behera, L. (2015). Multiple f0 estimation and source clustering of polyphonic music audio using plca and hmrfs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23*(2), 278–287. [Cited on page 81.]

Atal, B. S. (2018). From vocoders to code-excited linear prediction: Learning how we hear what we hear. In *Proc. Interspeech 2018*, p. 1. [Cited on page 60.]

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015. [Cited on page 43.]

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49. JMLR Workshop and Conference Proceedings. [Cited on page 40.]

Banno, H., Hata, H., Morise, M., Takahashi, T., Irino, T., & Kawahara, H. (2007). Implementation of realtime straight speech manipulation system: Report on its first implementation. *Acoustical science and technology*, *28*(3), 140–146. [Cited on page 61.]

Beerends, J. G., Schmidmer, C., Berger, J., Obermann, M., Ullmann, R., Pomy, J., & Keyhl, M. (2013). Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment. *Journal of the Audio Engineering Society*, *61*(6), 366–384. [Cited on page 93.]

Beerends, J. G. & Stemerdink, J. A. (1992). A perceptual audio quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, *40*(12), 963–978. [Cited on page 92.]

Beerends, J. G. & Stemerdink, J. A. (1994). A perceptual speech-quality measure based on a psychoacoustic sound representation. *Journal of the Audio Engineering Society*, *42*(3), 115–123. [Cited on page 92.]

Ben-Shalom, A. & Dubnov, S. (2004). Optimal filtering of an instrument sound in a mixed recording given approximate pitch prior. In *ICMC*. [Cited on page 104.]

Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., & Simonyan, K. (2019). High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*. [Cited on pages 93 and 145.]

Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*. [Cited on page 93.]

Bittner, R., Salamon, J., Essid, S., & Bello, J. (2015). Melody extraction by contour classification. In *International Conference on Music Information Retrieval (ISMIR)*. [Cited on page 80.]

Bittner, R. M. & Bosch, J. J. (2019). Generalized metrics for single-f0 estimation evaluation. In *ISMIR*, pp. 738–745. [Cited on page 114.]

Bittner, R. M., McFee, B., Salamon, J., Li, P., & Bello, J. P. (2017). Deep salience representations for f0 estimation in polyphonic music. In *ISMIR*, pp. 63–70. [Cited on pages 81 and 112.]

Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014). Medleydb: A multitrack dataset for annotation-intensive mir research. In *ISMIR*, vol. 14, pp. 155–160. [Cited on pages 86, 97, and 141.]

Blaauw, M. & Bonada, J. (2016). Modeling and transforming speech using variational autoencoders. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. [Cited on pages XXI, XXII, 61, 62, 67, 68, 93, 107, 109, 112, 121, 126, 128, 130, 195, and 197.]

Blaauw, M. & Bonada, J. (2017). A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs. *Applied Sciences*, *7*(12), 1313. [Cited on pages XXI, XXV, 61, 64, 67, 68, 107, 109, 110, 112, and 131.]

Blaauw, M., Bonada, J., & Daido, R. (2019). Data Efficient Voice Cloning for Neural Singing Synthesis. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [Cited on pages XXI, 69, 107, 132, and 198.]

Black, A. W., Zen, H., & Tokuda, K. (2007). Statistical parametric speech synthesis. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, pp. IV–1229–IV–1232. [Cited on page 58.]

Black, D. A., Li, M., & Tian, M. (2014). Automatic identification of emotional cues in chinese opera singing. *ICMPC, Seoul, South Korea*. [Cited on page 88.]

Bonada, J. & Blaauw, M. (2020). Hybrid neural-parametric f0 model for singing synthesis. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7244–7248. [Cited on page 121.]

Bonada, J., Umbert Morist, M., & Blaauw, M. (2016). Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016. *Morgan N, editor. Interspeech 2016; 2016 Sep 8-12; San Francisco, CA.[place unknown]: ISCA; 2016. p. 1230-4.* [Cited on page 64.]

Bosch, J. J., Bittner, R. M., Salamon, J., & Gómez Gutiérrez, E. (2016). A comparison of melody extraction methods based on source-filter modelling. In *17th International Society for Music Information Retrieval Conference; 2016 Aug 7-11; New York, United States.* International Society for Music Information Retrieval (ISMIR). [Cited on page 80.]

Bosch, J. J., Janer, J., Fuhrmann, F., & Herrera, P. (2012). A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals. In *ISMIR*, pp. 559–564. [Cited on pages 87 and 88.]

Bous, F. & Roebel, A. (2019). Analysing deep learning-spectral envelope prediction methods for singing synthesis. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE. [Cited on page 67.]

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press. [Cited on pages 18, 19, and 23.]

Brown, G. J. & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech & Language*, *8*(4), 297–336. [Cited on page 20.]

Bugler, A., Pardo, B., & Seetharaman, P. (2020). A study of transfer learning in music source separation. *arXiv preprint arXiv:2010.12650*. [Cited on page 159.]

Camacho, A. & Harris, J. G. (2008). A sawtooth waveform inspired pitch estimator for speech and music. *The Journal of the Acoustical Society of America*, *124*(3), 1638–1652. [Cited on page 79.]

Cañadas Quesada, F., Ruiz Reyes, N., Vera Candeas, P., Carabias, J. J., & Maldonado, S. (2010). A multiple-f0 estimation approach based on gaussian spectral modelling for polyphonic music transcription. *Journal of New Music Research*, *39*(1), 93–107. [Cited on page 81.]

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, *58*(3), 1–37. [Cited on pages 20, 36, and 157.]

Cannam, C., Figueira, L. A., & Plumbley, M. D. (2012). Sound software: Towards software reuse in audio and music research. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2745–2748. IEEE. [Cited on page 209.]

Cano, E., FitzGerald, D., & Brandenburg, K. (2016). Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1758–1762. IEEE. [Cited on pages 96, 119, and 197.]

Carabias-Orti, J. J., Cobos, M., Vera-Candeas, P., & Rodríguez-Serrano, F. J. (2013). Nonnegative signal factorization with learnt instrument models for sound source separation in close-microphone recordings. *EURASIP Journal on Advances in Signal Processing*, *2013*(1), 1–16. [Cited on page 35.]

Caro Repetto, R. & Serra, X. (2014). Creating a corpus of jingju (beijing opera) music and possibilities for melodic analysis. In *Proceedings of the 15th Conference of the International Society for Music/nInformation Retrieval (ISMIR 2014); 2014 Oct 27-31; Taipei, Taiwan. Taipei: International Society for Music Information Retrieval; 2014.* International Society for Music Information Retrieval (ISMIR). [Cited on page 88.]

Chakraborty, K., Talele, A., & Upadhya, S. (2014). Voice recognition using mfcc algorithm. *International Journal of Innovative Research in Advanced Engineering (IJIRAE)*, *1*(10), 2349–2163. [Cited on page 13.]

Chan, T.-S., Yeh, T.-C., Fan, Z.-C., Chen, H.-W., Su, L., Yang, Y.-H., & Jang, R. (2015). Vocal activity informed singing voice separation with the iKala dataset. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 718–722. IEEE. [Cited on pages 86, 97, and 113.]

Chandna, P. (2016). Audio Source Separation Using Deep Neural Networks, Master Thesis, Universitat Pompeu Fabra. [Cited on pages XX, 49, and 112.]

Chandna, P., Miron, M., Janer, J., & Gómez, E. (2017). *Monoaural audio source separation using deep convolutional neural networks*, vol. 10169 LNCS. [Cited on pages XXI, XXI, XXII, XXII, 116, 117, and 118.]

Chandna, P., Ramires, A., & andE. Gómez, X. S. (2021). Loopnet: Musical loop synthesis conditioned on intuitive musical parameters. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [Cited on page 189.]

Chen, W.-R., Whalen, D., & Tiede, M. K. (2021). A dual mechanism for intrinsic f0. *Journal of Phonetics*, *87*, 101063. [Cited on page 15.]

Cherry, C. (1966). On human communication. [Cited on pages 18 and 19.]

Chiu, B., Keogh, E., & Lonardi, S. (2003). Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 493–498. [Cited on page 72.]

Chou, J.-c., Yeh, C.-c., & Lee, H.-y. (2019). One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*. [Cited on pages 73 and 77.]

Chou, J.-c., Yeh, C.-c., Lee, H.-y., & Lee, L.-s. (2018). Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. [Cited on pages 72, 73, and 136.]

Chung, J. & Schafer, R. (1989). A 4.8 k bps homomorphic vocoder using analysis-by-synthesis excitation analysis. In *International Conference on Acoustics, Speech, and Signal Processing,*, pp. 144–147 vol.1. [Cited on page 61.]

Chung, J. H. & Schafer, R. W. (1990). Excitation modeling in a homomorphic vocoder. In *International Conference on Acoustics, Speech, and Signal Processing*, pp. 25–28. IEEE. [Cited on page 61.]

Cuesta, H., Gómez, E., & Chandna, P. (2019). A framework for multi-f<inf>0</inf>modeling in SATB choir recordings. [Cited on pages 201 and 202.]

Cuesta, H., Gómez, E., Martorell, A., & Loáiciga, F. (2018). Analysis of Intonation in Unison Choir Singing. In *Proceedings of the International Conference of Music Perception and Cognition ({ICMPC})*, pp. 125–130. Graz, Austria. [Cited on pages 87, 97, 154, 170, 174, and 181.]

Cuesta, H., McFee, B., & Gómez, E. (2020). Multiple F0 Estimation in Vocal Ensembles using Convolutional Neural Networks. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*. Montreal, Canada (Virtual). [Cited on pages 81 and 154.]

Dai, J., Mauch, M., & Dixon, S. (2015). Analysis of intonation trajectories in solo singing. In *Proceedings of the 16th ISMIR Conference*, vol. 421, p. 29. [Cited on page 88.]

Dauphin, Y. N., Fan, A., Auli, M., & Grangier, D. (2017). Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR. [Cited on pages 56 and 111.]

Davis, S. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357–366. [Cited on page 13.]

De Cheveigné, A. & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, *111*(4), 1917–1930. [Cited on page 78.]

De Krom, G. (1995). Timing and accuracy of fundamental frequency changes in singing. *ICPhS 95 Stockholm, Session 10.2*, *1*, 206–209. [Cited on page 17.]

de Martinville, S. (1860). Édouard-léon. the phonautographic manuscripts of édouard-léon scott de martinville. [Cited on page 6.]

Défossez, A., Usunier, N., Bottou, L., & Bach, F. (2019). Music source separation in the waveform domain. *arXiv preprint arXiv:1911.13254*. [Cited on pages 55, 161, and 202.]

Défossez, A., Zeghidour, N., Usunier, N., Bottou, L., & Bach, F. (2018). Sing: Symbol-to-instrument neural generator. *arXiv preprint arXiv:1810.09785*. [Cited on pages 46 and 132.]

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798. [Cited on page 82.]

Demirel, E., Ahlbäck, S., & Dixon, S. (2020). Automatic lyrics transcription using dilated convolutional neural networks with self-attention. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE. [Cited on pages 71, 134, 187, and 198.]

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. [Cited on page 35.]

Diehl, R. L. (2008). Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 965–978. [Cited on page 14.]

Dieleman, S., Oord, A. v. d., & Simonyan, K. (2018). The challenge of realistic music generation: Modelling raw audio at scale. In *Proc. of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, p. 8000–8010. Red Hook, NY, USA: Curran Associates Inc. [Cited on pages 46 and 132.]

Donahue, C., McAuley, J., & Puckette, M. (2019). Adversarial Audio Synthesis. In *International Conference on Learning Representations*. [Cited on page 125.]

Dong, M., Wu, J., & Luan, J. (2019). Vocal pitch extraction in polyphonic music using convolutional residual network. In *INTERSPEECH*, pp. 2010–2014. [Cited on page 81.]

Doras, G., Esling, P., & Peeters, G. (2019). On the use of u-net for dominant melody estimation in polyphonic music. In *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, pp. 66–70. [Cited on pages 81 and 112.]

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(4), 247–255. [Cited on page 52.]

Dredze, M., Jansen, A., Coppersmith, G., & Church, K. (2010). Nlp on spoken documents without asr. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 460–470. [Cited on page 72.]

Dressler, K. & Fraunhofer, I. (2009). Audio melody extraction for mirex 2009. *5th Music Inform. Retrieval Evaluation eXchange (MIREX)*, *79*, 100–115. [Cited on page 79.]

Duan, Z., Fang, H., Li, B., Sim, K. C., & Wang, Y. (2013). The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9. IEEE. [Cited on pages xix, 16, 19, 86, 97, 123, 128, 197, and 211.]

Duan, Z., Zhang, Y., Zhang, C., & Shi, Z. (2008). Unsupervised single-channel music source separation by average harmonic structure modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(4), 766–778. [Cited on page 104.]

Dubnowski, J., Schafer, R., & Rabiner, L. (1976). Real-time digital hardware pitch detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(1), 2–8. [Cited on page 78.]

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, *12*(Jul), 2121–2159. [Cited on page 39.]

Dudley, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, *11*(2), 169–177. [Cited on pages 12, 59, and 60.]

Dudley, H. (1940). The vocoder—electrical re-creation of speech. *Journal of the Society of Motion Picture Engineers*, *34*(3), 272–278. [Cited on page 60.]

Dudley, H. (1958). Phonetic pattern recognition vocoder for narrow-band speech transmission. *The Journal of the Acoustical Society of America*, *30*(8), 733–739. [Cited on pages 59 and 60.]

Dudley, H., Riesz, R. R., & Watkins, S. S. (1939). A synthetic speaker. *Journal of the Franklin Institute*, *227*(6), 739–764. [Cited on pages 59 and 60.]

Dudley, H. W. (1938). System for the artificial production of vocal or other sounds. US Patent 2,121,142. [Cited on pages 59 and 60.]

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W. et al. (2019). The zero resource speech challenge 2019: Tts without t. *arXiv preprint arXiv:1904.11469*. [Cited on pages 71, 72, and 73.]

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., & Dupoux, E. (2017). The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 323–330. IEEE. [Cited on page 71.]

Dunbar, E., Karadayi, J., Bernard, M., Cao, X.-N., Algayres, R., Ondel, L., Besacier, L., Sakti, S., & Dupoux, E. (2020). The zero resource speech challenge 2020: Discovering discrete subword and word units. *arXiv preprint arXiv:2010.05967*. [Cited on page 71.]

Duong, N. Q., Vincent, E., & Gribonval, R. (2010). Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(7), 1830–1840. [Cited on page 53.]

Durrieu, J.-L., David, B., & Richard, G. (2011). A musically motivated mid-level representation for pitch estimation and musical audio source separation. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1180–1191. [Cited on page 35.]

Durrieu, J.-L., Ozerov, A., Févotte, C., Richard, G., & David, B. (2009). Main instrument separation from stereophonic audio signals using a source/filter model. In *2009 17th European Signal Processing Conference*, pp. 15–19. IEEE. [Cited on page 35.]

Durrieu, J.-L., Richard, G., David, B., & Févotte, C. (2010). Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE transactions on audio, speech, and language processing*, *18*(3), 564–575. [Cited on page 105.]

Edwards, B. (2007). The future of hearing aid technology. *Trends in amplification*, *11*(1), 31–45. [Cited on page 22.]

Ellis, D. P. & Poliner, G. E. (2006). Classification-based melody transcription. *Machine Learning*, *65*(2), 439–456. [Cited on page 80.]

Emiya, V., Vincent, E., Harlander, N., & Hohmann, V. (2011). Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2046–2057. [Cited on pages 96 and 215.]

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*. [Cited on page 125.]

Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020a). Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*. [Cited on page 63.]

Engel, J. H., Hantrakul, L., Gu, C., & Roberts, A. (2020b). DDSP: differentiable digital signal processing. In *Proc. of the 8th International Conference on Learning Representations*. OpenReview.net. [Cited on pages 46 and 132.]

Falk, T. H., Zheng, C., & Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(7), 1766–1774. [Cited on page 92.]

Fan, Z.-C., Lai, Y.-L., & Jang, J.-S. R. (2018). Svsgan: Singing voice separation via generative adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 726–730. IEEE. [Cited on page 56.]

Fang, F., Yamagishi, J., Echizen, I., & Lorenzo-Trueba, J. (2018). High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5279–5283. IEEE. [Cited on page 136.]

Fant, G. (2001). T. chiba and m. kajiyama, pioneers in speech acoustics (< feature articles> sixtieth anniversary of the publication of the vowel, its nature and structure by chiba and kajiyama). *Journal of the Phonetic Society of Japan*, *5*(2), 4–5. [Cited on page 12.]

Févotte, C., Bertin, N., & Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, *21*(3), 793–830. [Cited on pages 21, 35, and 157.]

Févotte, C. & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, *23*(9), 2421–2456. [Cited on pages 21 and 35.]

Fitzgerald, D. (2011). Upmixing from mono-a source separation approach. In *2011 17th International Conference on Digital Signal Processing (DSP)*, pp. 1–7. IEEE. [Cited on page 22.]

Flanagan, J. L. & Golden, R. (1966). Phase vocoder. *Bell System Technical Journal*, *45*(9), 1493–1509. [Cited on page 63.]

Foote, J. & Uchihashi, S. (2001). The beat spectrum: A new approach to rhythm analysis. In *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.*, pp. 224–224. IEEE Computer Society. [Cited on page 34.]

French, N. R. & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, *19*(1), 90–119. [Cited on page 92.]

Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Neural Information Processing Systems (NIPS)*. [Cited on page 82.]

Fujihara, H., Goto, M., Kitahara, T., & Okuno, H. G. (2010). A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(3), 638–648. [Cited on page 104.]

Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2005). Singer identification based on accompaniment sound reduction and reliable frame selection. In *ISMIR*, pp. 329–336. [Cited on page 104.]

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks*, *1*(2), 119–130. [Cited on page 42.]

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, *17*(1), 2096–2030. [Cited on pages 72, 73, and 211.]

Gao, Y., Zhang, X., & Li, W. (2021). Vocal melody extraction via hrnet-based singing voice separation and encoder-decoder-based f0 estimation. *Electronics*, *10*(3), 298. [Cited on page 81.]

Gfeller, B., Frank, C., Roblek, D., Sharifi, M., Tagliasacchi, M., & Velimirović, M. (2020). Spice: Self-supervised pitch estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1118–1128. [Cited on page 81.]

Glass, J. (2012). Towards unsupervised speech processing. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 1–4. IEEE. [Cited on pages 71, 72, and 134.]

Gómez Gutiérrez, E. (2006). Tonal description of music audio signals. [Cited on page 190.]

Gonzalvo, X., Tazari, S., Chan, C.-a., Becker, M., Gutkin, A., & Silen, H. (Proc. Interspeech 2016). Recent advances in google real-time hmm-driven unit selection synthesizer. [Cited on page 58.]

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*. [Cited on pages 45 and 124.]

Goto, M. (2004). A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, *43*(4), 311–329. [Cited on page 79.]

Goto, M. (2007). Active music listening interfaces based on signal processing. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–1441. IEEE. [Cited on page 22.]

Goto, M. et al. (2004). Development of the rwc music database. In *Proceedings of the 18th International Congress on Acoustics (ICA 2004)*, vol. 1, pp. 553–556. [Cited on page 88.]

Gover, M. & Depalle, P. (2019). *Score-informed source separation of choral music*. Ph.D. thesis, Ph. D. dissertation, McGill University. [Cited on pages 154 and 159.]

Gover, M. & Depalle, P. (2020). Score-informed source separation of choral music. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, pp. 231–239. [Cited on pages 154 and 159.]

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pp. 369–376. [Cited on page 65.]

Griffin, D. & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236–243. [Cited on pages 63 and 66.]

Griffin, D. W. & Lim, J. S. (1988). Multiband excitation vocoder. *IEEE Transactions on acoustics, speech, and signal processing*, *36*(8), 1223–1235. [Cited on page 105.]

Gupta, C., Li, H., & Wang, Y. (2018a). Automatic pronunciation evaluation of singing. In *Interspeech*, pp. 1507–1511. [Cited on pages 71 and 187.]

Gupta, C., Tong, R., Li, H., & Wang, Y. (2018b). Semi-supervised lyrics and solo-singing alignment. In *ISMIR*, pp. 600–607. [Cited on pages 71 and 187.]

Hansen, J. K. & Fraunhofer, I. (2012). Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. In *9th Sound and Music Computing Conference (SMC)*, pp. 494–499. [Cited on page 88.]

Hennequin, R., Khlif, A., Voituret, F., & Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, *5*(50), 2154. Deezer Research. [Cited on page 51.]

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. [Cited on page 41.]

Hollier, M. P., Hawksford, M. J., Guard, D. et al. (1993). Characterization of communications systems using a speechlike test stimulus. *Journal of the Audio Engineering Society*, *41*(12), 1008–1021. [Cited on page 92.]

Holt, L. L. & Lotto, A. J. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics*, *72*(5), 1218–1227. [Cited on pages 13 and 14.]

Hono, Y., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2019). Singing Voice Synthesis Based on Generative Adversarial Networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6955–6959. IEEE. [Cited on pages 69, 125, 132, and 198.]

House, A. S. & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America*, *25*(1), 105–113. [Cited on page 15.]

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., & Wang, H.-M. (2016). Voice conversion from non-parallel corpora using variational auto-encoder. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–6. IEEE. [Cited on page 136.]

Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y., & Wang, H.-M. (2017). Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*. [Cited on page 136.]

Hsu, C.-L. & Jang, J.-S. R. (2009). On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 310–319. [Cited on page 86.]

Huang, P.-S., Chen, S. D., Smaragdis, P., & Hasegawa-Johnson, M. (2012). Singing-voice separation from monaural recordings using robust principal component analysis. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 57–60. IEEE. [Cited on pages 21, 36, and 157.]

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014). Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks. In *ISMIR*, pp. 477–482. [Cited on page 49.]

Huang, W.-C., Hayashi, T., Watanabe, S., & Toda, T. (2020). The sequence-to-sequence baseline for the voice conversion challenge 2020: Cascading asr and tts. *arXiv preprint arXiv:2010.02434*. [Cited on page 77.]

Huang, W.-C., Hwang, H.-T., Peng, Y.-H., Tsao, Y., & Wang, H.-M. (2018). Voice conversion based on cross-domain features using variational auto encoders. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 51–55. IEEE. [Cited on page 136.]

Huber, R. & Kollmeier, B. (2006). Pemo-q—a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(6), 1902–1911. [Cited on pages 92 and 96.]

Huq, A., Cartwright, M., & Pardo, B. (2010). Crowdsourcing a real-world on-line query by humming system. In *Proceedings of the Sixth Sound and Music Computing Conference (SMC 2010)*. Citeseer. [Cited on page 88.]

Hwang, Y., Cho, H., Yang, H., Won, D.-O., Oh, I., & Lee, S.-W. (2020). Mel-spectrogram augmentation for sequence to sequence voice conversion. *arXiv preprint arXiv:2001.01401*. [Cited on page 77.]

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, *10*(3), 626–634. [Cited on page 20.]

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, *13*(4-5), 411–430. [Cited on pages 20 and 34.]

Ioffe, S. & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. PMLR. [Cited on page 39.]

Janer, J. & Marxer, R. (2013). Separation of unvoiced fricatives in singing voice mixtures with semi-supervised nmf. In *Proc. 16th Int. Conf. Digital Audio Effects*, pp. 2–5. [Cited on page 35.]

Jansen, A., Church, K., & Hermansky, H. (2010). Towards spoken term discovery at scale with zero resources. In *Eleventh Annual Conference of the International Speech Communication Association*. [Cited on page 72.]

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., Feldman, N., Hermansky, H., Metze, F., Rose, R. et al. (2013). A summary of the 2012 jhu clsp workshop on zero resource speech technologies and models of early language acquisition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8111–8115. IEEE. [Cited on pages 71, 72, and 134.]

Jansson, A., Bittner, R. M., Ewert, S., & Weyde, T. (2019). Joint singing voice separation and f0 estimation with deep u-net architectures. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE. [Cited on pages 81 and 112.]

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks. [Cited on pages XX, XXIII, XXV, 50, 51, 141, 144, 160, 165, 201, and 215.]

Jeong, I.-Y. & Lee, K. (2014). Vocal separation using extended robust principal component analysis with schatten p/l p-norm and scale compression. In *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE. [Cited on page 36.]

Joder, C. & Schuller, B. (2012). Score-informed leading voice separation from monaural audio. In *Proceedings 13th International Society for Music Information Retrieval Conference, ISMIR 2012*. [Cited on page 35.]

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In *Advances in psychology*, vol. 121, pp. 471–495. Elsevier. [Cited on page 40.]

Juvela, L., Bollepalli, B., Yamagishi, J., & Alku, P. (2019a). Gelp: Gan-excited linear prediction for speech synthesis from mel-spectrogram. *arXiv preprint arXiv:1904.03976*. [Cited on page 63.]

Juvela, L., Bollepalli, B., Yamagishi, J., & Alku, P. (2019b). Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919. IEEE. [Cited on page 63.]

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018a). ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder. [Cited on page 136.]

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018b). StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–273. IEEE. [Cited on pages 77 and 136.]

Kaneko, T. & Kameoka, H. (2017). Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*. [Cited on page 136.]

Kaneko, T. & Kameoka, H. (2018). Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2100–2104. IEEE. [Cited on page 136.]

Kaneko, T., Kameoka, H., Hojo, N., Ijima, Y., Hiramatsu, K., & Kashino, K. (2017a). Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *Proc. ICASSP*, vol. 2017, pp. 4910–4914. [Cited on page 125.]

Kaneko, T., Kameoka, H., Tanaka, K., & Hojo, N. (2019a). Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6820–6824. IEEE. [Cited on page 136.]

Kaneko, T., Kameoka, H., Tanaka, K., & Hojo, N. (2019b). Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279*. [Cited on page 136.]

Kaneko, T., Takaki, S., Kameoka, H., & Yamagishi, J. (2017b). Generative adversarial network-based postfilter for STFT spectrograms. In *Proceedings of Interspeech*. [Cited on page 125.]

Karaiskos, V., King, S., Clark, R. A., & Mayo, C. (2008). The blizzard challenge 2008. In *Proc. Blizzard Challenge Workshop*. Citeseer. [Cited on page 91.]

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, *27*(3-4), 187–207. [Cited on pages 67 and 69.]

Kawahara, H., Morise, M., Nisimura, R., & Irino, T. (2012). An interference-free representation of group delay for periodic signals. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4. IEEE. [Cited on page 61.]

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936. IEEE. [Cited on page 61.]

Kenmochi, H. & Ohshita, H. (2007). Vocaloid-commercial singing synthesizer based on sample concatenation. In *Eighth Annual Conference of the International Speech Communication Association*. [Cited on pages 64 and 121.]

Kilgour, K., Zuluaga, M., Roblek, D., & Sharifi, M. (2019). Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pp. 2350–2354. [Cited on page 93.]

Kilmer, A. D. & Civil, M. (1986). Old babylonian musical instructions relating to hymnody. *Journal of Cuneiform Studies*, *38*(1), 94–98. [Cited on page 9.]

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018a). Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165. IEEE. [Cited on pages 80, 112, and 188.]

Kim, J. W., Salamon, J., Li, P., & Bello, J. P. (2018b). {C}repe: {A} Convolutional Representation for Pitch Estimation. In *Proceedings of the {IEEE} International Conference on Acoustics, Speech and Signal Processing ({ICASSP})*, pp. 161–165. Calgary, Canada. [Cited on page 171.]

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. [Cited on pages 39, 111, and 142.]

Kingma, D. P. & Welling, M. (2014). Auto-encoding variational bayes. [Cited on page 44.]

Klapuri, A. (2006). Multiple fundamental frequency estimation by summing harmonic amplitudes. In *ISMIR*, pp. 216–221. [Cited on page 79.]

Klatt, D. H. & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, *87*(2), 820–857. [Cited on page 15.]

Koguchi, J., Takamichi, S., & Morise, M. (2020). Pjs: phoneme-balanced japanese singing-voice corpus. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 487–491. IEEE. [Cited on page 88.]

Kraft, S. & Zölzer, U. (2014). BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality. In *Linux Audio Conference, Karlsruhe, DE*. [Cited on pages XXI, 97, and 98.]

Kruspe, A. M. & Fraunhofer, I. (2016). Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *ISMIR*, pp. 358–364. [Cited on pages 71 and 187.]

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. [Cited on page 71.]

Kum, S., Oh, C., & Nam, J. (2016). Melody extraction on vocal segments using multi-column deep neural networks. In *ISMIR*, pp. 819–825. [Cited on page 81.]

Lagrange, M., Martins, L. G., Murdoch, J., & Tzanetakis, G. (2008). Normalized cuts for predominant melodic source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 278–290. [Cited on page 104.]

Lagrange, M. & Tzanetakis, G. (2007). Sound source tracking and formation using normalized cuts. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1, pp. I–61. IEEE. [Cited on page 104.]

Lahat, M., Niederjohn, R., & Krubsack, D. (1987). A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*(6), 741–750. [Cited on page 78.]

Lea, C., Vidal, R., Reiter, A., & Hager, G. D. (2016). Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, pp. 47–54. Springer. [Cited on pages 43, 54, and 109.]

Lee, D. D. & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. [Cited on pages 21 and 35.]

Lee, J., Choi, H.-S., Jeon, C.-B., Koo, J., & Lee, K. (2019a). Adversarially trained end-to-end korean singing voice synthesis system. *arXiv preprint arXiv:1908.01919*. [Cited on pages 69, 132, and 198.]

Lee, K. & Nam, J. (2019). Learning a Joint Embedding Space of Monophonic and Mixed Music Signals for Singing Voice. *Proceedings of the International Society for Music Information Retrieval Conference ({ISMIR})*. [Cited on pages 82, 137, and 140.]

Lee, S., Ha, J., & Kim, G. (2019b). Harmonizing Maximum Likelihood with {GAN}s for Multimodal Conditional Generation. In *International Conference on Learning Representations*. [Cited on page 127.]

Liu, Z., Chen, K., & Yu, K. (2020). Neural homomorphic vocoder. *Proc. Interspeech 2020*, pp. 240–244. [Cited on page 63.]

Liutkus, A. & Badeau, R. (2015a). Generalized wiener filtering with fractional power spectrograms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 266–270. [Cited on page 20.]

Liutkus, A. & Badeau, R. (2015b). Generalized wiener filtering with fractional power spectrograms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 266–270. IEEE. [Cited on page 53.]

Liutkus, A., Fitzgerald, D., & Rafii, Z. (2015). Scalable audio separation with light kernel additive modelling. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 76–80. IEEE. [Cited on page 86.]

Liutkus, A., Fitzgerald, D., Rafii, Z., Pardo, B., & Daudet, L. (2014a). Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, *62*(16), 4298–4310. [Cited on pages 21 and 34.]

Liutkus, A., Rafii, Z., Pardo, B., Fitzgerald, D., & Daudet, L. (2014b). Kernel spectrogram models for source separation. In *2014 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, pp. 6–10. IEEE. [Cited on pages 21 and 34.]

Liutkus, A., Rohlfing, C., & Deleforge, A. (2018). Audio source separation with magnitude priors: the beads model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 56–60. IEEE. [Cited on page 51.]

Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., & Fontecave, J. (2017). The 2016 signal separation evaluation campaign. In *International conference on latent variable analysis and signal separation*, pp. 323–332. Springer. [Cited on pages 52 and 86.]

Lluís, F., Pons, J., & Serra, X. (2019). End-to-end music source separation: Is it possible in the waveform domain? *Proc. Interspeech 2019*, pp. 4619–4623. [Cited on page 202.]

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M. (2019). Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*. [Cited on page 93.]

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*. [Cited on page 72.]

Lu, Y. & Lu, J. (2020). A universal approximation theorem of deep neural networks for expressing probability distributions. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.) *Advances in Neural Information Processing Systems*, vol. 33, pp. 3094–3105. Curran Associates, Inc. [Cited on page 38.]

Luo, Y., Chen, Z., Hershey, J. R., Le Roux, J., & Mesgarani, N. (2017). Deep clustering and conventional networks for music separation: Stronger together. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 61–65. IEEE. [Cited on page 159.]

Luo, Y. & Mesgarani, N. (2018). Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700. IEEE. [Cited on pages xx, 54, and 56.]

Luo, Y. & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, *27*(8), 1256–1266. [Cited on pages xx, 54, 56, 161, 166, 201, and 202.]

Ma, S., Mcduff, D., & Song, Y. (2019). A generative adversarial network for style modeling in a text-to-speech system. In *International Conference on Learning Representations*. [Cited on page 125.]

Maher, R. C. (1989). A approach for the separation of voices in composite musical signals. *Ph. D. Thesis*. [Cited on pages 21 and 104.]

Maiti, S. & Mandel, M. I. (2019). Speech denoising by parametric resynthesis. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6995–6999. [Cited on page 108.]

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, *63*(4), 561–580. [Cited on page 60.]

Manilow, E., Wichern, G., Seetharaman, P., & Le Roux, J. (2019). Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. [Cited on page 159.]

Mathews, M., Miller, J. E., & David Jr, E. (1961). Pitch synchronous analysis of voiced sounds. *The Journal of the Acoustical Society of America*, *33*(2), 179–186. [Cited on page 61.]

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., & Dixon, S. (2015). Computer-aided melody note transcription using the tony software: Accuracy and efficiency. In *Proceedings of the First International Conference on Technologies for Music Notation and Representation*. Accepted. [Cited on page 79.]

Mauch, M. & Dixon, S. (2014). pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*. In press. [Cited on page 79.]

Mauch, M., Fujihara, H., & Goto, M. (2011). Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 200–210. [Cited on page 88.]

McAulay, R. & Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *34*(4), 744–754. [Cited on pages 11, 59, and 63.]

McCree, A. & Barnwell, T. (1995). A mixed excitation lpc vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and Audio Processing*, *3*(4), 242–250. [Cited on page 61.]

McLeod, A., Schramm, R., Steedman, M., & Benetos, E. (2017). Automatic transcription of polyphonic vocal music. *Applied Sciences*, *7*(12). [Cited on page 89.]

Medsker, L. R. & Jain, L. (2001). Recurrent neural networks. *Design and Applications*, *5*. [Cited on page 40.]

Meron, Y. & Hirose, K. (1998). Separation of singing and piano sounds. In *Fifth International Conference on Spoken Language Processing*. [Cited on page 104.]

Mesaros, A. (2013). Singing voice identification and lyrics transcription for music information retrieval invited paper. In *2013 7th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–10. IEEE. [Cited on pages 71 and 187.]

Mesaros, A. & Virtanen, T. (2009). Adaptation of a speech recognizer for singing voice. In *2009 17th European Signal Processing Conference*, pp. 1779–1783. IEEE. [Cited on page 71.]

Mesaros, A., Virtanen, T., & Klapuri, A. (2007). Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *ISMIR*, pp. 375–378. [Cited on page 104.]

Meseguer-Brocal, G., Cohen-Hadria, A., & Peeters, G. (2019). Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. *arXiv preprint arXiv:1906.10606*. [Cited on pages 87 and 88.]

Meseguer-Brocal, G. & Peeters, G. (2019). Conditioned-u-net: Introducing a control mechanism in the u-net for multiple source separations. *arXiv preprint arXiv:1907.01277*. [Cited on pages 51, 159, and 215.]

Miller, N. J. (1973). Removal of noise from a voice signal by synthesis. Tech. rep., UTAH UNIV SALT LAKE CITY DEPT OF COMPUTER SCIENCE. [Cited on pages 21 and 107.]

Miron, M., Carabias-Orti, J. J., Bosch, J. J., Gómez, E., & Janer, J. (2016). Score-informed source separation for multichannel orchestral recordings. *Journal of Electrical and Computer Engineering*, *2016*. [Cited on page 36.]

Mirza, M. & Osindero, S. (2014). Conditional Generative Adversarial Nets. *CoRR*, *abs/1411.1784*. [Cited on pages 46 and 124.]

Mohammadi, S. H. & Kain, A. (2017). An overview of voice conversion systems. *Speech Communication*, *88*, 65–82. [Cited on pages 71, 134, and 135.]

Monroe, P., Halaki, M., Kumfor, F., & Ballard, K. J. (2020). The effects of choral singing on communication impairments in acquired brain injury: A systematic review. *International journal of language & communication disorders*, *55*(3), 303–319. [Cited on page 6.]

Moore, B., Tyler, L., & Marslen-Wilson, W. (2009). *The perception of speech: from sound to meaning*. Oxford University Press. [Cited on pages 8 and 13.]

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill. [Cited on page 19.]

Mora, J., Gómez, F., Gómez, E., Escobar-Borrego, F., & Díaz-Báñez, J. M. (2010). Characterization and melodic similarity of a cappella flamenco cantes. In *Proceedings of ISMIR*, pp. 9–13. [Cited on page 88.]

Moris, M. (). "tohoku kiritan singing voice corpus. [Cited on page 88.]

Morise, M. (2015a). Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, *67*, 1–7. [Cited on page 61.]

Morise, M. (2015b). Cheaptrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, *67*, 1–7. [Cited on page 109.]

Morise, M. (2016). D4c, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, *84*, 57–65. [Cited on pages 61 and 109.]

Morise, M., Kawahara, H., & Katayose, H. (2009). Fast and reliable f0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech. In *Audio Engineering Society Conference: 35th International Conference: Audio for Games*. Audio Engineering Society. [Cited on pages 61, 79, and 109.]

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, *99*(7), 1877–1884. [Cited on pages xx, 61, 62, 67, 109, 126, 137, 145, and 195.]

Mouchtaris, A., Van der Spiegel, J., & Mueller, P. (2004). Non-parallel training for voice conversion by maximum likelihood constrained adaptation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–1. IEEE. [Cited on page 136.]

Moulines, E. & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*(5-6), 453–467. [Cited on page 63.]

Müller, M. (2007). *Information retrieval for music and motion*, vol. 2. Springer. [Cited on pages 13 and 70.]

Muscariello, A., Gravier, G., & Bimbot, F. (2009). Audio keyword extraction by unsupervised word discovery. In *Tenth Annual Conference of the International Speech Communication Association*. [Cited on page 72.]

Nachmani, E. & Wolf, L. (2019). Unsupervised singing voice conversion. *arXiv preprint arXiv:1904.06590*. [Cited on pages 72, 73, 137, and 211.]

Naderi, B., Jiménez, R. Z., Hirth, M., Möller, S., Metzger, F., & Hoßfeld, T. (2020). Towards speech quality assessment using a crowdsourcing approach: evaluation of standardized methods. *Quality and User Experience*, *6*(1), 1–21. [Cited on page 91.]

Nakamura, K., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2019). Singing voice synthesis based on convolutional neural networks. *arXiv preprint arXiv:1904.06868*. [Cited on page 69.]

Narayanaswamy, V., Thiagarajan, J. J., Anirudh, R., & Spanias, A. (2020). Unsupervised audio source separation using generative priors. *arXiv preprint arXiv:2005.13769*. [Cited on pages 21 and 56.]

Navarro, F. L. (2013). *Life Soundtrack Recovery for Alzheimer's disease patients*. Master's thesis. [Cited on page 6.]

Nercessian, S. (2020). Zero-shot singing voice conversion. In *Proceedings of the International Society for Music Information Retrieval Conference*. [Cited on page 138.]

Ney, H., Suendermann, D., Bonafonte, A., & Höge, H. (2004). A first step towards text-independent voice conversion. In *Eighth International Conference on Spoken Language Processing*. [Cited on page 136.]

Nieto, O. (2013). Unsupervised clustering of extreme vocal effects. In *Proc. 10th Int. Conf. Advances in Quantitative Laryngology*, p. 115. [Cited on page 146.]

Nishimura, M., Hashimoto, K., Oura, K., Nankaku, Y., & Tokuda, K. (2016). Singing voice synthesis based on deep neural networks. In *Interspeech*, pp. 2478–2482. [Cited on page 67.]

Noll, A. M. (1967). Cepstrum pitch determination. *The journal of the acoustical society of America*, *41*(2), 293–309. [Cited on page 78.]

Nugraha, A. A., Liutkus, A., & Vincent, E. (2016). Multichannel music separation with deep neural networks. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1748–1752. IEEE. [Cited on pages 54 and 161.]

Ogawa, I. & Morise, M. (2021). Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs. *Acoustical Science and Technology*, *42*(3), 140–145. [Cited on page 132.]

Oncley, P. (1971). Frequency, amplitude, and waveform modulation in the vocal vibrato. *The Journal of the Acoustical Society of America*, *49*(1A), 136–136. [Cited on page 17.]

Oord, A. v. d., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*. [Cited on page 73.]

Oppenheim, A. & Schafer, R. (1968). Homomorphic analysis of speech. *IEEE Transactions on Audio and Electroacoustics*, *16*(2), 221–226. [Cited on page 107.]

Oura, K., Mase, A., Nankaku, Y., & Tokuda, K. (2012). Pitch adaptive training for hmm-based singing voice synthesis. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5377–5380. IEEE. [Cited on page 64.]

Ozerov, A., Vincent, E., & Bimbot, F. (2012). A General Flexible Framework for the Handling of Prior Information in Audio Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1118–1133. [Cited on pages XXI, XXI, XXII, XXII, 21, 35, 112, 116, 117, 118, and 157.]

Paiva, R. P., Mendes, T., & Cardoso, A. (2006). Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Computer Music Journal*, *30*(4), 80–98. [Cited on page 79.]

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE. [Cited on page 122.]

Park, A. S. & Glass, J. R. (2007). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 186–197. [Cited on page 72.]

Pearce, A., Brookes, T., & Mason, R. (2017). Timbral attributes for sound effect library searching. In *AES International Conference on Semantic Audio*. [Cited on page 190.]

Perez, E., Strub, F., De Vries, H., Dumoulin, V., & Courville, A. (2018a). Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32. [Cited on page 51.]

Perez, E., Strub, F., de Vries, H., Dumoulin, V., & Courville, A. C. (2018b). FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. [Cited on page 159.]

Petermann, D., Chandna, P., Cuesta, H., Bonada, J., & Gómez, E. (2020). Deep Learning Based Source Separation Applied To Choir Ensembles. In *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*. Montreal, Canada (Virtual). [Cited on pages XXIII, 154, 159, and 160.]

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., & Miller, J. (2018). Deep voice 3: 2000-speaker neural text-to-speech. *Proc. ICLR*, pp. 214–217. [Cited on page 66.]

Poliner, G. E. & Ellis, D. P. (2005). A classification approach to melody transcription. [Cited on page 80.]

Poliner, G. E., Ellis, D. P., Ehmann, A. F., Gómez, E., Streich, S., & Ong, B. (2007). Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1247–1256. [Cited on pages 78 and 79.]

Pons, J., Janer, J., Rode, T., & Nogueira, W. (2016). Remixing music using source separation algorithms to improve the musical experience of cochlear implant users. *The Journal of the Acoustical Society of America*, *140*(6), 4338–4349. [Cited on page 22.]

Potter, J. & Sorrell, N. (2012). A history of singing. *A History of Singing*, pp. 1–349. [Cited on page 5.]

Prétet, L., Hennequin, R., Royo-Letelier, J., & Vaglio, A. (2019). Singing voice separation: A study on training data. In *ICASSP 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 506–510. IEEE. [Cited on page 51.]

Pétermann, D. A. (2020). Satb voice segregation for monoaural recordings. In *Masters' thesis, Universitat Pompeu Fabra*. [Cited on page 215.]

Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., & Cox, D. (2020). Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pp. 7836–7846. PMLR. [Cited on page 77.]

Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019). AUTOVC: Zero-Shot Voice Style Transfer with Only Autoencoder Loss. In *International Conference on Machine Learning*, pp. 5210–5219. [Cited on pages xxi, xxii, xxiv, 13, 72, 73, 75, 77, 82, 134, 136, 137, 138, 142, 147, 198, 211, and 212.]

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural Networks*, *12*(1), 145–151. [Cited on page 39.]

Quatieri, T. F. & McAulay, R. J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40*(3), 497–510. [Cited on page 104.]

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *CoRR*, *abs/1511.06434*. [Cited on pages 46, 126, and 127.]

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., & Ellis, D. P. (2014). mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer. [Cited on pages 114, 177, and 204.]

Rafii, Z., Liutkus, A., & Pardo, B. (2014). Repet for background/foreground separation in audio. In *Blind Source Separation*, pp. 395–411. Springer. [Cited on pages 21 and 34.]

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2017). The MUSDB18 corpus for music separation. [Cited on page 86.]

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., FitzGerald, D., & Pardo, B. (2018). An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(8), 1307–1335. [Cited on pages 12 and 21.]

Rafii, Z. & Pardo, B. (2011). A simple music/voice separation method based on the extraction of the repeating musical structure. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 221–224. IEEE. [Cited on page 21.]

Rafii, Z. & Pardo, B. (2012). Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing*, *21*(1), 73–84. [Cited on pages 21 and 34.]

Rahman, M. S. & Shimamura, T. (2010). Pitch determination using autocorrelation function in spectral domain. In *Eleventh Annual Conference of the International Speech Communication Association*. [Cited on page 78.]

Ramires, A., Chandna, P., Favory, X., Gómez, E., & Serra, X. (2020). Neural percussive synthesis parameterised by high-level timbral features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 786–790. [Cited on page 189.]

Rao, P., Nayak, N., & Adavanne, S. (2014). Singing voice separation using adaptive window harmonic sinusoidal modeling. *The Music Information Retrieval Exchange MIREX 2014*. [Cited on page 105.]

Raphael, L. J., Borden, G. J., & Harris, K. S. (2007). *Speech science primer: Physiology, acoustics, and perception of speech*. Lippincott Williams & Wilkins. [Cited on page 14.]

Recht, B., Fazel, M., & Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, *52*(3), 471–501. [Cited on page 21.]

Reddy, C. K., Gopal, V., & Cutler, R. (2020). Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. *arXiv preprint arXiv:2010.15258*. [Cited on page 93.]

Reindl, K., Zheng, Y., & Kellermann, W. (2010). Speech enhancement for binaural hearing aids based on blind source separation. In *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1–6. IEEE. [Cited on page 22.]

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). Fastspeech: Fast, robust and controllable text to speech. *arXiv preprint arXiv:1905.09263*. [Cited on page 66.]

Rennie, S. J., Achan, K., Frey, B. J., & Aarabi, P. (2005). Variational speech separation of more sources than mixtures. In *AISTATS*. Citeseer. [Cited on page 51.]

Rethage, D., Pons, J., & Serra, X. (2018). A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5069–5073. IEEE. [Cited on page 109.]

Rigaud, F. & Radenen, M. (2016). Singing voice melody transcription using deep neural networks. In *ISMIR*, pp. 737–743. [Cited on page 81.]

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2, pp. 749–752. IEEE. [Cited on page 92.]

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer. [Cited on pages xxii, 50, 126, 127, and 197.]

Rosenzweig, S., Cuesta, H., Weiss, C., Scherbaum, F., Gómez, E., & Müller, M. (2020). Dagstuhl ChoirSet: A Multitrack Dataset for MIR Research on Choral Singing. *Transactions of the International Society for Music Information Retrieval (TISMIR)*, *3*(1), 98–110. [Cited on pages 89, 97, 154, and 202.]

Ross, M., Shaffer, H., Cohen, A., Freudberg, R., & Manley, H. (1974). Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *22*(5), 353–362. [Cited on page 79.]

Roux, J. L., Wichern, G., Watanabe, S., Sarroff, A. M., & Hershey, J. R. (2018). Phasebook and friends: Leveraging discrete representations for source separation. *CoRR*, *abs/1810.01395*. [Cited on pages 48 and 51.]

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science. [Cited on pages 39 and 40.]

Ryynänen, M. P. & Klapuri, A. P. (2008). Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, *32*(3), 72–86. [Cited on pages 78 and 79.]

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). An hmm-based singing voice synthesis system. In *Ninth International Conference on Spoken Language Processing*. [Cited on page 64.]

Salamon, J. & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1759–1770. [Cited on pages xxi, 79, 80, 112, and 114.]

Salamon, J., Serra, J., & Gómez, E. (2013a). Tonal representations for music retrieval: from version identification to query-by-humming. *International Journal of Multimedia Information Retrieval*, *2*(1), 45–58. [Cited on page 88.]

Salamon, J. J. et al. (2013b). *Melody extraction from polyphonic music signals*. Ph.D. thesis, Universitat Pompeu Fabra. [Cited on pages xxv, 79, 112, 114, and 115.]

Salaün, Y., Vincent, E., Bertin, N., Souviraà-Labastie, N., Jaureguiberry, X., Tran, D. T., & Bimbot, F. (2014). The flexible audio source separation toolbox version 2.0. In *ICASSP*. [Cited on page 35.]

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *arXiv preprint arXiv:1606.03498*. [Cited on page 93.]

Salimans, T., Karpathy, A., Chen, X., & Kingma, D. P. (2017). Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *ArXiv*, *abs/1701.05517*. [Cited on page 67.]

Salimans, T. & Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *arXiv preprint arXiv:1602.07868*. [Cited on page 40.]

Samuel, D., Ganeshan, A., & Naradowsky, J. (2020). Meta-learning extractors for music source separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 816–820. IEEE. [Cited on pages 55, 161, and 202.]

Santos, J. F. & Falk, T. H. (2014). Updating the srmr-ci metric for improved intelligibility prediction for cochlear implant users. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, *22*(12), 2197–2206. [Cited on page 93.]

Santos, J. F., Senoussaoui, M., & Falk, T. H. (2014). An updated objective intelligibility estimation metric for normal hearing listeners under noise and reverberation. In *Proc. Int. Workshop Acoust. Signal Enhancement*, pp. 55–59. [Cited on page 93.]

Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2018). How does batch normalization help optimization? *arXiv preprint arXiv:1805.11604*. [Cited on page 39.]

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). *Vocal expression of emotion.* Oxford University Press. [Cited on page 5.]

Schoeffler, M., Stöter, F.-R., Edler, B., & Herre, J. (2015). Towards the next generation of web-based experiments: A case study assessing basic audio quality following the itu-r recommendation bs. 1534 (mushra). In *1st Web Audio Conference*, pp. 1–6. [Cited on pages 90 and 94.]

Schramm, R. & Benetos, E. (2017). Automatic Transcription of a cappella Recordings from Multiple Singers. Audio Engineering Society. [Cited on page 89.]

Scirea, M. & Brown, J. A. (2015). Evolving four part harmony using a multiple worlds model. In *Proceedings of the 7th International Joint Conference on Computational Intelligence (IJCCI)*, vol. 1, pp. 220–227. IEEE. [Cited on page 154.]

Seashore, C. E. (1932). The vibrato, volume i of university of iowa studies in the psychology of music. [Cited on page 17.]

Serrà, J., Pascual, S., & Segura Perales, C. (2019). Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.) *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. [Cited on page 77.]

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21. [Cited on page 60.]

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., & Others (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. IEEE. [Cited on page 66.]

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. [Cited on pages 63, 75, and 144.]

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). Tobi: A standard for labeling english prosody. In *Second international conference on spoken language processing*. [Cited on page 14.]

Siu, M.-h., Gish, H., Lowe, S., & Chan, A. (2011). Unsupervised audio patterns discovery using hmm-based self-organized units. In *Twelfth Annual Conference of the International Speech Communication Association*. [Cited on page 72.]

Slaney, M., Naar, D., & Lyon, R. (1994). Auditory model inversion for sound separation. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. II–77. IEEE. [Cited on page 105.]

Smaragdis, P., Raj, B., & Shashanka, M. (2007). Supervised and semi-supervised separation of sounds from single-channel mixtures. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 414–421. Springer. [Cited on page 35.]

Smith, J. C. (2013). *Correlation analyses of encoded music performance*. Stanford University. [Cited on pages 87 and 88.]

Smith, J. O. & Serra, X. (1987). Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. In *Proceedings of the 1987 International Computer Music Conference, ICMC; 1987 Aug 23-26; Champaign/Urbana, Illinois.[Michigan]: Michigan Publishing; 1987. p. 290-7.* International Computer Music Conference. [Cited on page 105.]

Sorensen, C., Kavalekalam, M. S., Xenaki, A., Boldt, J. B., & Christensen, M. G. (2017). Non-intrusive intelligibility prediction using a codebook-based approach. In *EUSIPCO*, pp. 216–220. [Cited on page 93.]

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y. (2017). Char2Wav: End-to-end speech synthesis. [Cited on page 66.]

Southall, C., Stables, R., & Hockman, J. (2017). Automatic drum transcription for polyphonic recordings using soft attention mechanisms and convolutional neural networks. In *Proc. of the 18th International Society for Music Information Retrieval Conference*. Suzhou, China. [Cited on page 190.]

Sprechmann, P., Bronstein, A. M., & Sapiro, G. (2012). Real-time online singing voice separation from monaural recordings using robust low-rank modeling. In *ISMIR*, pp. 67–72. [Cited on pages 21 and 36.]

Steeneken, H. J. & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America*, *67*(1), 318–326. [Cited on page 92.]

Stevens, K. N. (1960). Toward a model for speech recognition. *The Journal of the Acoustical Society of America*, *32*(1), 47–55. [Cited on pages 8 and 23.]

Stevens, K. N. (1972). Segments, features, and analysis by synthesis. [Cited on pages 8 and 23.]

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, *8*(3), 185–190. [Cited on pages 11, 12, 13, and 15.]

Stillings, L. (2021). Singing voice melody estimation from polyphonic signals. [Cited on page 115.]

Stoller, D., Durand, S., & Ewert, S. (2019). End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 181–185. IEEE. [Cited on page 88.]

Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. [Cited on pages xx, 48, 52, 53, 55, 126, 159, 160, 165, 189, 201, and 215.]

Stöter, F.-R., Liutkus, A., & Ito, N. (2018). The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 293–305. Springer. [Cited on page 52.]

Stöter, F.-R., Uhlich, S., Liutkus, A., & Mitsufuji, Y. (2019). Open-unmix - a reference implementation for music source separation. *Journal of Open Source Software*. [Cited on pages xx, 54, 55, 161, 166, 188, 201, 202, and 215.]

Su, L. (2018). Vocal melody extraction using patch-based cnn. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 371–375. IEEE. [Cited on page 81.]

Sundberg, J. (1987). *The Science of the Singing Voice*. Northern Illinois University Press. [Cited on page 16.]

Sundberg, J. & Rossing, T. D. (1990). The science of singing voice. [Cited on pages 9, 11, 12, 16, and 17.]

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4214–4217. [Cited on page 92.]

Takahashi, N., Agrawal, P., Goswami, N., & Mitsufuji, Y. (2018a). Phasenet: Discretized phase modeling with deep neural networks for audio source separation. In *INTERSPEECH*, pp. 2713–2717. [Cited on page 52.]

Takahashi, N., Goswami, N., & Mitsufuji, Y. (2018b). Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 106–110. IEEE. [Cited on page 54.]

Takahashi, N. & Mitsufuji, Y. (2017). Multi-scale multi-band densenets for audio source separation. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 21–25. [Cited on page 54.]

Talkin, D. & Kleijn, W. B. (1995). A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, *495*, 518. [Cited on page 79.]

Tamaru, H., Takamichi, S., Tanji, N., & Saruwatari, H. (2020). Jvs-music: Japanese multispeaker singing-voice corpus. *arXiv preprint arXiv:2001.07044*. [Cited on pages 87 and 88.]

Ternström, S. (1991). Perceptual evaluations of voice scatter in unison choir sounds. *STL-Quarterly Progress and Status Report*, *32*, 41–49. [Cited on pages 28, 170, 171, 173, 179, 181, 182, and 204.]

Ternström, S. (2002). Choir acoustics – an overview of scientific research published to date. *Speech, Music and Hearing Quarterly Progress and Status Report*, *43*(April), 1–8. [Cited on page 87.]

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., & Colomes, C. (2000). Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, *48*(1/2), 3–29. [Cited on page 92.]

Tokuda, K., Kobayashi, T., Masuko, T., & Imai, S. (1994). Mel-generalized cepstral analysis-a unified approach to speech spectral estimation. In *Third International Conference on Spoken Language Processing*. [Cited on page 62.]

Tsai, C.-P., Tuan, Y.-L., & Lee, L.-s. (2018). Transcribing lyrics from commercial song audio: the first step towards singing content processing. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5749–5753. IEEE. [Cited on pages 71 and 187.]

Typke, R. (2007). Music retrieval based on melodic similarity. ASCI. [Cited on page 78.]

Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., & Mitsufuji, Y. (2017). Improving music source separation based on deep neural networks through data augmentation and network blending. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265. IEEE. [Cited on page 54.]

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2017). Instance normalization: The missing ingredient for fast stylization. [Cited on page 74.]

Umbert, M., Bonada, J., & Blaauw, M. (2013). Systematic database creation for expressive singing voice synthesis control. In *Eighth ISCA Workshop on Speech Synthesis*. [Cited on page 88.]

Valin, J.-M. & Skoglund, J. (2019). Lpcnet: Improving neural speech synthesis through linear prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5891–5895. IEEE. [Cited on page 63.]

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*. [Cited on page 110.]

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., & Kavukcuoglu, K. (2016a). WaveNet: A Generative Model for Raw Audio. In *SSW*. [Cited on pages xxi, 46, 63, 65, 67, 75, 76, 108, 109, 111, 122, 126, and 211.]

van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., & Others (2016b). Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, pp. 4790–4798. [Cited on page 110.]

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11). [Cited on page 178.]

van Niekerk, B., Nortje, L., & Kamper, H. (2020). Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. *arXiv preprint arXiv:2005.09409*. [Cited on page 74.]

Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4052–4056. IEEE. [Cited on page 82.]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. [Cited on pages xxi, 43, and 69.]

Veaux, C., Yamagishi, J., MacDonald, K. et al. (2017). Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. [Cited on page 122.]

Velasco, G. A., Holighaus, N., Dörfler, M., & Grill, T. (2011). Constructing an invertible constant-q transform with non-stationary gabor frames. *Proceedings of DAFX11, Paris*, *33*. [Cited on page 81.]

Vembu, S. & Baumann, S. (2005). Separation of vocals from polyphonic audio recordings. In *ISMIR*, pp. 337–344. Citeseer. [Cited on pages 21 and 35.]

Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The zero resource speech challenge 2015. In *Sixteenth annual conference of the international speech communication association*. [Cited on page 71.]

Vích, R. & Vondra, M. (). Voice conversion based on spectral envelope transformation. [Cited on page 61.]

Vincent, E. (2012). Improved perceptual metrics for the evaluation of audio source separation. In *International Conference on Latent Variable Analysis and Signal Separation*, pp. 430–437. Springer. [Cited on pages 96 and 215.]

Vincent, E., Araki, S., & Bofill, P. (2009). The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *International Conference on Independent Component Analysis and Signal Separation*, pp. 734–741. Springer. [Cited on page 52.]

Vincent, E., Araki, S., Theis, F., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, V., Lutter, D., & Duong, N. Q. (2012). The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, *92*(8), 1928–1936. [Cited on page 52.]

Vincent, E., Gribonval, R., & Fevotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, *14*(4), 1462–1469. [Cited on pages 53, 94, 113, 119, 163, 196, and 201.]

Vinyes, M. (2008). MTG MASS database. http://www.mtg.upf.edu/static/mass/resources. [Cited on page 86.]

Virtanen, T. (2007). Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, *15*(3), 1066–1074. [Cited on pages 21, 35, and 157.]

Von Helmholtz, H. (1912). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Longmans, Green. [Cited on page 18.]

Voran, S. (1999). Objective estimation of perceived speech quality. i. development of the measuring normalizing block technique. *IEEE Transactions on speech and audio processing*, *7*(4), 371–382. [Cited on page 92.]

Wan, C. Y., Rüüber, T., Hohmann, A., & Schlaug, G. (2010). The therapeutic effects of singing in neurological disorders. *Music perception*, *27*(4), 287–295. [Cited on page 6.]

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883. IEEE. [Cited on pages 82, 137, 140, and 178.]

Wang, A. (1994). *Instantaneous and frequency-warped techniques for auditory source separation*. Ph.D. thesis, Ph. D. dissertation, Stanford Univ., Stanford, CA, USA. [Cited on pages 21 and 104.]

Wang, A. L. (1995). Instantaneous and frequency-warped techniques for source separation and signal parametrization. In *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, pp. 47–50. IEEE. [Cited on pages 21 and 104.]

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, pp. 181–197. Springer. [Cited on pages 20 and 53.]

Wang, Q., Fan, H., Sun, G., Cong, Y., & Tang, Y. (2019). Laplacian pyramid adversarial network for face completion. *Pattern Recognition*, *88*, 493–505. [Cited on page 46.]

Wang, Q., Fan, H., Sun, G., Ren, W., & Tang, Y. (2020a). Recurrent generative adversarial network for face completion. *IEEE Transactions on Multimedia*, *23*, 429–442. [Cited on page 46.]

Wang, R., Ding, Y., Li, L., & Fan, C. (2020b). One-shot voice conversion using star-gan. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7729–7733. IEEE. [Cited on page 136.]

Wang, S., Sekey, A., & Gersho, A. (1992). An objective measure for predicting subjective quality of speech coders. *IEEE Journal on selected areas in communications*, *10*(5), 819–829. [Cited on page 92.]

Wang, Y., Narayanan, A., & Wang, D. (2014). On Training Targets for Supervised Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1849–1858. [Cited on page 48.]

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. In *INTERSPEECH*. [Cited on page 66.]

Ward, D., Wierstorf, H., Mason, R. D., Grais, E. M., & Plumbley, M. D. (2018). Bss eval or peass? predicting the perception of singing-voice separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 596–600. IEEE. [Cited on page 96.]

Weinstein, C. & Oppenheim, A. (1971). Predictive coding in a homomorphic vocoder. *IEEE Transactions on Audio and Electroacoustics*, *19*(3), 243–248. [Cited on page 61.]

Weintraub, M. (1985). *A theory and computational model of auditory monaural sound separation*. Ph.D. thesis, Stanford University Stanford, CA. [Cited on pages 19 and 20.]

Weiss, C., Schelcht, S. J., Rosenzweig, S., & Müller, M. (2019). Towards Measuring Intonation Quality of Choir Recordings: A Case Study on Bruckner's Locus Iste. In *Proceedings of the 20th International Society for Music Information Retrieval Conference ({ISMIR})*, pp. 276–283. Delft, The Netherlands. [Cited on page 170.]

Wester, M., Wu, Z., & Yamagishi, J. (2016). Analysis of the voice conversion challenge 2016 evaluation results. In *Interspeech*, pp. 1637–1641. [Cited on page 72.]

Wiener, N. (1950). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. MIT press Cambridge. [Cited on page 20.]

Wiener, N., Wiener, N., Mathematician, C., Wiener, N., Wiener, N., & Mathématicien, C. (1949). Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications. [Cited on page 20.]

Wilkins, J., Seetharaman, P., Wahl, A., & Pardo, B. (2018). Vocalset: A singing voice dataset. In *ISMIR*, pp. 468–474. [Cited on page 88.]

Williamson, D. S., Wang, Y., & Wang, D. (2015). Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, *24*(3), 483–492. [Cited on page 51.]

Wu, D.-Y., Chen, Y.-H., & Lee, H.-Y. (2020). Vqvc+: One-shot voice conversion by vector quantization and u-net architecture. *arXiv preprint arXiv:2006.04154*. [Cited on pages XXIV, 72, 73, 74, 82, 136, 137, 147, and 213.]

Wu, D.-Y. & Lee, H.-y. (2020). One-shot voice conversion by vector quantization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7734–7738. IEEE. [Cited on pages XXIV, 72, 73, 74, 82, 136, 147, and 213.]

Wu, Z., Xie, Z., & King, S. (2019). The blizzard challenge 2019. In *Proc. Blizzard Challenge workshop*, vol. 2019. [Cited on page 91.]

Yamagishi, J., Veaux, C., MacDonald, K. et al. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). [Cited on pages XIX and 15.]

Yang, S., Xie, L., Chen, X., Lou, X., Zhu, X., Huang, D., & Li, H. (2017). Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 685–691. IEEE. [Cited on page 125.]

Yang, Y.-H. (2013). Low-rank representation of both singing voice and music accompaniment via learned dictionaries. In *ISMIR*, pp. 427–432. [Cited on page 36.]

Yi, Y.-H., Ai, Y., Ling, Z.-H., & Dai, L.-R. (2019). Singing voice synthesis using deep autoregressive neural networks for acoustic modeling. *arXiv preprint arXiv:1906.08977*. [Cited on page 67.]

Yi, Z., Huang, W.-C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., Ling, Z., & Toda, T. (2020). Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion—. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 80–98. [Cited on page 72.]

Young, E. D. (2008). Neural representation of spectral and temporal information in speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 923–945. [Cited on page 10.]

Ze, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *2013 ieee international conference on acoustics, speech and signal processing*, pp. 7962–7966. IEEE. [Cited on page 64.]

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*. [Cited on page 39.]

Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F., & Szczepaniak, P. (2016). Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices. [Cited on page 64.]

Zen, H., Toda, T., Nakamura, M., & Tokuda, K. (2007). Details of the nitech hmm-based speech synthesis system for the blizzard challenge 2005. *IEICE transactions on information and systems*, *90*(1), 325–333. [Cited on page 58.]

Zhang, Y. & Glass, J. R. (2010). Towards multi-speaker unsupervised speech pattern discovery. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4366–4369. IEEE. [Cited on page 72.]

Zhao, R., Lee, S. W., Huang, D.-Y., & Dong, M. (2014). Soft constrained leading voice separation with music score guidance. In *The 9th International Symposium on Chinese Spoken Language Processing*, pp. 565–569. IEEE. [Cited on page 35.]

Zhao, Y., Takaki, S., Luong, H.-T., Yamagishi, J., Saito, D., & Minematsu, N. (2018). Wasserstein GAN and Waveform Loss-based Acoustic Model Training for Multi-speaker Text-to-Speech Synthesis Systems Using a WaveNet Vocoder. *IEEE Access*, *6*, 60478–60488. [Cited on page 125.]

Zhou, X., Ling, Z.-H., & King, S. (2020). The blizzard challenge 2020. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pp. 1–18. [Cited on page 91.]

Zukowski, Z. & Carr, C. (2018). Generating black metal and math rock: Beyond bach, beethoven, and beatles. *arXiv preprint arXiv:1811.06639*. [Cited on pages 46 and 132.]