COMPREHENSIVE EVALUATION OF SINGING QUALITY

CHITRALEKHA GUPTA

NATIONAL UNIVERSITY OF SINGAPORE

2019

COMPREHENSIVE EVALUATION OF SINGING QUALITY

CHITRALEKHA GUPTA
(M.Tech. in Electrical Engg., IIT Bombay, India,
B.Eng. in Electronics Engg., MSU Baroda, India)

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING (NGS), WITH DEPT. OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2019

Supervisors:
Associate Professor Ye Wang, Main Supervisor
Professor Haizhou Li, Co-Supervisor

Examiners:
Associate Professor Roger Zimmermann,
Associate Professor Ng Teck Khim,
Professor Xavier Serra, Universitat Pompeu Fabra

Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Chitralekha Gupta
January 2019

Acknowledgments

First and foremost, I would like to express my heartfelt gratitude to my supervisors, Prof. Haizhou Li and Prof. Ye Wang, for their continuous guidance and support throughout this journey. They inspired and encouraged me to always pursue quality research. I sincerely thank them for the excellent guidance and the timely constructive criticism that has helped in the shaping of this thesis.

I owe my sincere thanks to my Thesis Advisory Committee (TAC), Prof. Lonce Wyse, and Prof. Ng Teck Khim, for the insightful discussions and their feedback at different stages of this thesis, which kept me going in the right direction.

I would like to express special thanks to all my collaborators, Prof. Preeti Rao, Tong Rong, David Grunberg, and Bidisha Sharma, for their help and support in various sub-parts of this thesis.

I would like to sincerely thank NGS and School of Computing, for supporting my education and research. I am grateful to NGS for introducing and guiding me through interdisciplinary research that has a highly positive impact on my research outlook.

I thank my wonderful lab members and friends, Karim Magdi, Dania Murad, Praveen Kumar, Rajat Mishra, Shruti Tople, Xing Zhe, Cai Jingli, Duan Zhiyan, Fang Jiakun, Wei Wei, Sai Phai, Michael Mustaine, and many others, for their company and support that made life in Singapore cheerful. I would especially like to thank Boyd Anderson, and Prachee Priyadarshinee for being my rock-solid friends and constant support, without whom this journey would have been difficult.

And last but definitely not the least, I am highly indebted to my family members - mom, dad, sister, brother-in-law, and niece for being the strong pillars in my life. Your unconditional love and faith in me has made it possible to overcome all the hurdles and successfully complete this thesis work.

Abstract

Singing is a popular medium of entertainment, and a desirable skill to develop. But singing pedagogy remains heavily dependent on human music experts, who are few in number. In this thesis, we study the methodology for automatic evaluation of singing voice with respect to two broad aspects of singing quality - prosody and pronunciation.

We study various prosody-related parameters of singing quality judgment, as identified by music experts, such as pitch, rhythm, vibrato, and timbre, and objectively characterize them. We also incorporate a cognitive modeling theory inspired by the telecommunication standard PESQ (Perceptual Evaluation of Speech Quality) to provide a perceptually valid objective score to assess singing quality. Furthermore, we design a reference-independent method of evaluation that leverages on the large amounts of singing data available to compute inter-singer statistics. These algorithms result in an automatic generation of a leaderboard of singers, without relying on an ideal or reference singing sample for comparison.

Another aspect of singing quality is pronunciation. Phonetic modeling of singing voice has applications in language learning, speech therapy, and karaoke. We address the problem of the lack of lyrics-aligned singing voice datasets by designing a strategy to automatically build such datasets with the help of imperfect transcriptions from automatic speech recognition (ASR) and non-aligned published lyrics. With this dataset, we design algorithms to build singing-adapted speech acoustic models that take into account the differences between speech and singing, such as the duration of vowels and pitch dynamics. We show the usability of these models in pronunciation evaluation in singing voice, and automatic lyrics-to-audio alignment.

List of Publications

1. **Chitralekha Gupta**, Haizhou Li, and Ye Wang, "Perceptual Evaluation of Singing Quality", *In Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Kuala Lumpur, 2017. **(Best Student Paper Award)**
2. **Chitralekha Gupta**, Haizhou Li, and Ye Wang, "A Technical Framework for Automatic Perceptual Evaluation of Singing Quality", *Transactions of APSIPA*, 2018.
3. **Chitralekha Gupta**, Haizhou Li, and Ye Wang, "Automatic Evaluation of Singing Quality without a Reference", *In Proceedings of APSIPA*, Hawaii, 2018.
4. **Chitralekha Gupta**, Haizhou Li, and Ye Wang, "Automatic Leaderboard: Evaluation of singing quality without a standard reference", to be submitted to IEEE Transactions on Audio, Speech, and Language Processing.
5. **Chitralekha Gupta**, Rong Tong, Haizhou Li, and Ye Wang, "Semi-supervised lyrics and solo-singing alignment", *In Proceedings of International Society of Music Information Retrieval (ISMIR)*, Paris, 2018.
6. **Chitralekha Gupta**, David Grunberg, Preeti Rao, and Ye Wang, "Towards Automatic Mispronunciation Detection in Singing", *In Proceedings of ISMIR*, Suzhou, 2017.
7. **Chitralekha Gupta**, Haizhou Li, and Ye Wang, "Automatic Pronunciation Evaluation of Singing", *In Proceedings of Interspeech*, Hyderabad, 2018.
8. **Chitralekha Gupta***, Bidisha Sharma*, Haizhou Li, and Ye Wang, "Automatic Lyrics-to-Audio Alignment on Polyphonic Music Using Singing-Adapted Acoustic Models", *Submitted to ICASSP*, 2019.(*equal contributors)
9. Douglas Turnbull, **Chitralekha Gupta**, Dania Murad, Michael Barone, and Ye Wang, "Using Music Technology to Motivate Foreign Language Learning", *In Proceedings of International Conference on Orange Technologies (ICOT)*, Singapore, 2017.
10. Michael Mustaine, Karim Ibhrahim, **Chitralekha Gupta**, and Ye Wang, "Empirically weighing the importance of decision factors when selecting music to sing", *In Proceedings of ISMIR*, Paris, 2018.
11. Karim Magdi, David Grunberg, Kat Agres, **Chitralekha Gupta**, and Ye Wang, "Intelligibility of Sung Lyrics: A Pilot Study", *In Proceedings of ISMIR*, Suzhou, 2017.
12. Zhiyan Duan, **Chitralekha Gupta**, Graham Percival, David Grunberg, and Ye Wang, "SEC-CIMA: Singing and Ear Training for Children with Cochlear Implants via a Mobile Application", *In Proceedings of Sound and Music Computing (SMC)*, Helsinki, 2017.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "In the beginning was the voice. Voice is sounding breath, the audible sign of life." – Otto Jespersen, *Language, its Nature, Development and Origin.*
>
> "We may ordinarily regard the voice organ as a tool for making sound; a singer uses this tool as a musical instrument." – Johan Sundberg, *The Science of the Singing Voice*

1.1 Why do we sing?

Singing, the vocal production of musical tones, is so basic to humans that today, we rarely wonder about its origins. Voice is presumed to be the oldest musical instrument, and there is evidence of singing being universally present in human culture since antiquity [Koo94]. Scientists have argued that singing in its earliest form may have preceded speech [Mon17]. The ability to produce something melodic, such as humming and mother-infant vocalizations, may have preceded the ability to form the consonants and vowels to make meaningful speech. But what is the purpose of singing? Before written language, stories were passed down to generations through songs, as songs are often more memorable. Chants and hymns were part of religious rituals, and recounts of history and heroics were often in the form of ballads and epics. From a broad evolutionary perspective, there are three theories about why singing was beneficial for humans [Eco08].

First, the Shakespearean anecdotal theory, that music is "one of the foods of love", has a strong claim to be true [Eco08]. The more mellifluous the singer, the more mates he attracts. Charles Darwin [Dar88] suggested that a lot of the features in animals have been selected not to aid its survival, but to aid its need to find a mate, such as the tail of the peacock. He suggested that human features too, such as the ability to sing, might be selected in this way. However, this theory does not seem like the whole story. There has to be more as to why we sing because people enjoy singing and listening to music even without the intention of finding a mate.

The second more controversial idea suggested by Pinker [Pin97] is that music was an

accidental invention and is a consequence of the abilities that evolved for other purposes, such as language. That is, language led to music, where music was originally non-functional, and subsequently has been exploited by evolution and made functional. This idea has been widely criticized by scientists [JH05, Kon10] stating evidence that suggest precedence of music for rudimentary communication over natural language, and justifying why music is more than just an "auditory cheesecake" [Eco08].

The third more popular hypothesis is that music binds groups of people together, and has originated from the social benefits of group-living. Hagen et al. [HH09] suggested that the evolution of human music and dance is rooted in coordinated auditory and visual territorial advertisement for group defence. Music has been used for personal and communal entertainment, dancing, communication, and religious rituals [Mon17]. In general, music is believed to result in bonding – between mother and child, between groups of people, for example who are working together [Mon17, MMW+06]. Work songs have even been a cohesive element in pre-industrial societies, i.e. everyone of the group moves together with the rhythm of the song to make a mundane repetitive job more interesting. It is even suggested that it was music that brought individuals and groups of people together who might otherwise have led solitary lives, scattered at random over the landscape.

Although the origins of singing and music are still debated, what cannot be denied is its ability to evoke emotions and the role it plays in everyone's lives. We listen to an upbeat song to change our mood, or hum to our favorite song in the shower. The immense popularity of singing idol shows, music channels, radio stations or online platforms such as YouTube, Spotify or karaoke applications such as Smule Sing! show how we are so surrounded by music in our daily lives today. The music industry generated a whopping 17.3 billion dollar revenue in 2017 [IFP18]. Moreover, we witness the binding power of music that brings people together, transcending the boundaries of countries and languages. Remember how the Spanish song *Despacito* topped every international music chart in 2017?

Music has tremendous educational, entertainment, and therapeutic value. However, for this value to actually reach the masses, there are various problems to be addressed. For example, many of us would like to learn to sing like our favorite singer, but do not have the time to take music lessons. Or many of us would like to learn a new language but get demotivated by the traditional methods of language learning and want a more effective solution, where music could help. This thesis proposes and investigates various technological methods to overcome these hurdles to make music and its benefits more accessible to the people.

In this chapter, we discuss about the singing voice production mechanism, a review of the singing skill assessment criteria in singing pedagogy, followed by the goal of this thesis.

1.2 Singing voice production mechanism

Singing is a skill that requires highly developed muscle reflexes and a high degree of muscle coordination. Individuals can develop their voices further through the careful and systematic practice of both songs and vocal exercises. In its physical aspect, singing and speech share the underlying voice organ. The voice organ consists of three units [SR90]: the breathing apparatus, the vocal folds, and the vocal tract. The purpose of the breathing apparatus is to compress the air in the lungs, which act as an air supply or bellows, so that an air stream is generated past the vocal folds and the vocal tract. The vocal folds vibrate with the passage of air generating a primary sound called voice source. The frequency of vibration of the vocal folds is equal to the frequency of the generated note, i.e. pitch. The voice source passes through the vocal tract which "shapes" it acoustically. The vocal tract consists of the oral cavity, palate, teeth, lips, and tongue. The nature of the shaping depends on the vocal tract configuration, which is controlled by articulation. As the source-filter model of speech production [RS78] suggests, the vocal tract acts as a resonator that resonates at certain frequencies, called formant frequencies, according to its shape. Vocal fold vibration along with vocal tract configurations results in the production of voiced sounds such as vowels and voiced consonants. When the vocal folds do not vibrate, rather the air stream from the lungs is forced to pass through a narrow slit or constriction, the air stream becomes turbulent and noisy, lacking pitch, which gives rise to the unvoiced sounds. For example the unvoiced sound /f/ is a result of the slit between the lower lip and the front teeth that work as the noise-generating oscillator. Thus pitch is altered by the vocal folds, while articulation is controlled by the vocal tract.

If speech and singing share a common voice production organ and mechanism, then the question arises how are the two phenomena different. The degree of variations in the units of the voice organ manifest as the differences between speech and singing voice [FGOO11, MEG14]. For example, the pitch variations in singing voice is far more than the relatively monotonic speech voice. A controlled manipulation of the vocal fold vibrations result in the pitch variations in singing voice. Another big difference between singing and speech is the duration of the vowels. In singing, the vowels are stretched in time to sustain the musical notes, which is dictated by the musical score of the song. The duration of vowels in speech is comparatively small and less varying. Moreover, singing voice often contains embellishments in pitch such as vibrato, which is not present in speech. Therefore, singing voice has its own defining properties that need to be understood and characterized.

1.3 Singing quality assessment

Learning to sing is an activity that benefits from the involvement of an instructor. However, singing pedagogy remains heavily dependent on human music experts, who are a few in

number. The evaluation criterion for singing relies on subjective expert judgments, which are not always conveniently available to people who desire to learn singing. Thus, a system for automatic and reliable evaluation of singing could serve as an aid to singing pedagogy, as well as to singing contests, and karaoke systems, in turn making singing training more accessible to the masses.

Assessing singing voice quality is a subjective matter, no one knows what is absolutely right or wrong. Yet there is some kind of conformity in what sounds pleasing, and hence the right way of singing. Studies have shown that music experts can evaluate singing quality with high consensus when the melody or the song is unknown to them [NGH06b, NGH06a]. This suggests that there are inherent properties of singing quality that are independent of a reference singer or melody, which help music-experts judge singing quality with high conformity without a reference. Therefore, despite its subjective nature, singing quality assessment has been objectively broken down into various perceptual parameters.

Singing quality can be characterized broadly and independently from two main aspects: *prosody* and *articulation*. Prosodic attributes such as pitch, and rhythm represent the way the musical notes of a song are sung. Articulation or pronunciation related attributes represent the way the phonetic segments (vowels and consonants) of the lyrics of a song are uttered. In singing quality judgment in the context of singing training or singing contests, the prosodic attributes are observed to be more important than pronunciation. However, the articulation of lyrics is important in music-based language learning and speech therapies. Modeling the phonetics of singing voice has many other applications such as automatic lyrics recognition and alignment.

It is important to note here that the singing quality assessment criteria may vary across the different singing style categories, such as Western classical, opera, and Indian classical, as well as across different genres such as pop, rock, metal, rap, and jazz. The work in this thesis focuses particularly on Western English popular music.

1.3.1 *Prosody* aspect of singing voice

Singing quality assessment often refers to the degree to which a particular vocal production meets professional standards of excellence. For reliable assessment, it is important to identify vocal attributes that relate to human perceptual ratings and objectively define singing excellence. Past studies have identified several perceptual parameters pertaining to singing voice that play a significant role in subjective evaluation of singing skill. One study described twelve generally accepted criteria used in the evaluation of Western classical singing by expert music teachers [WE97], which are: *appropriate vibrato, resonance/ring, color/warmth, intensity, dynamic range, efficient breath management, evenness of registration, flexibility, freedom throughout vocal range, intonation accuracy, legato line,* and *diction*. Oates et al. [OBD+06] proposed an auditory-perceptual rating scale for operatic singing voice, which

consisted of five perceptual parameters - *appropriate vibrato, ring, pitch accuracy, evenness throughout the range,* and *strain*, and these parameters were proven to be unambiguous and covered all aspects of operatic voice.

However, all of the above parameters may not be suitable for evaluating a non-trained or a novice singer. For example, as pointed out by Tsai et al. [TL12], the ringing voice quality is typically observed in operatic style of singing. However, operatic style is a specific way of singing that, one may argue, can be unsuitable and undesirable for singing lessons or karaoke performances, especially for beginners. Chuan et al. [CLLY08] defined and verified six perceptual parameters that were of most relevance for assessing non-trained singers. These parameters were: *Intonation accuracy*, described as singing in tune, where suitable key transposition is allowed; *Rhythm consistency*, described as singing with appropriate tempo speed, where slight tempo variation is allowed; *Timbre brightness*, described as brilliance of tone, a sensation of brightness of spectrum; *Appropriate vibrato*, described as regular and smooth undulation of frequency of the tone; *Dynamic Range*, described as the pitch range that the subject is able to sing freely throughout, without inappropriate change in voice quality or any external effort; and *Vocal Clarity*, described as vocal vibrations of a clear, well-produced tone. In this work, we explore different features of the audio signal that represent some of these perceptual parameters for singing evaluation, to develop an objective methodology for singing assessment.

1.3.2 *Pronunciation* aspect of singing voice

Automatic pronunciation evaluation of singing is an essential technology in a wide-range of applications. Lyrics play an important role in music, serving as a cue for detecting a song's identity, or its mood or genre [AP06, BAB+11]. Therefore, correctly pronouncing the lyrics of a song becomes an important component of a singing performance. Automatic pronunciation assessment in such a scenario is possible with accurate phonetic modeling for singing voice.

Phonetic modeling of singing voice has many important applications. Singing is shown to be helpful in improving pronunciation in foreign language learning classes [NS11, GRS15b]. Educators recommend singing as a fun and effective language learning aid [DS11]. The use of songs and karaoke is helpful in teaching and improving pronunciation in adult second language (L2) classes [Tea, NS11]. Scientific studies have shown that there is a connection between the ability of phonemic production of a foreign language and singing ability [MPTE10], and singing ability often leads to better imitation of phrases in an unknown language [MR13]. More recently, evidence from experimental psychology suggests that learning a new language through songs helps improve vocabulary gain, memory recall, and pronunciation [GRS15a]. Additionally, singing releases the need to focus on prosody, as melody of the song overrides the prosodic contrasts while singing [Leh04]. So, given a familiar melody, all the attention can

be on articulating the lyrics correctly. Thus automatic phonetic modeling and pronunciation evaluation will help in providing feedback to language learners.

Another application is in speech rehabilitation. Music and speech therapists apply a therapeutic process called Melodic Intonation Therapy (MIT) to treat patients with speech disorders, such as non-fluent aphasia [NZMS09, ASH73]. The therapy is designed based on practicing intoned or sung short phrases. It uses the musical elements of speech (melody & rhythm) to improve expressive language by capitalizing on the preserved capabilities, i.e. singing, and engaging language-capable regions in the undamaged right hemisphere [VDMDSKM+16, NZMS09]. Given the number of therapists is relatively small with regard to the large patient base, it is not feasible for so few therapists to provide timely and regular therapy sessions for all.

Computer-aided pronunciation training (CAPT) for speech has been an active area of research [NFDW00, ZSG+05]. But automatic pronunciation evaluation of singing is still a relatively unexplored area. The state-of-the-art automatic speech recognition (ASR) technology cannot be directly applied for singing pronunciation evaluation because of the mismatch between speech and singing. The acoustic characteristics of singing and speech differ in many ways, such as pitch range, vibrato, and phoneme durations [FG12, LCB99]. Thus to build a pronunciation evaluation algorithm for singing, the ASR needs to be adapted to singing voice. Moreover, the applicability of the traditional speech pronunciation scoring methods for evaluating singing pronunciation needs to be investigated.

1.4 What causes people to sing *badly*?

Before diving deep into the methods of assessing singing quality, one may wonder what causes people to sing "badly". By bad singing, we broadly mean out of tune singing, or incorrect timing/rhythm, or wrong lyrics/pronunciation. There is a lack of formal investigations on this topic, however some informed speculations can be made about it [SR90].

One cause of singing out-of-tune would be that the singer does not really know in advance how the pitch of the next tone sounds, therefore cannot imagine the next pitch before starting to change the pitch. This may be because of unfamiliarity with the song. Another reason would be an imprecise imagination, i.e. the singer imagines the wrong pitch and arrives at it. This could be due to incorrect recollection of the notes of the song. A third reason might be lack of training, i.e. the singer miscalculates the muscular activities needed in order to change the pitch to the imagined next target [SR90].

Out-of-timing is a manifestation of prosody. It often occurs when there is a lack of sense of the rhythm structure of the song, or when the song is unfamiliar. The singer in such cases often stumbles over words and misses the timing.

Out-of-vocabulary words and incorrect pronunciation may occur because of unfamiliarity with the lyrics or if the singer is a non-native speaker of the language.

The way to fix these problems is practice. An automated system for singing quality evaluation is going to aid this process.

1.5 Literature Overview

Objective evaluation of singing has been an area of interest in the recent past. There have been multiple attempts to develop automatic singing quality evaluation algorithms based on prosody-related parameters such as pitch, rhythm, expression, and volume related features, as well as pronunciation-related parameters. Most of the existing methods for prosody-based evaluation are reference dependent, i.e. a test singing sample is compared against an ideal singing sample [Lal06, MBG+13] or the MIDI notes of the song [TL12]. Although MIDI notes approximately represent the sequence of sung notes, they are unable to represent human voice that also comprises of pitch transitions, modulations, and voice timbre. Most of the previous scientific studies and patents have focused on a combination of volume, pitch, and rhythm-based distance measures [MBG+13, TL12, Cha07, Tan99]. However, we need a unified evaluation system that finds the appropriate weighting of all the perceptually relevant parameters while also incorporating ideas of human perception of singing quality.

There have been studies that suggest that there are inherent properties of singing quality that are independent of a reference singer or melody, which help music-experts judge singing quality for an unknown melody or without a reference [NGH06b]. However, there have been limited studies that explore a reference-independent way of evaluating singing quality. Only pitch interval accuracy, and vibrato quality are the parameters that have been explored for reference-independent evaluation of singing quality for prosody-based parameters [NGH06a, NDAL12, BES+17]. A more comprehensive framework for reference-independent evaluation needs to be investigated.

Pronunciation evaluation of singing voice needs reliable acoustic modeling of phonemes in singing. Computer-aided pronunciation training for speech has been an active area of research [NFDW00, ZSG+05]. But automatic pronunciation evaluation of singing is still a relatively unexplored area. Acoustic modeling of sung phonemes has many applications in music information retrieval, such as lyrics transcription, and audio-to-lyrics alignment [MEG14, Kru16b]. However, acoustic modeling in singing face challenges due to the differences between sung and spoken voices [FGOO11, MEG14] that makes the direct use of ASR unsuitable for singing voice. Moreover, there is a lack of transcribed singing data to train phonetic models for singing voice, that has been a bottleneck for research in this field.

A brief overview of the existing studies is provided in Table 1.1. An in-depth study of the literature is done for each of the sub-topics, provided separately under the corresponding

Table 1.1: Summary of literature

Perspective	Topics and Methods	References	Summary
Prosody aspect	Reference-dependent	[Lal06, MBG+13] [TL12, MBG+13] [Cha07, Tan99]	A test singing sample is compared against an ideal singing sample or MIDI notes; Features: a combination of volume, pitch, and rhythm-based distance measures
	Reference-independent	[NGH06b, NGH06a], [NDAL12, BES+17]	Evaluate a test singing sample without relying on a reference; Features: kurtosis and skew of pitch interval accuracy based histogram, and vibrato quality
Pronunciation aspect	Acoustic modeling	[MV08, Kru16a]	Adapting speech acoustic models to singing voice with small data; training singing acoustic models with large but noisy data
	Pronunciation evaluation	[JR12, GS18, PGS17]	Pitch-based vowel classification, musical score- and duration-informed syllable evaluation
	Audio-to-lyrics alignment	[WKN+04, MFG10, Dzh17], [Kru16a, GCOC15]	music-structure informed and, acoustic modeling of sung phonemes based lyrics alignment

chapters of this thesis.

1.6 Goal and Summary

In this thesis, we investigate objective and automatic methods to characterize and assess singing quality. We design techniques to model singing voice and its automatic evaluation methods with respect to prosody and pronunciation, under different plausible contexts.

We have organized the content of this thesis into two parts: In Part I, we discuss the prosody aspect of singing quality, the evaluation techniques, and our proposed methods of evaluation. First in Chapter 2, we explore the idea of evaluating singing quality compared to an ideal singing reference. Next in Chapter 3, we explore the context where an ideal singing reference is not available.

In Part II, we study the lyrics pronunciation or the phonetic aspect of singing voice. In Chapter 4, we develop a method to automatically build acoustic models for sung phonemes. In Chapter 5 and 6, we observe the application of these acoustic models for two use-cases: for assessing the pronunciation quality in singing voice and for lyrics-to-audio alignment, respectively. Finally, we conclude and discuss some of the future work in Chapter 7.

Part I

Prosody aspect of singing voice

Summary

In this part, we discuss our ideas on prosody-based singing quality evaluation. It consists of two chapters: Chapter 2 and 3. Chapter 2 talks about our proposed reference-dependent evaluation techniques, i.e. the methods where we compare a test singer against a reference or ideal singing rendition. In Chapter 3, we discuss the reference-independent methods of singing quality evaluation, where a test singer is automatically evaluated without relying on a reference rendition.

The research findings of Chapter 2 have appeared in two of our publications:

- Chitralekha Gupta, Haizhou Li, and Ye Wang, "Perceptual Evaluation of Singing Quality", *In Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Kuala Lumpur, 2017
- Chitralekha Gupta, Haizhou Li, and Ye Wang, "A Technical Framework for Automatic Perceptual Evaluation of Singing Quality", *Transactions of APSIPA*, 2018.

And parts of Chapter 3 were reflected in the following publications:

- Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic Evaluation of Singing Quality without a Reference", *In Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Hawaii, 2018
- Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic Leaderboard: Evaluation of singing quality without a standard reference", to be submitted to IEEE Transactions on Audio, Speech, and Language Processing.

Chapter 2

Reference-based Singing Quality Evaluation

Singing quality assessment often refers to the degree to which a particular vocal production meets professional standards of excellence. Music experts assess singing on the basis of musical knowledge and perceptual relevance. However, personal biases and preferences are unavoidable in human judgment. Therefore, an automatic objective singing quality assessment framework can be a useful tool in this context, and could have applications such as karaoke singing scoring, and a practice tool for singing learners.

As discussed in Section 1.3.1, Chuan et al. [CLLY08] defined six perceptual parameters that characterize the main prosody-related properties of singing for assessing non-trained singers: *Intonation accuracy*, *Rhythm consistency*, *Appropriate vibrato*, *Timbre brightness*, *Dynamic Range* and *Vocal Clarity*. In this work, we provide a unified evaluation framework that applies music theory and human perception to find the appropriate acoustic features for all of these perceptual parameters to obtain the overall evaluation score called Perceptual Evaluation of Singing Quality (PESnQ) score. Furthermore, we provide a systematic analysis of PESnQ to further answer the following research questions:

1. **Study of human perception for singing quality judgment:** In what way do humans combine the assessment of the perceptual parameters to give an overall singing quality score? Do the human scores for the individual perceptual parameters, i.e. intonation accuracy, rhythm consistency, appropriate vibrato etc. combine to predict the overall singing quality score?
2. **Prediction of perceptual parameters:** Based on the acoustic features, can a machine predict the perceptual parameters individually?
3. **Strategies for overall singing quality scoring:** Can human perception inspired strategy for singing quality assessment result in better scoring? We study and compare two different strategies for scoring: early fusion and late fusion of acoustic features.

The singing evaluation techniques considered in this chapter are reference-based methods, i.e. a test sample is compared against a template or reference sample, where the template can be either MIDI notes or a reference ideal singing rendition, which in our case is reference singing.

In Section 2.1, we present a critical review of the related previous work and techniques on the prosody-related aspect of singing quality evaluation, the challenges in the area, and formulate the problem of perceptual singing assessment, while also summarizing our contributions. In Section 2.2, we discuss how singing quality is characterized along with our feature design approach. In Section 2.3, we describe our framework of evaluation. Section 2.4 describes our experiment methodology of subjective and objective evaluation, and discusses our experiment results. Section 2.5 provides a summary of this study and suggestions for the future work.

2.1 Background studies and our contributions

Objective evaluation of singing has been an area of interest in the recent past. There have been multiple attempts to develop automatic singing quality evaluation algorithms based on prosody-related parameters such as pitch, rhythm, expression, and volume related features. But there are some technical challenges in each of these algorithms, that we will try to address in this work.

Intonation accuracy or pitch accuracy evaluation is the most common method for singing assessment. In one study, Lal [Lal06] proposed a pitch-based similarity measure to compare a test singing clip to the reference singing clip. But reliable and automatic pitch estimation is a challenging task, and errors in pitch estimation can result in incorrect automatic score. In another study, Tsai and Lee [TL12] proposed an automatic evaluation system for karaoke singing in which they compared MIDI (Musical Instrument Digital Interface) notes of test singing to that of the intended reference song to compute the pitch accuracy rating. Although MIDI notes approximately represent the sequence of sung notes, they are unable to represent human voice. Apart from steady notes, singing voice comprises of pitch transitions, modulations, and voice timbre, that are not captured by digitally generated MIDI notes. Also, when singing without background accompaniments, the singers tend to sing at a key they are comfortable in, which may or may not be the same as that of the reference song. In such a scenario, singing the correct sequence of notes with a key transposition should not be penalized [CLLY08]. The possibility of key transposition has not been considered in [TL12], because the key of a song is inherently fixed in karaoke singing due to background accompaniments.

Rhythm consistency is another important feature for singing evaluation. Tsai and Lee [TL12] evaluated rhythm by comparing the note-onset strength of the background accompaniment of karaoke to that of the test singing. But when we consider the case of singing without background accompaniments (or free singing), such methods cannot be directly applied. Similar to key transposition, in free singing, the singers can also show a slight tempo variation from the reference singing, i.e. slightly but uniformly faster or slower rhythm than the reference song. Such tempo variations should not be penalized [CLLY08]. Molina et

al. [MBG+13] and Lin et al. [LLCW14] measured rhythm accuracy without penalizing for a rhythm different from the reference. They evaluated rhythm by aligning test pitch contour with the reference pitch contour using Dynamic Time Warping (DTW), and obtained the rhythm score by computing the deviation of the optimal path from a straight line fit in the cost matrix of the DTW between the pitch contours. This line-fit may be different from the ideal 45 degree straight line, in turn compensating for tempo difference. But aligning test and reference singing using pitch contour makes rhythm assessment dependent on pitch correctness. This poses a problem if the test singer sings with inaccurate pitch (off-tune) but maintains a good rhythm. Inaccurate pitch estimation also creates the same problem. This method will give large deviation from the optimal path due to pitch inaccuracies, despite good rhythm.

Expressive elements such as *appropriate vibrato* are considered to be important cues to distinguish between a well-trained singer and a mediocre singer. Nakano et al. [NGH06a] computed acoustic features which are independent from specific characteristics of the singer or melody, such as vibrato features like rate and extent of pitch undulations, to evaluate singing in a case that has no reference singing. But while learning to sing a song, one would try to match their singing with a reference singing in every way possible. In such a case, vibrato evaluation in the presence of reference singing is needed. Vibrato detection and evaluation will also be affected by pitch estimation errors.

Timbre brightness is defined as the ring or brilliance of a tone [CLLY08], which often relates to voice quality. Singing power ratio (SPR), which is the ratio of highest spectral peak between 2 and 4 kHz and the highest spectral peak between 0 and 2 kHz in voiced segments, has been used previously to separate professional singers and non singers [OKC+96]. But as pointed out by Tsai and Lee [TL12], the ringing voice quality which is indicated by high SPR, is typically observed in operatic style of singing. However, operatic style is a specific way of singing that, one may argue, can be unsuitable and undesirable for singing lessons or karaoke performances, especially for beginners. Hence our work here does not consider SPR as a parameter for automatic singing evaluation. Prasert et al. [Pra08] developed a more general method to evaluate voice quality in singing based on timbral features, such as Mel Frequency Cepstral Coefficients (MFCC) and Filter Banks (FBANK), and found that MFCCs performed better. We will consider this direction in our study.

As illustrated in [TL12], most of the singing evaluation studies have been reported in patent documentation that do not discuss the rationale of their evaluation methods, and fail to show results of their qualitative analysis to validate their methods. Comparatively, the number of scientific studies in this area is fewer. Literature suggests that a combination of the various perceptual parameters, as described in [CLLY08], would result in the final judgment of a test singing clip. But both, the patents and the scientific studies, have managed to incorporate a set of objective acoustic cues that are relevant to only a subset of the perceptual parameters for singing evaluation. For example, patents such as [Tan99, Cha07] have used a combination

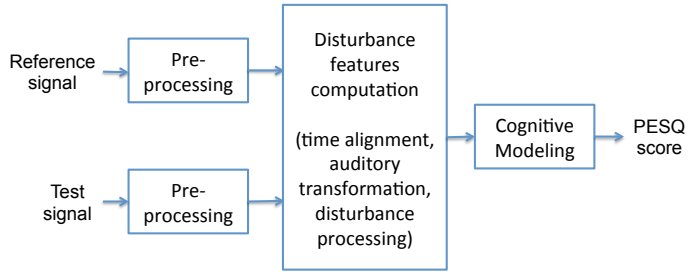


Figure 2.1: Overview of PESQ computation

of volume and pitch as evaluation features, while scientific studies such as [TL12] have used pitch, volume, and rhythm features. We need a unified evaluation system that finds the appropriate weighting of all the perceptually relevant parameters to obtain the final score.

Our idea of perceptual assessment of singing quality is motivated by the International Telecommunication Union (ITU) standard for quality assessment of speech in telephone networks, PESQ (Perceptual Evaluation of Speech Quality) [RBHH01]. PESQ is obtained by comparing the original speech signal with its degraded version (test signal), that is the result of passing the original signal through a communication channel, and predicting the perceived quality of the degraded signal, as shown in Figure 2.1. We note that objective measurement of signal quality doesn't always correlate with human perception. The ITU benchmark experiments report an average correlation of 0.935 between PESQ scores and human scores, that make PESQ an ideal objective metric. According to cognitive modeling literature, *localised errors* dominate perception of audio quality [HHG94], i.e. a highly concentrated error in time and frequency is found to have a greater subjective impact than a distributed error. This concept has been successfully used in assessing speech signal quality (PESQ), by using a higher weightage for localised distortions in PESQ score computation. Motivated by this approach, we apply this concept of audio quality perception in our work to obtain a novel singing quality assessment method.

2.1.1 Our contributions

- We improve our understanding about the importance of the perceptual parameters of singing skill evaluation and their representation in human mind to predict the singing quality scores
- We design a perceptually relevant score for singing quality called PESnQ, that adapts the cognitive modeling theory of PESQ for singing voice, providing aspiring singers a comprehensive and musically relevant objective feedback to help them improve their singing skills

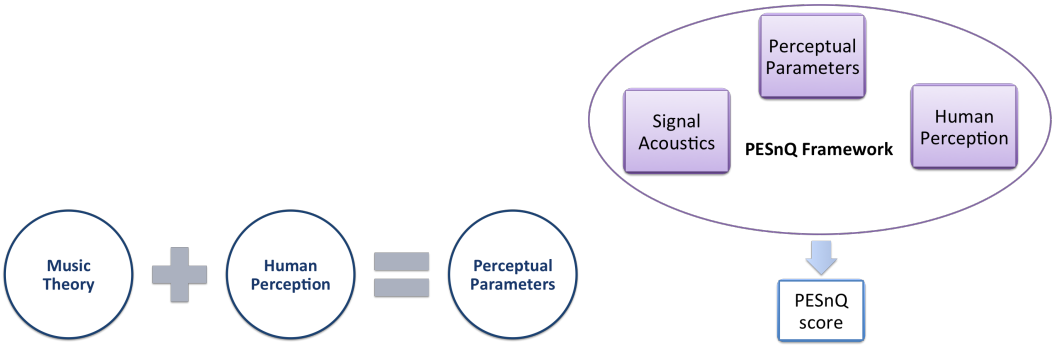


Figure 2.2: The concept of the PESnQ framework. The perceptual parameters are motivated by the rules of singing as dictated by music theory and music-experts human perception studies. Our proposed PESnQ framework comprises of elements from these perceptual parameters, along with signal acoustics, and an understanding of human perceptual judgment process to obtain a perceptually-valid score for singing quality called the PESnQ score.

2.2 Singing quality characterization and evaluation

In this study, we aim to develop a holistic scoring framework for automatic singing evaluation based on perceptual parameters, as recommended by music educators. We explore various low-level descriptors or acoustic features to characterize intonation, rhythm, vibrato, timbre, volume, and pitch dynamic range, to develop a measure for evaluating singing skill of a test singer as compared to an ideal reference singing of a song. We introduce methods to overcome the challenges of key-transposition, and rhythm variation. Moreover, we adopt the cognitive modeling theory from PESQ for singing quality evaluation. We believe that the evaluation framework should emulate human perception of judgment which will lead to an improvement in the performance of automatic singing quality scoring. We have termed it as Perceptual Evaluation of Singing Quality (PESnQ) score. We compare our results with the known baseline methods for singing evaluation. Figure 2.2 summarizes the idea of PESnQ. The perceptual parameters such as intonation accuracy, rhythm consistency, etc. are derived from music theory and human perception studies from music-experts. Our PESnQ computation framework comprises of information from these perceptual parameters, along with the characterization of signal acoustics and understanding of the human perceptual judgment process.

In this section, we characterize the perceptual parameters identified by human experts using various acoustic features. Also evaluation is the distance between the target and the reference singing characteristics. We describe the distance parameters defined and used for evaluation.

2.2.1 Intonation accuracy

Pitch is an auditory sensation in which a listener assigns musical tones to relative positions on a musical scale based primarily on their perception of the frequency of vibration [POF06]. Pitch of a musical note is characterized by the fundamental frequency $F0$ and its movements between high and low values. Intonation accuracy or "singing in tune" is directly related to the correctness of the pitch produced in comparison to reference singing. For developing an automatic system that evaluates intonation accuracy, estimation of reliable pitch contours becomes very important. Pitch estimation is an active research area, and various algorithms have been developed for pitch estimation in monophonic speech signals, such as ACF (autocorrelation function)[Rab77], YIN [YHP08], etc. But these methods need adaptations and post-processing to accurately detect pitch in singing waveforms. Babacan et al. [BDd+13] compared the different pitch detection algorithms for monophonic singing, and found that parameter settings specific to singing, such as increasing the $F0$ search range to account for wide vocal range of singing, as well as applying post-processing to pitch estimates lead to better pitch estimates. They also found that the autocorrelation-based PRAAT [B+02] pitch estimator gives best voicing boundaries even without post-processing, while the source-filter model-based STRAIGHT [KEF01] pitch estimator is the most robust algorithm in noisy conditions. The modified autocorrelation-based estimator YIN [YHP08] achieves the best accuracy of pitch detection but it requires a number of post-processing steps depending on the properties of the music type being analysed, as described in [Boz08].

In our work, we use the pitch estimates from PRAAT, with one generic post-processing step to remove unreliable pitch values. We first use the pitch estimates to determine the voicing boundaries, compute the pitch estimates over all the voiced frames, and then remove the frames with low periodicity, which is determined by harmonic-to-noise ratio (HNR). HNR, also computed in PRAAT, represents the degree of acoustic periodicity expressed in dB. For example, if 99% of the energy of the signal is in the periodic part, and 1% is noise, the HNR is $10\log_{10} 99/1 = 19.95\,\text{dB}$. In determining the valid pitch frames, we remove the ones with $< 98\%$ of energy in periodic part, i.e. $HNR < 10\log_{10} 98/2 \approx 16.9\,\text{dB}$. This threshold is set empirically. By choosing only the voiced segments and removing the frames with low periodicity, spurious $F0$ values are avoided and only reliable pitch values are used. Figure 2.3 shows an example of pitch contour before and after periodicity-based pitch clean-up. All pitch values in this study are calculated in the unit of cents (one semitone being 100 cents on equi-tempered octave) defined as,

$$p_{\text{cent}} = 1200 \times \log_2 \frac{p_{\text{Hz}}}{440}, \tag{2.1}$$

where 440 Hz (pitch-standard musical note A4) is considered as the base frequency.

For singing quality evaluation in terms of intonation accuracy, we first time-align the reference and test singing by using the alignment from DTW between their MFCC vectors. This

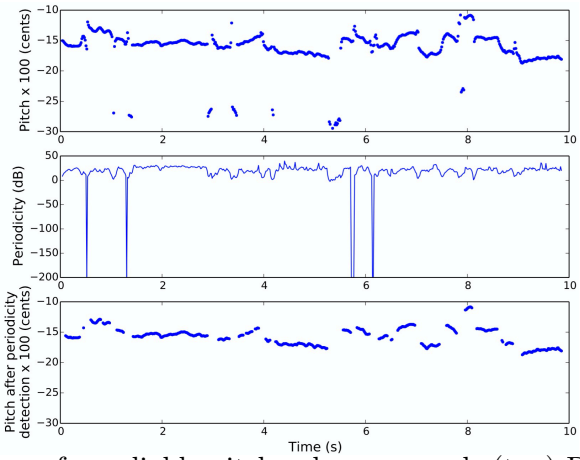


Figure 2.3: Illustration of unreliable pitch values removal. (top) Pitch contour extracted from voiced segments using PRAAT, (middle) periodicity values in dB, (bottom) pitch contour, after removal of low periodicity pitch values

compensates for any tempo differences or tempo errors between reference and test. Then, we compute the DTW distance between the pitch contours of the reference and test singing (termed as *pitch_ dist*) for evaluation, which would be an indicator of *intonation accuracy*, as previously used in [MBG+13, TL12, Lal06]. But this distance between pitch contours will penalise key transposition, although key transposition is allowed in case of singing without background accompaniments [CLLY08]. Hence we use two different methods to make the distance measure insensitive to key transposition:

Pitch Derivative

Derivatives of pitch contours of both the reference and the test singing make the resultant contours independent of key shifts. The derivative also emphasizes on the transitions between notes, in terms of the magnitude as well as the duration of the change. Note transition expressions, such as glissando, are considered to be a significant indicator of good singing, that get captured by this feature. For a pitch vector $\mathbf{p_a} = \begin{bmatrix} p_1 & p_2 & ... & p_N \end{bmatrix}^T$, where N is the number of frames, one frame derivative $\vec{\Delta p}$ is computed as

$$\mathbf{\Delta p} = \mathbf{p_a} - \mathbf{p_b}, \tag{2.2}$$

where $\vec{p_b}$ is the pitch vector shifted by one frame. Figure 2.4 shows an example of a pitch contour and its derivative.

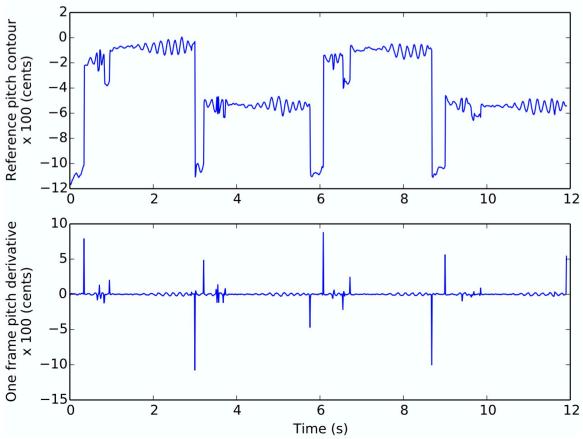


Figure 2.4: (top) Pitch contour extracted from voiced segments using PRAAT and after removal of low periodicity frames, and (bottom) corresponding pitch contour derivative with one frame shift.

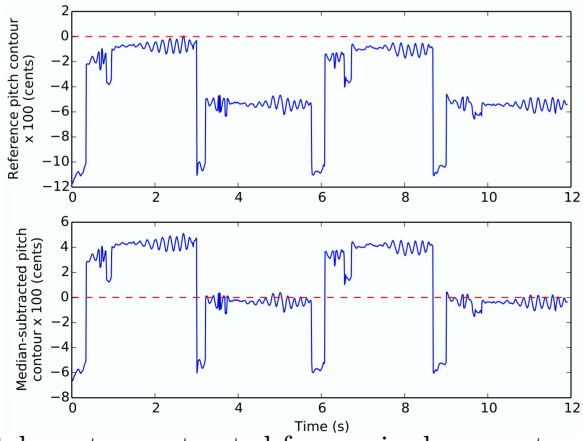


Figure 2.5: (top) Pitch contour extracted from voiced segments using PRAAT and after removal of low periodicity frames, and (bottom) corresponding median-subtracted pitch contour.

Median-subtracted Pitch

Subtracting the median of the pitch values of an audio segment is another way to make the pitch contour independent of key-transposition. Here, median is preferred over mean because averaging over all pitch values might get affected by infrequent outlier pitch values, which is avoided by the median. The median-subtracted pitch for a pitch vector $\vec{p}$, is computed as

$$p_{\vec{medsub}} = \vec{p} - \text{median}\{\vec{p}\}. \tag{2.3}$$

Figure 2.5 shows an example of a pitch contour and its median-subtracted version.

We apply the cognitive modeling theory to these frame-level modified pitch vectors (pitch-derivative and median-subtracted pitch) to obtain the pitch evaluation between reference and test singing (Section 2.3.1).

2.2.2 Rhythm Consistency

Rhythm is defined as the regular repeated pattern in music, that relates to the timing of the notes sung, and is often referred to as tempo. Rhythm consistency refers to the similarity of tempo between reference and test singing. As mentioned earlier, a slight variation in tempo is allowed, i.e. uniformly faster or slower tempo compared to the reference. Molina et al. [MBG+13] proposed DTW as a procedure for automatic rhythm assessment, and accounted for rhythm variation. They computed the DTW between the reference and the test pitch contours, and analyzed the shape of the optimal path in the cost matrix of DTW. A 45° straight line would represent a perfect rhythmic performance with respect to reference melody, while straight line with an angle $\neq 45°$ would represent good rhythmic performance in a different tempo. So they fit a straight line on the optimal path in the cost matrix of the DTW, and computed the root mean square error of this straight line fit from the optimal path (termed as *molina_rhythm_pitch_dist*),

$$\epsilon = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\epsilon_k^2}, \tag{2.4}$$

where ϵ_k is the error in linear fit at frame k, and N is the total number of frames.

But aligning test and reference singing using pitch contour makes rhythm assessment dependent on pitch correctness. So if the singer maintains a good rhythm but sings with inaccurate pitch, this algorithm will give a poor score for rhythm. Thus this method works well only when the test singing pitch is same as that of reference singing, even if words are spoken incorrectly. But this method will give a large deviation from the optimal path if the pitch is inaccurate, despite good rhythm.

We propose a modified version of Molina's rhythm deviation measure. Instead of using pitch contour, we use 13 MFCC feature vectors to compute DTW between reference and test singing. MFCCs capture the short-term power spectrum of the audio signal that represents the shape of the vocal tract and thus the phonemes uttered. So when we compute DTW between MFCC vectors, we assume that the sequences of phonemes and words are uttered correctly, thus making this measure independent of off-tune pitch. So we obtain a modified Molina's rhythm deviation measure (termed as *molina_rhythm_mfcc_dist*) that measures the root mean square error (Equation 2.4) of the linear fit of the optimal path of DTW matrix computed using MFCC vectors.

We also introduce another way to compute rhythm deviation, while accounting for allowable rhythm variations. We compute 13 MFCC vectors over a 32 ms long window for every 16 ms of the reference singing, and then compute the corresponding frame rate for the test singing such that the number of frames in reference and test are the same. This way we compensate for constant rhythm difference between reference and test singing, and thus the number of

MFCC vectors in reference and test are equal. Then we apply cognitive modeling theory to these frame-equalized MFCC feature vectors to obtain the rhythm evaluation between reference and test singing (see Section 2.3.1).

2.2.3 Voice Quality

Timbre is related to the voice quality and describes the perceived quality of a tone produced by the singer. Perception of timbre is physically represented by spectral envelope of the sound, which, as mentioned earlier, is captured well by MFCC vectors, as illustrated in [Pra08]. MFCCs also represent phonetic quality, which relates to pronunciation. Thus, we compute the distance between reference and test singing timbre (termed as *timbral_dist*) by computing the DTW distance between their 13 MFCC vectors.

2.2.4 Appropriate Vibrato

Vibrato is the rapid periodic undulations in pitch on a steady note while singing. Studies have found that vibrato oscillations are within 5-8 Hz, and their extent is between 30-150 cents [SR90]. Vibrato is considered to be a fair indicator of the quality of singing, hence we would like to evaluate it. For a fully automated evaluation system, the idea is to first detect the vibrato sections in the reference, then find the corresponding time-aligned pitch segments in the test, and finally compute measures to compare the reference and test vibrato segments. Another way could be to compare vibrato-specific feature vectors of every frame from test and reference. However, the frames in test that correspond to those in reference that do not contain vibrato, should not be given a high score, as we are not giving marks for "absence of vibrato". Thus detection of vibrato sections as the first step is necessary.

Nakano et al. [NGH06a] applied short-term Fourier transform to the first order differential of $F0$ and defined vibrato likeliness $P_v(t)$ as the product of power $\Psi_v(t)$ and sharpness $S_v(t)$ as:

$$\Psi_v(t) = \int_{F_L}^{F_H} \hat{X}(f,t)df, \qquad S_v(t) = \int_{F_L}^{F_H} \hat{X}(f,t)f\,df, \tag{2.5}$$

$$P_v(t) = \Psi_v(t)S_v(t), \tag{2.6}$$

where (F_L,F_H) is the range of vibrato rate set as 5 and 8 Hz respectively, and $\hat{X}(f,t)$ is the power spectrum $X(f,t)$ normalized over f:

$$\hat{X}(f,t) = \frac{X(f,t)}{\int X(f,t)df}. \tag{2.7}$$

If the value of vibrato likeliness is greater than a threshold, the section is detected as *vibrato section*. However, the problem with this measure of vibrato likeliness is that the obtained

likeliness values are not normalized, which makes it difficult to set a singer-independent threshold for vibrato detection. In this study, we have modified the vibrato likeliness measure as the ratio of energy in the power spectrum of $F0$, $X(f,t)$, between 5 to 8 Hz (F_L,F_H) to the total energy in the spectrum (Equation 2.8). A similar feature was used by Amir et al. [AEGF09].

$$P_{v_{\text{mod}}}(t) = \frac{\int_{F_L}^{F_H} X(f,t)df}{\int X(f,t)df}. \tag{2.8}$$

This measure gives a normalized score between 0 and 1, unlike the score obtained by Nakano et al. Also it is a good indicator of concentration of energy in the vibrato oscillation frequency range. We compute this modified vibrato likeliness score over every 512 ms frame (i.e. 32 samples, similar to [NGH06a]) of the reference singing segment, and empirically set a threshold of 0.4 to detect the valid vibrato segments in the reference singing as shown in Figure 2.6.

We map the time stamps of the detected vibrato segments in reference to that of the aligned test pitch contour to obtain potential vibrato segments in the test. For these segments, we compute three vibrato-related features - modified vibrato likeliness ($P_{v_{\text{mod}}}(t)$ from Equation 2.8), extent, and rate. The extent and rate features are the ones defined by Nakano et al.:

$$\frac{1}{\text{rate}} = \frac{1}{N}\sum_{n=1}^{N} R_n \qquad\qquad \text{extent} = \frac{1}{2N}\sum_{n=1}^{N} E_n, \tag{2.9}$$

where R_n (in seconds) is the time period of n^{th} oscillation, computed as the difference between alternate zero-crossing time instants, and E_n (in cents) is the difference between the maximum and the minimum pitch value in the n^{th} oscillation. As a post-processing step, we discard any detected reference vibrato section from vibrato evaluation that does not have at least one whole oscillation present. Thus we have modified vibrato likeliness, rate, and extent features for every valid reference vibrato section and corresponding test pitch section. We compute the Euclidean distance of these features between the reference and the test to obtain the vibrato distance score (termed as *vib_segment_dist*) for evaluation.

2.2.5 Volume

Dynamics of volume reflect the relative loudness or softness of different parts of the song. It is expected that there will be a similar pattern of volume variations across time when different singers perform the same song [TL12]. Apart from Tsai and Lee's work, various singing evaluation patents have incorporated volume as an acoustic cue in their systems [Tan99, Cha07]. In this study, we implement the volume feature used by Tsai and Lee's system, i.e. short-term log energy over 30 ms window, and then compute the DTW distance of this feature between the reference and the test (termed as *volume_dist*) for evaluation.

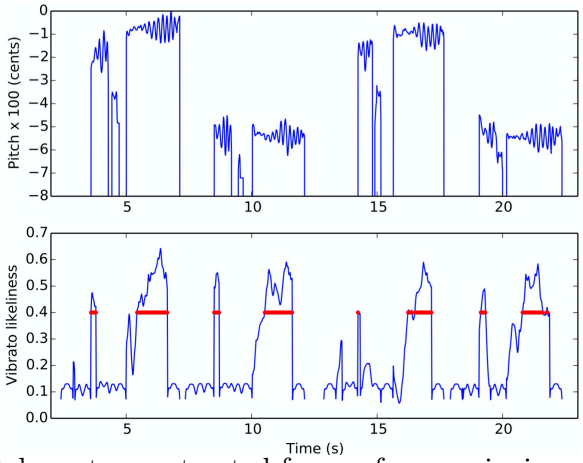


Figure 2.6: (top) Pitch contour extracted from reference singing, and (bottom) modified vibrato likeliness $P_{v_{\text{mod}}}(t)$, vibrato sections marked in red.

2.2.6 Pitch Dynamic Range

The pitch range that a subject is able to sing freely throughout is a good indicator of quality of singing [CLLY08]. Thus we compute the absolute difference between the highest and the lowest pitch values in an audio segment as a feature for pitch dynamic range. The distance of this feature between reference and test singing (termed as *pitch_ dynamic_ dist*) is an indicator of the similarity of the test singing range to the expected singing range, and is used for singing quality evaluation. We note that the pitch dynamic range is also a function of the song being sung. However, in this study, we only compare the pitch range of a test singer to that of a reference singer for any given song. Although this feature captures the ability of the singer in singing the pitch range of the given song, it does not indicate the general pitch range of the singer.

2.3 Framework of evaluation

Singing quality evaluation was once considered to be a highly subjective task. But in the recent years, research has convincingly provided perceptual parameters based on which humans evaluate singing quality. These perceptual parameters are defined on the basis of music theory and human perception, as discussed in the previous section. For example, the perceptual parameter *intonation accuracy* involves the concepts of musical notes, pitch, and what a music-expert perceives as accurate intonation. The aim of an automatic singing evaluation framework is to objectively characterize these perceptual parameters based on signal acoustics to predict the singing quality score as would be given by music-experts based on their perception.

The framework for automatic singing quality evaluation can be broadly considered from

two perspectives: human perception and signal acoustics. Human perception of singing quality is the subjective ground truth, whereas the signal acoustics provide the objective characterization of the perceptual parameters. The aim of the evaluation framework is to build a bridge between the objective characterizations and the subjective judgments. In this section, we discuss the strategies for building this framework with respect to these two perspectives.

2.3.1 Human Perception

Our singing quality evaluation framework is motivated by human perceptual relevance. Our goal is to teach a machine to appreciate singing quality in the same way as human music-experts would do. Thus, based on the identified perceptual parameters for singing judgments, we have considered two different aspects of human perceptual judgments:

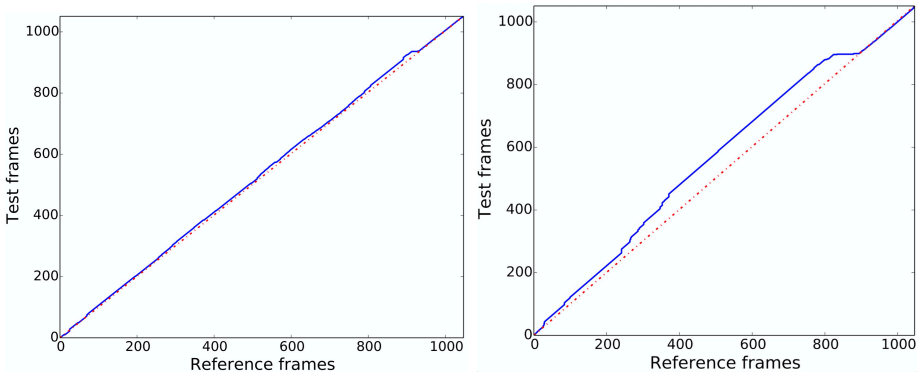
Cognitive Modeling: localized vs. distributed errors

The International Telecommunication Union (ITU) standard for quality assessment of speech in telephone networks, PESQ (Perceptual Evaluation of Speech Quality) [RBHH01] incorporates the cognitive modeling theory that a *localized error* in time has a larger subjective impact than a distributed error [HHG94]. PESQ combines the frame-level disturbance values of an audio file by computing the L_6 norm over split-second intervals, i.e. over 20 frames (320 ms) window (with 50% overlap and no window function), and the L_2 norm over all these split-second disturbance values over the length of the speech file. The value of p in L_p norm is higher for averaging over split-second intervals, to give more weightage to localized disturbances. L_p norm is computed as:

$$L_p \text{ norm} = \left(\frac{1}{\mathrm{N}} \sum_{\mathrm{m}=1}^{\mathrm{N}} \text{disturbance}[\mathrm{m}]^{\mathrm{p}} \right)^{\frac{1}{\mathrm{p}}}. \tag{2.10}$$

where N is the total number of disturbance values, over index m. Similarly in singing, errors are time-localized; for example, only certain notes may become off-tune or only certain sections may be sung with bad rhythm. Therefore, in this study we explore the possibility of applying the same cognitive modeling concept as in PESQ, for singing quality evaluation.

We first compute the frame-level disturbance values of the following singing features: pitch derivative $\mathbf{\Delta p}$, median-subtracted pitch $p_{\vec{medsub}}$, and frame-equalized MFCC feature vectors for rhythm. That is, we compute the optimal path in the cost matrix of DTW between the respective feature vectors of reference and test. If the pitch or rhythm in test singing matches with that of the reference, it would give a 45° optimal path in the corresponding DTW cost matrix. Figure 2.7 illustrates the optimal path of singing with good and poor rhythm accuracy. Deviation of the best alignment path from the diagonal represents error



(a) (b)

Figure 2.7: Optimal path in DTW cost matrix for (a) good rhythm (b) poor rhythm. Red broken diagonal line shows the ideal rhythm.

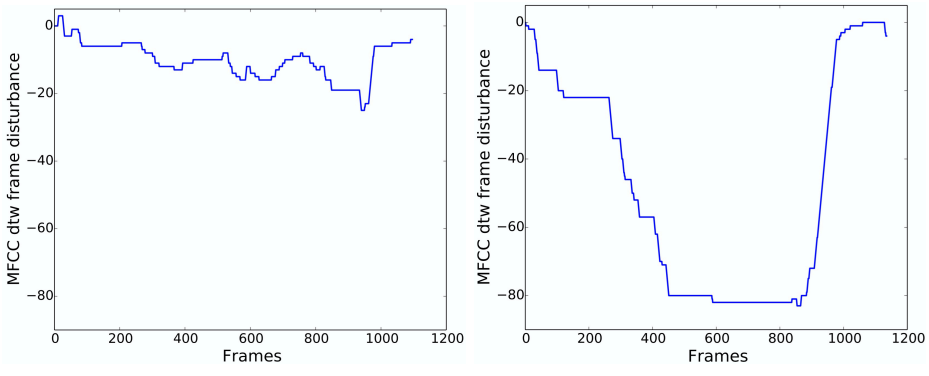
in that characteristic (pitch or rhythm). We compute the deviation of the optimal path from the ideal diagonal path for every frame, and term it as the *frame-level disturbance*. We compute the frame-level disturbance vectors for the different features: *rhythm frame disturbance* R_d, *pitch derivative frame disturbance* $P_{d_{\mathbf{\Delta p}}}$, and *median-subtracted pitch frame disturbance* $P_{d_{\vec{p}_{medsub}}}$. Larger deviations indicate poor intonation/rhythm accuracy. Figure 2.8 shows an example of the frame disturbance of good and poor rhythm.

Next, we compute the L_6 norm over split-second intervals and L_2 norm over all these split-second disturbance values over the length of the sung file for all of the frame-level disturbance values mentioned above - $P_{d_{\mathbf{\Delta p}}}$ (termed as *pitch_der_L6_L2*), $P_{d_{\vec{p}_{medsub}}}$ (termed as *pitch_med_L6_L2*), and R_d (termed as *rhythm_L6_L2*). And for performance comparison, we also compute the L_2 norm of all the disturbance values over the entire file, to observe the effectiveness of the cognitive modeling method for singing evaluation. To summarize, we compute various acoustic features from reference and test singing vocals and compare them to obtain three kinds of distance features: L_2 norm, $L_6 + L_2$ norm (PESQ-based), and DTW distance (feature groups: L2, L6+L2, and dist respectively). The summary of evaluation features is listed in Table 2.1.

We define *Perceptual Evaluation of Singing Quality* (PESnQ) as the score generated from a system comprising of a combination of PESQ-based, L_2 norm, and DTW based distance features. In this work, we will explore different combinations of these features to build various singing evaluation systems and investigate the factors that can impact their performance, such as type and definition of features, the PESQ-based perceptual distance features, and their combinations.

Table 2.1: Acoustic features, distance features and their description corresponding to the human perceptual parameters for singing quality evaluation.

Perceptual Parameters	**Acoustic Features**	**Distance Features**	**Description**
A) Intonation accuracy	pitch contour	pitch_dist	DTW distance between pitch contours
	pitch contour derivative **Δp**	pitch_der_L2	L2-norm of frame disturbances of DTW between pitch derivative contours
		pitch_der_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between pitch contour derivatives
		pitch_der_dist	DTW distance between pitch derivative contours
	median subtracted pitch contour $\vec{p}_{medsub}$	pitch_med_L2	L2-norm of frame disturbances of DTW between median subtracted pitch contour
		pitch_med_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between median subtracted pitch contour
		pitch_med_dist	DTW distance between median-subtracted pitch contours
B) Rhythm consistency	pitch contour	molina_rhythm_pitch_dist	Rhythm distance computed by the method in [MBG+13]
	MFCC	rhythm_L6_L2	L6+L2-norm (PESQ method) of frame disturbances of DTW between MFCC vectors
		rhythm_L2	L2-norm of frame disturbances of DTW between MFCC vectors
		molina_rhythm_mfcc_dist	Modified version of the method in [MBG+13] by computing rhythm distance using mfcc vectors instead of pitch
C) Voice quality & Pronunciation	MFCC	timbral_dist	DTW distance between MFCC features
D) Appropriate Vibrato	pitch contour	vib_segment_dist	DTW distance between vibrato features of only the valid vibrato segments
E) Volume	energy contour	volume_dist	DTW between log energy contours
F) Pitch Dynamic Range	pitch contour	pitch_dynamic_dist	Comparison of the difference between max and min pitch values



(a) (b)

Figure 2.8: Frame disturbance for (a) good rhythm (b) poor rhythm.

Human Parametric Judgment Equation

In the recent years, research has provided various perceptual parameters that explain the physical and acoustic implications of human judgments. One such work [CLLY08] investigates the importance of every perceptual parameter and its contribution to the overall assessment of singing clips. They found that *intonation accuracy* is the most important perceptual contributor to human judgment, when assessing untrained singers.

In the cognitive psychology field of judgment and decision making, studies have found that people often construct a mental representation of the object, person, or situation about which they are making a judgment [HP95]. Before making a judgment, people construct an internally consistent model that explains as much of the judgment-relevant evidence as possible. This model is central to the judgment process, and plays a causal role in determining the ultimate decision and the degree of confidence that is associated with the decision.

According to music psychology studies, song perception-production in humans is a two-stage process. Humans first convert the perceived audio into a mental *symbolic representation*, and then convert it into a vocal-motor representation to produce the same sound [HM13]. This means that the underlying mental process of perception is in the form of a symbolic representation, similar to the cognitive psychology studies. This leads us to our hypothesis that humans follow a similar two-stage process to judge the overall singing quality. Humans first convert the perceived singing audio into a weighted representation of the identified perceptual parameters as the *symbolic representation*, and this representation is then mapped to the overall singing quality judgment score.

In cognitive algebra, such mental processes that determine human judgments are modeled as equations [HP95, Hof60, SBN90]. For example, Hoffman [Hof60] fitted linear regression models to predict judgments on the basis of five to ten cues presented in the cases to be judged, concluding that the observable judgments were well fit by algebraic equations. This motivates us to explore the possibility of expressing overall singing quality judgment score in the form of a linear parametric equation in terms of the perceptual parameters. In this paper, we will validate this theory and formulate a parametric equation that models the overall human perceptual scoring as a function of a set of human perceptual parameters.

We obtain human judgments for overall singing quality as well as for the seven perceptual parameters relevant to singing quality, that are summarized in Table 2.2 and discussed in detail in Section 2.2. Our hypothesis is that the overall singing quality score can be approximated by a linear combination of the perceptual parameter scores, that will look like

Table 2.2: List of perceptual parameters.

Perceptual Parameters	**Symbol**
Intonation Accuracy	P
Rhythm Consistency	R
Appropriate Vibrato	Vib
Volume	Vol
Voice Quality	$VQual$
Pronunciation	$Pronun$
Pith Dynamic Range	PDR

this:

$$\begin{aligned} score = C_1 \times P + C_2 \times R + C_3 \times Vib + \\ C_4 \times Vol + C_5 \times VQual + C_6 \times Pronun \\ + C_7 \times PDR \end{aligned} \tag{2.11}$$

We train the linear regression model given in Eq.(2.11) to obtain the weights C_i of each of these parameters as discussed further in the experiment in Section 2.4.

2.3.2 Signal Acoustics

Literature shows that human judgment of overall singing quality is based on a set of perceptual parameters such as intonation accuracy, rhythm consistency, etc. [CLLY08]. These perceptual parameters can be represented by acoustic features extracted from the singing signal, as discussed in Section 2.2. However, no studies have elaborated on how these perceptual parameters and the signal acoustics map to the overall singing quality judgment score.

Psychology studies show that the human speech/singing perception-production model converts the perceived audio into some form of symbolic representation, and then converts it into vocal-motor representation to produce or mimic the sound [HM13]. This indicates that the human perceptual model converts an audio signal into a parametric representation before making a judgment. Based on this theory, we explore methods of obtaining the singing quality judgment from the signal acoustics.

Figure 2.9 summarizes our evaluation framework. The identified perceptual parameters for singing quality assessment is objectively represented through acoustic features, as discussed in Section 2.2. We apply the cognitive modeling theory and compute various distance features between the acoustic features from the reference and the test singing vocals, as described in

Section 2.3.1. We explore two methods of mapping these objective features to the human judgments: the *early fusion method*, where the features are directly mapped to the overall singing quality judgment; and the *late fusion method*, where the features are mapped to the perceptual parameters, that are further mapped to the overall judgment.

Early Fusion

The idea of this method is to combine the distance features directly to predict human overall singing quality judgment score. This method is the standard way of computing the overall singing quality judgment score as reported in [MBG+13, TL12, Lal06]. In this work, we generate the overall singing quality judgment score from different combinations of the cognitive model-based and DTW distance-based features. We report and compare the performance of various combinations of these features.

Late Fusion

This method is inspired from the human perception-production model. As discussed in Section 2.3.1, we believe that humans first convert the perceived singing audio into a weighted representation of the identified perceptual parameters, which is then mapped to the overall singing quality judgment score. In the same manner, we propose that our machine should first predict the perceptual parameters independently with the help of the distance features, and then apply the human parametric judgment equation from Section 2.3.1 to fuse these perceptual parameters to give the final overall singing quality judgment score. We believe that this late fusion approach best resembles the process of how humans perceive and judge, and thus will lead to better results.

2.4 Experiments

For evaluating our hypothesis of a perception-driven singing evaluation framework, we conduct four experiments: a) Study of human perception for singing quality judgment b) Study of combinations of acoustic features to predict the overall score c) Prediction of Perceptual Parameters d) Strategies for scoring. In this section, we briefly discuss our dataset, and the four sets of experiments along with their results.[1]

[1]The code base for feature extraction and the dataset can be found here: https://github.com/chitralekha18/PESnQ_APSIPA2017.git

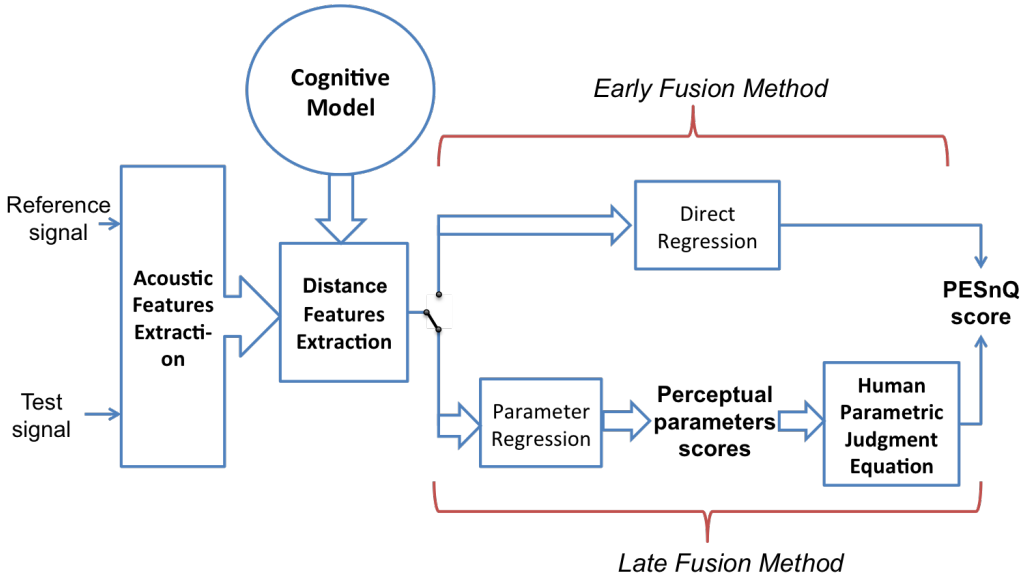


Figure 2.9: The diagram of PESnQ scoring with different approaches: early fusion, and late fusion.

2.4.1 Data

To test our methodology for singing evaluation, we chose two popular English songs - "I have a dream" by ABBA (∼2 min), and "Edelweiss" (∼1 min) from the movie "The Sound of Music". These songs have steady rhythm throughout the song, and are rich in long steady notes and vibrato. We needed monophonic sung recordings of these songs from singers with a range of singing ability - poor to professional. Duan et al. [DFL+13] recorded these two songs from 20 singers, but the range of singing ability in that dataset was limited to mediocre to good level, where the singers had some exposure to vocal training or were talented to sing in tune and rhythm. However, to cover the entire spectrum of singing ability, we needed samples from the two extremes - poor singers and professionally trained excellent singers. So we first obtained the dataset from Duan et al. and then recruited a few subjects to fill the gap at the two ends of the spectrum. We recruited two professionally trained singers and five students from NUS with no past experience in singing. These subjects were a mix of native and non-native English speakers, but all were proficient in speaking in English, similar to [DFL+13]. To be consistent with the previous dataset [DFL+13], we followed their procedure for collecting audio data from the new recruits. Subjects were asked to familiarize themselves with the two songs beforehand. Audio data was collected in a sound-proof audio recording studio at 16-bit and 16 kHz. A metronome was fed to the subject via headphone to serve as a guide for singing: "I have a dream" at 56 bpm, and "Edelweiss" at 32 bpm. These settings were same as that in [DFL+13]. Except for the metronome beats, no other accompaniment was provided. Lyrics for the songs were provided for the subject's reference while recording.

From the previous and the newly collected dataset, we selected 20 recordings for singing quality evaluation. Each was sung by a different singer with singing abilities ranging from poor to professional. Ten singers sang the song “I have a dream", and the other ten sang “Edelweiss". We obtained subjective evaluation ratings from music experts for these 20 recordings and ensured that this dataset is well representative of the singing skill spectrum (see Section 2.4.2). We also obtained objective evaluation scores for these recordings using our features and methods, and the known baseline methods.

2.4.2 Subjective Evaluation

We developed a website[2] to collect subjective ratings for this dataset. The task was to listen to the audio recordings and evaluate the singers' singing quality, compared to a professionally trained reference singer (also provided on the website). The reference singing of both the songs were from one professional singer, different from our test singing evaluation dataset of 20 singers. Five professional musicians were the human judges to complete this task. These judges have been trained in vocal and/or musical instruments in different genres of music such as jazz, contemporary, and Chinese orchestra, and all of them were stage performers. One of them has also been a music teacher for more than 2 years. The task could be done in multiple sittings, a few recordings each time. Their evaluations were saved in our database, that they could revisit later.

The website had two songs, each with 10 audio tracks, sung by different individuals (as described in Section 2.4.1). For every track, the corresponding lyrics were displayed on the screen. This is followed by a questionnaire, where the judges were asked to give an overall singing quality score out of 5 to each of these audio recordings compared to the reference singing of the song. The judges were also asked to separately rate each of the recordings based on pitch (*intonation accuracy*), rhythm (*rhythm consistency*), expression/vibrato (*appropriate vibrato*), voice quality (*timbre brightness*), articulation, relative volume, and pitch dynamic range on a likert scale of 1 to 5. Additionally, an optional question was asked to know if the music expert considers any other parameters that the singer could improve upon, apart from the ones already listed.

The average inter-judge (Pearson's) correlation of the overall singing quality question was 0.82, which shows a high agreement of singing quality assessment amongst the music experts. Table 2.3 shows the inter-judge correlation of all the questions that used a likert scale. Most of the questions showed correlation of higher than 0.60. Thus these parameters are judged by music experts coherently. However, the questions on pronunciation and volume showed lower inter-judge correlation. Since the lyrics were already provided to the singers, there was little room for mispronouncing words because of unfamiliar lyrics. The only way

[2] https://slions.smcnus.org/welcome.php

Table 2.3: Inter-judge correlation for the questionnaire questions.

Question	Inter-judge correlation
Overall singing quality	0.82
How would you rate the singer in terms of pitch accuracy?	0.81
How would you rate the singer in terms of rhythm accuracy?	0.75
How would you rate the singer in terms of vibrato/expression quality?	0.65
How would you rate the singer in terms of voice quality?	0.68
How would you rate the singer in terms of pronunciation quality?	0.53
How would you rate the singer in terms of relative volume?	0.46
How would you rate the singer in terms of pitch dynamic range?	0.67

Table 2.4: Number of singers with different levels of overall singing ability, categorized based on average human ratings.

Avg. score range	1.0 - 1.8	1.8 - 2.6	2.6 - 3.4	3.4 - 4.2	4.2 - 5.0
# of singers	5	3	8	2	2

mispronunciations could have happened was due to mother-tongue influence in non-native English. A possible reason for less agreement on pronunciation ratings is unclear definition of mispronunciation in singing, which leads to influence of other factors on this rating. An example of disagreement was when a singer, whose mother-tongue was English, but who had poor singing skills, was rated poorly for pronunciation by a couple of judges, while the other three judges rated the singer high for the same parameter. So in this case, poor singing seems to influence the perception of pronunciation. We believe that the reason for disagreement in case of the relative volume question is also because of lack of clear definition. As seen in Section 1.3.1 and 2.1, volume never showed up in the literature on subjective assessment of singing in non-trained singers [CLLY08, OBD+06], but volume was one of the key features in the objective evaluation literature [TL12, Tan99] because this measure is easy to compute objectively and pattern-match with a reference template, but difficult to rate subjectively. This explains the low agreement on volume parameter.

We computed the average of the overall singing quality score given to each of the 20 singers over the 5 human judges. We found that this data represents the complete singing skill spectrum. Table 2.4 shows the number of singers with different overall singing abilities categorized by average human ratings.

2.4.3 Pre-processing

As a pre-processing step, we first split every audio recording into shorter segments or *utterances* of approximately 20 sec duration. This is done by using DTW to align MFCC feature vectors of the test audio with that of the reference audio that is marked with segment boundaries. Rough segment boundaries for test audio file are obtained from this method, and then a quick manual check and correction of these segments is done, if needed. We need these short audio segments because alignment errors propagation is expected to be less in short duration segments compared to relatively longer segments. From here on, each of the features are computed for each of these segments. The subjective evaluation for a test audio recording is assumed to hold for every segment of that recording. We have 80 such segments for the song "I have a dream", and 40 segments for the song "Edelweiss", in total 120 test singing segments.

2.4.4 Study of Human Perception for Singing Quality Judgment

As discussed in Section 2.3.1, based on psychology studies, we hypothesize that humans follow a two-stage process to judge the overall singing quality, i.e. we first convert the perceived singing audio into a weighted representation of the identified perceptual parameters, and then map this representation to the overall singing quality judgment score.

To evaluate this theory, we first observe the effect of the individual perceptual parameters on the overall singing quality scores in our dataset, and then train a regression model to estimate the weights of the parametric equation that approximates the overall score (Eq.(2.11)). This will give us insight about the important perceptual parameters that affect the human judgment. We also verify if the trends observed are comparable to that reported in the literature. This is an important validation step because these human ratings would serve as the references for the rest of our experiments. Table 2.5 shows the matrix of Pearson's correlation between the individual perceptual parameters. The last column of Table 2.5 shows the Pearson's correlation between the average of the overall ratings by 5 judges and the individual perceptual parameter scores.

Results and Interpretation

Table 2.5 shows that intonation accuracy is the highest contributing factor to the overall performance rating. Chuan et al. [CLLY08] also report this same observation. Moreover, we find that appropriate vibrato also has a strong correlation with the overall score. Chuan et al. did not consider vibrato as an evaluation criterion in their experiments because it rarely

Table 2.5: Pearson's correlation between individual perceptual parameters human scores. (*vq*: voice quality, *pronun*: pronunciation, *pdr*: pitch dynamic range)

	intonation	**rhythm**	**vibrato**	**volume**	**vq**	**pronun**	**pdr**	**overall**
intonation	1	0.919	0.91	0.786	0.89	0.793	0.965	**0.961**
rhythm	-	1	0.863	0.847	0.86	0.758	0.947	0.923
vibrato	-	-	1	0.684	0.923	0.839	0.871	0.96
volume	-	-	-	1	0.716	0.677	0.856	0.731
vq	-	-	-	-	1	0.822	0.889	0.945
pronun.	-	-	-	-	-	1	0.783	0.848
pdr	-	-	-	-	-	-	1	0.945
overall	-	-	-	-	-	-	-	1

appeared in their dataset that consisted of only untrained singers. However, our dataset consisted of a mix of trained and amateur singers, and literature suggests that appropriate vibrato is an important cue to distinguish between them [NGH06a, WE97, OBD+06]. This is also reflected from the strong correlation between intonation and vibrato (0.910), meaning that good intonation is likely to have good vibrato. Thus high contribution of vibrato in determining singing quality is reasonable. The other parameters also show high correlation with the overall score, except the volume parameter. We believe that the reason for relatively weaker correlation for volume is lack of clear definition, as observed in Section 2.4.2.

Then we train a linear regression model in 10-fold cross validation using WEKA [HFH+09] to estimate the weights C_i of the Eq.(2.11) that predicts the overall singing quality score from the individual perceptual parameter scores. We check the correlation of these predictions with the subjective ground-truths. This experiment tests the possibility of defining singing quality judgment score in the form of a parametric equation. The linear regression model describes how humans relate the individual human perceptual parameters with the overall song-level scoring. The 10 fold cross validation resulted in a Pearson's correlation of 0.966 with the human overall judgment, and the linear equation that gives this prediction is:

$$\begin{aligned} score = 0.161 \times P + 0.189 \times R + 0.297 \times Vib \\ -0.27 \times Vol + 0.181 \times VQual + 0.115 \times Pronun \\ +0.311 \times PDR - 0.061 \end{aligned} \tag{2.12}$$

where $P, R, Vib, Vol, VQual, Pronun, PDR$ are the human scores for the perceptual parameters intonation accuracy, rhythm consistency, appropriate vibrato, volume, voice quality, pronunciation, and pitch dynamic range, respectively. This is the human parametric judgment equation for singing quality evaluation.

The weights of the perceptual parameters in this equation show the contribution of each of the parameters to the overall score. We see that intonation, rhythm, vibrato, voice quality, and pitch dynamic range are high contributors, while volume and pronunciation are low contributors. This trend is consistent with the trend observed in the last column of Table 2.5.

The strong correlation of the predicted overall score with the human ratings based on this linear parametric equation indicates the possibility that humans evaluate singing quality in a two stage manner, i.e. first evaluate the individual perceptual parameters and then express the overall judgment as a weighted combination of these perceptual parameters. Therefore to build an automatic system that emulates this two stage process of singing quality evaluation, we should first predict the perceptual parameters and then use this parametric judgment equation to predict the score.

2.4.5 Study of combinations of distance features to predict the overall singing quality score

Here we describe automatic systems built using combinations of the features from Section 2.2. Our automatic singing evaluation framework is the same as that of PESQ (Figure 2.1).

We then compare each of the corresponding reference and test audio segments in terms of intonation, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range related objective features. The methods to compute these features are described in Section 2.2. Table 2.1 summarizes the acoustic and the distance features extracted corresponding to the perceptual parameters.

To investigate the factors that influence the performance of machine-based singing quality evaluation, we use combinations of the various L2-norm based, PESQ based, and DTW based distance features to design two baseline and 9 test evaluation systems (Table 2.6). Baselines A and B are the systems purely consisting of features extracted from the singing evaluation literature. Baseline A consists of pitch distance feature [MBG+13, TL12, Tan99, Cha07] and Molina et al.'s pitch-based rhythm feature [MBG+13], while Baseline B has an additional volume distance feature [TL12, Tan99]. So Baselines A and B are the comparison benchmarks of this study. Also these systems would reveal the impact of the additional volume feature. Systems 1 and 2 are modified-baselines A and B respectively with the difference of the pitch-based rhythm feature [MBG+13] being replaced with the MFCC-based modified version (see Section 2.2.2). These systems will provide insight about the definition of the objective feature for rhythm consistency, i.e. if the MFCC-based rhythm feature is better than the pitch-based version. System 4 contains PESQ-based L6+L2 norm features along with DTW distance features but no L2-norm feature, while System 5 is the one with L2-norm features but without L6+L2 features. System 6 contains only the DTW distance features. Systems 4, 5, and 6 should show the impact of the PESQ-based perceptual features, compared to

Table 2.6: The distance features that describe the various singing evaluation systems.

System Name	Description	Feature List
Baseline A	Consists of pitch distance and pitch-based rhythm distance [MBG+13] features	pitch_dist, molina_rhythm_pitch_distance
Baseline B	Consists of Baseline A features along with volume-based distance features	Baseline A + volume_dist
System 1	Modified Baseline A - modification: pitch-based rhythm distance feature [MBG+13] changed to MFCC-based rhythm distance feature	pitch_dist, molina_rhythm_mfcc_distance
System 2	Modified Baseline B - modification: pitch-based rhythm distance feature [MBG+13] changed to MFCC-based rhythm distance feature	Baseline B + volume_dist
System 3	Consists of L2, L6+L2, and dist features, except pitch-based [MBG+13] and MFCC-based rhythm distance features	rhythm_L2, pitch_der_L2, pitch_med_L2, rhythm_L6_L2, pitch_der_L6_L2, pitch_med_L6_L2, timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 4	Consists of L6+L2, and dist features, except pitch-based [MBG+13] and MFCC-based rhythm distance features	rhythm_L6_L2, pitch_der_L6_L2, pitch_med_L6_L2, timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 5	Consists of L2, and dist features, except pitch-based [MBG+13] and MFCC-based rhythm distance features	rhythm_L2, pitch_der_L2, pitch_med_L2, timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 6	Consists of only dist features, except pitch-based [MBG+13] and MFCC-based rhythm distance features	timbral_dist, pitch_dist, pitch_der_dist, pitch_med_dist, vib_segment_dist, volume_dist, pitch_dynamic_dist
System 7	Consists of union set of features from System 2 and System 3	System 2 + System 3
System 8	Consists of union set of features from Baseline B and System 3	Baseline B + System 3
System 9	Consists of union set of features from Baseline B, System 2 and System 3	Baseline B + System 2 + System 3

the distance features commonly used in singing evaluation literature. System 3 consists of PESQ-based (L6+L2) features as well as all other DTW distance and L2-norm based features, except for the rhythm distance feature of [MBG+13] and its modified version. System 7 adds the MFCC-based modified rhythm distance feature to System 3, while System 8 adds the pitch-based rhythm feature [MBG+13] to System 3. System 9 adds both these rhythm distance features to System 3. Comparison of Systems 3, 6, 7, 8, and 9 will provide insight about the interaction between the distance features that they comprise of, in terms of their performance in predicting the overall singing quality rating.

Systems 3, 4, 6-9 consist of combinations of PESQ-based, L2-norm, and DTW distance-based features. Thus the score generated from these systems is termed as the PESnQ score.

We build a Linear Regression (LR) model and a Multi-Layer Perceptron (MLP) model with one hidden layer for each of these systems using Weka [HFH+09], in two modes: A) train and test on overall singing quality score averaged over the 5 judges in 10-fold cross validation, and B) Leave-one-judge-out, i.e. train on 4 judges in 10-fold cross validation, and test on 1 judge. The R-squared correlation values (computed in Weka) between the various system outputs and human ratings are shown in Table 2.7.

Results and Discussion

From the subjective evaluation, we wanted to see if the parameters pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range perceptually combine to predict the overall singing quality score. We trained and tested the two models in the leave-one-judge-out mode (mode B). We used the average subjective ratings for each of these parameters to predict the average subjective overall singing quality rating. Mode A, i.e. using the average score over all the judges, is not applicable in this case because of overlap between train and test data. The predictions showed the maximum average leave-one-judge-out correlation of 0.87 (Table 2.7). This is the maximum correlation achieved amongst human judges, thus it is also the upper bound of the achievable performance of machine-based singing quality evaluation. We asked an optional question to the human judges to find out if there are other perceptual features that are important to singing quality assessment. Most of the answers were associated with one of these seven parameters, e.g. "key changes in the middle of the song" is indicated by the pitch accuracy parameter, etc. But there were a few comments which were indeed not covered in those seven parameters, such as "inability to sustain long notes". Nonetheless, with the high correlation between parameter-based prediction of overall score and the actual overall score, we can safely consider that the current set of seven perceptual features are good predictors of the overall singing quality. So we designed objective methods to obtain automatic scores for each of these parameters for building an automatic singing quality evaluation system.

Training and testing the various singing evaluation systems (Table 2.7) on average overall score (Mode A) shows that System 8 performs the best with a correlation of 0.59 with the average human ratings, as compared to 0.30 of Baseline B. This shows that a combination of PESQ-based, L2-norm, and distance-based pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range related objective features, can predict the overall singing rating with an improvement of ∼96% over the current baseline systems (Baseline B) that use only a subset of these features. This result shows that our designed features that take key transposition and rhythm variations into account have a positive impact on the system performance.

Addition of volume feature in Baseline B shows only a slight improvement over Baseline A. System 1, which is the modified version of Baseline A, performs better than the baselines.

Table 2.7: Correlation between system output and human overall singing quality ratings.

System configs	Average Overall Score		Leave out judge 1		Leave out judge 2		Leave out judge 3		Leave out judge 4		Leave out judge 5		Avg. leave-one-judge-out	
	LR	MLP	LR	MLP	LR	MLP	LR	MLP	LR	MLP	LR	MLP	LR	MLP
Human judge	-	-	0.93	0.96	0.87	0.87	0.78	0.75	0.95	0.88	0.83	0.83	0.87	0.86
Baseline A	0.30	0.26	0.36	0.39	0.30	0.34	0.45	0.36	0.31	0.36	0.35	0.37	0.35	0.36
Baseline B	**0.30**	0.29	0.36	0.40	0.30	0.40	0.45	0.36	0.31	0.39	0.35	0.36	0.35	**0.38**
System 1	**0.36**	0.27	0.38	0.50	0.35	0.55	0.43	0.40	0.37	0.50	0.40	0.36	0.39	**0.46**
System 2	0.34	0.30	0.38	0.36	0.35	0.39	0.43	0.30	0.37	0.36	0.34	0.32	0.37	0.35
System 3	0.48	0.55	0.61	0.66	0.57	0.56	0.54	0.54	0.61	0.66	0.46	0.51	0.56	0.59
System 4	0.50	0.55	0.61	0.64	0.57	0.56	0.54	0.53	0.60	0.65	0.48	0.48	0.56	0.57
System 5	0.49	0.55	0.61	0.65	0.57	0.56	0.54	0.53	0.61	0.66	0.49	0.50	0.56	0.58
System 6	0.53	0.47	0.58	0.62	0.56	0.68	0.52	0.52	0.58	0.67	0.41	0.43	0.53	0.58
System 7	0.48	0.53	0.61	0.66	0.57	0.67	0.51	0.53	0.61	0.71	0.46	0.54	0.55	0.62
System 8	0.42	**0.59**	0.61	0.68	0.58	0.70	0.54	0.57	0.61	0.73	0.45	0.54	0.56	0.64
System 9	0.43	0.56	0.61	0.69	0.59	0.72	0.53	0.57	0.61	0.74	0.47	0.56	0.56	**0.66**

This shows that our MFCC-based modified rhythm distance feature performs better than the baseline pitch-based rhythm distance feature [MBG+13]. This supports our theory that human errors in producing the correct pitch will degrade the baseline pitch-based rhythm distance feature. In our dataset, the subjects were proficient in English and had rehearsed the songs before recording. Thus, they made few mistakes in the lyrics while singing. However, they were restricted by their singing ability. Thus, the MFCC-based modified version of the baseline rhythm distance feature, which is robust to pitch errors, is more suitable in this case.

An interesting finding is that System 4 shows improvement over System 5. This is also evident when System 4 (PESQ-based and distance features) is compared to System 6 (DTW distance features alone). PESQ-based L6+L2 features provide an improvement of 3.7% over only distance features. Although earlier works relied on distance metric alone, our results show that adding features based on cognitive modeling theory improves machine correlation with human perceptual judgment.

The leave-one-judge-out experiments (Mode B) show that the output of our system trained on 4 judges correlates well with the 5th judge consistently. Thus, our system is able to generalize when trained on 4 judges. System 9 shows the best average correlation of 0.66. This is closer to the upper-bound of achievable correlation compared to the baseline system that shows correlation of 0.38. We also notice that the performance of some of our systems is comparable to that of the human judges. For example, System 9 shows correlation of 0.74 for leave-out-judge4 experiment, which is comparable to human judges' leave-out-judge3

correlation values. So, our system is close to reproducing judgments from a human music expert.

2.4.6 Prediction of Perceptual Parameters

In this experiment, we would like to observe how well we can predict the individual perceptual parameter scores using the best system that we designed in the previous experiment (Section 2.4.5). This experiment is motivated by our theory that the human evaluation process can be best emulated by predicting the perceptual parameters first and then combining these parameter scores. Providing scores for individual perceptual parameters also means giving more meaningful feedback to a learner, i.e. providing a breakdown of an overall score in terms of musical parameters understood by humans.

In Section 2.4.5, we tested various combinations of cognitive as well as distance-based acoustic-perceptual features to predict the overall singing quality judgment. We found that the *System 8* that consisted of a combination of PESQ-like, L2-norm, and distance-based pitch, rhythm, vibrato, voice quality, pronunciation, volume, and pitch dynamic range features performed the best in predicting the overall singing quality rating.

Now we use the features of System 8 to train (10-fold cross validation) linear regression models separately for each of the individual perceptual parameters. Prediction is done at *utterance-level* and *song-level*. Song-level scores are computed in two ways: a) by computing the *average* over all the utterance scores of a song, and b) by computing the *median* of all the utterance scores of a song. The median-based song-level scores avoid the impact of outlier utterance-scores. Figure 2.10 shows the correlation of the predictions of the individual parameters with the average human scores.

Results and Interpretation

Figure 2.10 shows that the song-level predictions correlate with the human scores more than the utterance-level predictions. Intuitively this trend can be understood from the idea that more data results in better statistics, so the machine gives better predictions with more data, i.e. at song-level. Short excerpts of the song may not represent the overall impression of the song, thus resulting in noisy scores. Thus a song-level automatic score for the song should correlate more with the song-level human judgment.

We observe that all the individual perceptual parameters are predicted well at the song-level with correlation of more than 0.7. Moreover, the average and the median score correlation at song-level evaluation are comparable. This means that the utterance scores did not have many outliers, hence these song-level scores were not affected drastically.

These results show that it is possible to predict the individual perceptual parameter scores reliably and hence provide a more meaningful feedback to a learner in terms of a breakdown

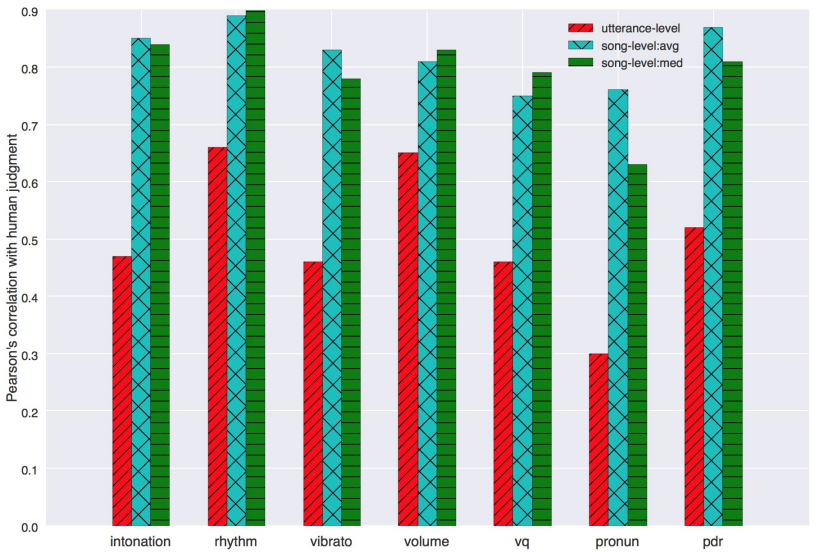


Figure 2.10: Performance of the best set of distance features from Section 2.4.5 in predicting the individual perceptual parameters when trained separately for each of them, at utterance-level, and song-level (average and median). (*vq*: voice quality, *pronun*: pronunciation, *pdr*: pitch dynamic range)

of an overall judgment score into perceptually and musically relevant scores. Moreover, having reliable objective measures for evaluation also helps in avoiding subjective biases that can creep into human judgments. For example, if a singer sings with good pronunciation, but has bad intonation, we observe that some judges tend to poorly rate pronunciation in such cases, because humans tend to get influenced by the overall impression. Objective evaluation helps in avoiding such biases.

2.4.7 Strategies for Scoring

In the experiment in Section 2.4.4, we found that the overall singing quality score can be modeled with a linear parametric equation, or Eq. 2.12. We showed that the human scores for the individual perceptual parameters collectively predict the overall scores via this model. This motivates us to apply this model on the machine scoring of Section 2.4.6 and obtain the overall singing quality score. This approach of late fusion for automatic overall scoring emulates the two-stage process of singing quality judgment by humans, as discussed in the previous sections. In this experiment, we compare this strategy with the early fusion strategy of scoring.

The distance feature set consists of the same features as in Section 2.4.6, i.e. a combination of PESQ-like and distance-based features. We compare the overall song-level singing quality scoring by two methods:

Table 2.8: Comparison of Pearson's correlation of the human overall judgment with the predicted overall PESnQ score by early and late fusion methods.

	Early fusion	**Late fusion**
Song-level: average	0.725	0.904
Song-level: median	0.747	0.855

Table 2.9: Comparison of Pearson's correlation of predicting the 5th judge in a leave-one-judge-out experiment by early and late fusion methods.

	Leave out judge 1	**Leave out judge 2**	**Leave out judge 3**	**Leave out judge 4**	**Leave out judge 5**	**Avg.**
Human judges	0.929	0.872	0.78	0.948	0.831	0.872
Early fusion	0.745	0.703	0.696	0.73	0.622	0.699
Late fusion	0.785	0.722	0.726	0.78	0.683	**0.739**

1. early fusion, i.e. by using the distance features directly (Section 2.4.5), and
2. late fusion, i.e. by using the individual parameter predictions from Section 2.4.6 and applying Eq. 2.12 for overall scoring.

Results and Interpretation

Table 2.8 shows the Pearson's correlation of the predictions obtained from two methods by 10-fold cross validation, where the ground-truth was the overall singing quality score averaged over all the 5 judges. All the correlation values are statistically significant with $p < 0.001$.

Table 2.9 shows a leave-one-judge-out experiment, where the two methods predict the singing quality score of the 5th judge i.e. train on 4 judges in 10-fold cross validation, and test on 1 judge.

From Table 2.8, we find that the overall singing quality scoring from the late fusion method is superior to that from the early fusion method. We obtain a song-level averaged score prediction correlation of 0.904 with that of human scores by late fusion method, compared to 0.725 by early fusion method. This shows that the late fusion method emulates human perception better than the early fusion method. This means that our hypothesis that the late fusion approach most resembles the process of how humans perceive and judge singing quality is valid.

Average and median song-level scores show comparable performance like in Section 2.4.6 indicating that there may not be many outliers in the utterance-level scores. Figure 2.11

demonstrates the correlation of the machine song-level scores from early and late fusion with the human scores. We observe that the correlation in case of late fusion is higher and there is no gross misjudgment by the machine. This is an encouraging result as it highlights the significance of modeling human perception for building a machine that assesses singing quality.

We also verify that the correlation between the machine scores for the individual parameters and the overall singing quality by late fusion method (Table 2.10) follow a trend similar to that observed in human scores (Table 2.5). Intonation accuracy shows the highest correlation with the overall singing quality, while volume shows low correlation, same as in human scores. This similarity in trends implies that our designed distance features and perceptually-driven framework for evaluation generates perceptually reliable scores both for overall singing quality as well as for the related musical parameters understood by humans.

In the leave-one-judge-out experiment, we see that the maximum correlation achieved by the human judges to predict all the song ratings of an unseen judge is 0.87, thus indicating an upper bound of the achievable performance of any system that tries to emulate a human judge. The leave-one-judge-out experiment (Table 2.9) is different from the overall singing quality scoring experiment (Table 2.8) by the fact that leave-one-judge-out predicts the scores given by a judge for all the songs when trained on four judges, while the other experiment predicts the score of any unseen singing when trained on all the judges. We find that late fusion is better at predicting the unseen judge than the early fusion method, thus showing that the perception-driven framework is able to generalize better. We also notice that the performance of our framework is sometimes comparable to that of the human judges. For example, in the leave-out-judge3 experiment, machine performance is comparable to the human judges' performance in predicting the unseen judge. These results show that our framework of evaluation closely emulates a human music expert.

2.5 Summary

In this chapter, we presented a technical framework for automatic perceptual evaluation of singing quality (PESnQ). From the subjective judgments of music experts, we found that intonation, rhythm, voice quality, vibrato, pronunciation, volume, and pitch dynamic range are the perceptual parameters that can reliably predict the overall singing quality. We designed objective distance features to automatically evaluate each of these perceptual parameters, while overcoming the challenges of the well-known baseline features. Our pitch evaluation features avoided penalizing for overall key transposition, and our rhythm evaluation features avoided penalizing for uniform rhythm variation, even when the pitch is off-tune. Also, we designed distance features according to the cognitive modeling theory for audio perception in speech, used in the PESQ standard. We found that this theory could be applicable for singing evaluation also. Based on these features, we compared various systems

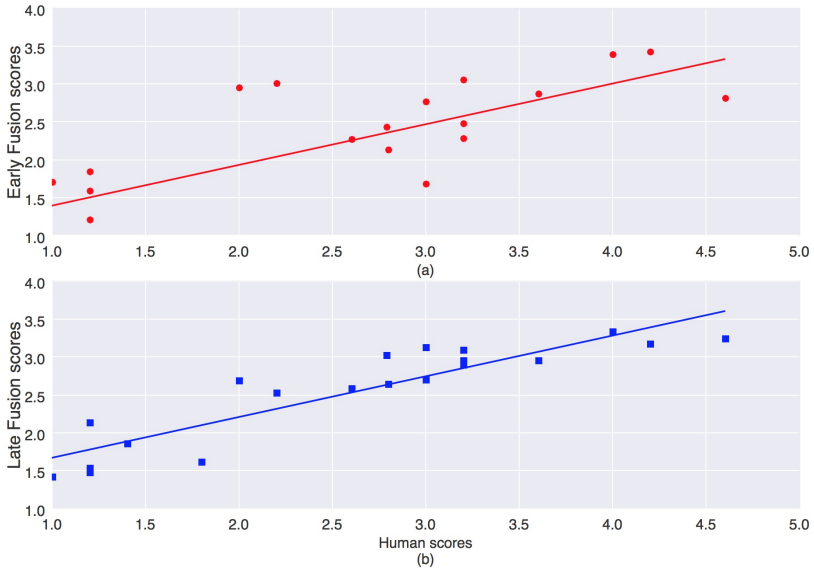


Figure 2.11: (a) Early Fusion versus (b) Late Fusion to obtain the PESnQ score. Pearson's correlation of early fusion method is 0.725 and that of late fusion method is 0.904, both with statistical significance of $p < 0.001$.

Table 2.10: Pearson's correlation between individual perceptual parameter score predictions and overall singing quality (PESnQ) scoring by late fusion method. (*vq*: voice quality, *pronun*: pronunciation, *pdr*: pitch dynamic range)

	P	**R**	**Vib**	**Vol**	**VQual**	**Pronun**	**PDR**	**Overall**
P	1	0.84	0.801	0.494	0.707	0.647	0.855	**0.952**
R	-	1	0.811	0.704	0.824	0.613	0.83	0.912
Vib	-	-	1	0.634	0.831	0.555	0.627	0.856
Vol	-	-	-	1	0.788	0.639	0.574	0.537
VQual	-	-	-	-	1	0.702	0.7	0.818
Pronun	-	-	-	-	-	1	0.653	0.672
PDR	-	-	-	-	-	-	1	0.896
Overall	-	-	-	-	-	-	-	1

trained to predict the overall singing quality. The predicted PESnQ score from a system having a combination of PESQ-based, L2-norm, and DTW based distance features (System 8), showed a correlation of 0.59 with human ratings. This is a ∼96% improvement over the system built on baseline features.

We obtained a human judgment parametric equation that predicts the overall singing quality human score from the individual perceptual parameter human scores. We showed that the best set of distance features can predict the individual perceptual parameter scores reliably. So then we applied the human judgment parametric equation on the predicted perceptual parameter scores to obtain the prediction of the overall singing quality score. We show that the singing quality score obtained from the late-fusion method has a higher correlation of 0.904 with human judgment than that of 0.725 from the early fusion method. Thus, we provided systematic analysis to show that a framework that emulates the human perceptual process of singing quality assessment results in better scoring.

Chapter 3

Reference-Independent Singing Quality Evaluation

Studies have shown that music experts can evaluate singing quality with high consensus when the melody or the song is unknown to them [NGH06b, NGH06a]. This suggests that there are inherent properties of singing quality that are independent of a reference singer or melody, which help music-experts judge singing quality without a reference. In this chapter, we explore these properties and propose methods to automatically evaluate singing quality without depending on an ideal reference singing.

This chapter is organized as follows. In Section 3.1, we discuss the motivation, and background of this topic, and briefly present our idea and contributions. In Section 3.2, we discuss the various music-motivated absolute measures, in Section 3.3, we discuss our idea and approach for inter-singer relative measure computation, and look at the various relative features. The computational framework, including data preparation, experimental setup, experiments, analysis and validation, is discussed in Sections 3.5 and 3.6.

3.1 Background studies and our contributions

Computer-assisted singing learning tools have been reported to be useful for singing lessons [HSD06, WHR89, HW89]. Recently, karaoke singing apps such as Smule Sing![Smu08], Starmaker[Sta10], and online platforms such as SoundCloud, and Youtube have provided a platform for people to showcase their singing talent, and a convenient way for amateur singers to practice and learn singing. They also provide an online competitive platform for singers to connect with other singers all over the world, and improve their singing skills. Automatic singing evaluation systems on such platforms typically compare a sample singing vocal with a standard reference such as a professional singing vocal [Cha07, Lal06, LLCW14] or the song melody notes [TL12, Tan99, MBG+13] to obtain an evaluation score. For example, in the previous chapter, we computed a singing quality score called Perceptual Evaluation of Singing Quality (PESnQ) that measured the similarity between a test singing and a reference singing in terms of pitch, rhythm, vibrato, etc. However, such methods are constrained either by the need for a professional grade singer, or the availability of digital sheet music for every song. The aesthetic perception of singing quality is very subjective and varies between evaluators,

so that even experts often disagree on the perfection of a certain performance [BES+17]. The choice of an *ideal* or *gold-standard* reference singer brings in a bias of subjective choice. Therefore, a reference-independent method of singing quality evaluation is desirable.

Amateur and promising singers also upload cover versions of their favorite songs on these online platforms, that are listened and liked by millions across the globe. However discovering talented singers from such huge datasets is a challenging task [NDAL12]. Moreover, often times the cover songs don't follow the original music scores, but rather demonstrate the creativity and singing style of individual singers. In such cases, reference singing or musical score based evaluation method is not an ideal choice.

There have been a few studies to objectively evaluate singing quality without a standard reference. Nakano et al. [NGH06a] designed a singing skill evaluation scheme based on pitch interval accuracy and vibrato, which are regarded as features that function independently from the individual characteristics of singer or melody. They used pitch interval accuracy to measure the consistency of the offset of the pitch values within a musical semitone grid. For computing the pitch interval accuracy, the fundamental frequency trajectory is fitted to a semitone (100 cents) width grid (corresponding to equal temperament in the Western Music Tradition), i.e. all the pitch values are wrapped on to a semitone. If the pitch values have a constant offset from this semitone grid throughout the song sequence, then the singing was considered to be of good quality. Although pitch interval accuracy is a fair indicator, it ignores other properties of a song. For example, if a singer sings only a single note throughout the song, pitch interval accuracy will classify it as good singing. Therefore it fails fundamentally, and overlooks the occurrence of several notes in a song and different notes being sustained for different durations.

Other studies primarily measure the quality of pitch histogram. A pitch histogram, wrapped on to a grid of 12 semitones or 1200 cents, preserves the information about the number of frequently hit notes in a song. And, sharp peaks in the pitch histogram capture note sustenance that indicates consistency of hitting the same notes. To measure the sharpness of the peaks, Nichols et al. [NDAL12] and Bohm et al. [BES+17] computed kurtosis and skew of the pitch histogram. They are overall statistical indicators that don't care much about the actual shape of the histogram which could be informative about the singing quality.

Humans are known to be better at relative judgments than absolute judgments, i.e. choosing the best and the worst among a small set of singers [LLI+13, ÇLT17], rather than giving an absolute rating. This leads us to the idea of generating a leaderboard of singers, where the singers are rank-ordered according to their singing quality relative to each other. With the immense amount of online uploads on singing platforms, we can now leverage on the comparative statistics between singers as well as music theory to derive such a leaderboard of singers. In this work, we propose a novel approach of discovering good singers from a large number of singers by assessing the similarities or the relative distances between the

singers. Based on the concept of *truth-finding*, we believe that good singers sing alike, but bad singers sing very differently to each other. If all singers sing the same song, the good singers would share many characteristics such as the frequently hit notes, the sequence of notes, and the overall consistency in the rhythm of the song. However different bad singers will deviate from the intended song in different amounts and ways. For example, a singer may be out-of-tune at certain notes, while another may be out-of-tune at some other notes. This idea exploits the inter-singer statistics of a song, instead of relying on a fixed reference sample. We propose a framework to combine these inter-singer relative distance measures with musically-motivated pitch histogram-based measures to provide a comprehensive singing quality assessment without relying on a standard reference. We assess the performance of our algorithm by comparing against human judgments.

In the context of singing pedagogy, it is important to provide detailed feedback to a learner about their performance with respect to the individual underlying perceptual parameters. Although humans are known to provide consistent overall judgments, they are not good at objectively judging the quality of individual underlying parameters. In this work, we show that the proposed singing quality evaluation scheme outperforms human judges in this regard.

In this chapter, we focus on only three of the seven perceptual parameters from Chapter 2: intonation accuracy, rhythm consistency, and voice quality. From the subjective judgments in Chapter 2, we find these parameters to be of high importance in singing quality evaluation. In the rest of this chapter, we refer to intonation accuracy as pitch, rhythm consistency as rhythm, and voice quality as timbre for ease of reference.

3.1.1 Our contributions

- We design a self-organizing *truth-finding* based method to rank-order large number of singing renditions based on singing quality without relying on a reference singer
- We propose and analyze various inter-singer relative distance measures and musically-motivated pitch histogram-based measures to characterize the inherent properties of singing quality
- We provide evidence that indicates that machines can provide a more unbiased assessment of the underlying parameters of singing quality compared to humans

3.2 Musically-Motivated Measures

In a subjective assessment study conducted by Nakano et al. [NGH06b], it was found that human judges could evaluate and rank singers with high consistency when they sang songs

that are unknown to the judges. This finding suggests that singing quality judgment depends more on common, objective features rather than reflecting subjective preference. This encourages us to explore methods to quantify singing quality in a reference-independent way.

From various subjective assessment studies, it has been found that the most important properties for singing quality evaluation are pitch and rhythm [CLLY08, Wel94]. Pitch is characterized by the fundamental frequency $F0$ and its movements between high and low values. Musical notes are the musical symbols that indicate the location and duration of pitch, i.e. the timing information or the rhythm of singing. In karaoke singing, visual cues to the lyric lines to be sung are provided that helps the singer to have more control over the rhythm of the song. Therefore, in the context of karaoke singing, rhythm is not expected to be a major contributor to singing quality assessment. Pitch, however, can be perceived and computed. Therefore, we will focus on the characterization of singing pitch in this section.

Pitch histograms are global statistical representations of the pitch content of a musical piece [TEC03]. They represent the distribution of pitch values in a sung rendition. A pitch histogram is computed as the count of the pitch values folded on to the 12 semitones in an octave. The melody of a song typically consists of a set of dominant musical notes (or pitch values). These are the notes that are hit frequently in the song and sometimes are sustained for long duration. These dominant notes of the song are a subset of the 12 semitones present in an octave. The other semitones may also be sung during the transitions between the dominant notes, but are comparatively less frequent and not sustained for long durations. Thus, in the pitch histogram of a good singing vocal of a song, these dominant notes should appear as the peaks, while the transition semitones appear in the valley regions.

To provide a fine representation of the pitch histogram, we divided each semitone into 10 bins, each bin representing 10 cents. All pitch values in this study are calculated in the unit of cents (one semitone being 100 cents on equi-tempered octave). Figure 3.1 shows the pitch histogram of MIDI (Musical Instrument Digital Interface) signal (a), a good singing vocal (b), and a poor singing vocal (c), all performing the same song, *I have a dream* by ABBA. The histogram is normalized to 1. The MIDI version contains the notes of the original composition, therefore represents the ideal pitch histogram of the song. It is apparent that the good singer histogram is close to the MIDI histogram. They have four sharp peaks showing that those pitch values are frequently and consistently hit, more than the rest of the pitch values. Since generally a song consists of only a set of dominant notes, the sharp, narrow, well-defined spikes of the pitch histogram of the good singer indicate that the notes of the song are being hit repeatedly and consistently. On the other hand, the poor singer has a dispersed distribution of pitch values, that reflect that the singer is unable to hit the dominant notes of the song consistently.

Previously, Nichols et al. [NDAL12] computed the tuning frequency to induce a grid of "correct" pitch frequencies based on an equi-tempered chromatic (12 semitone) scale, and

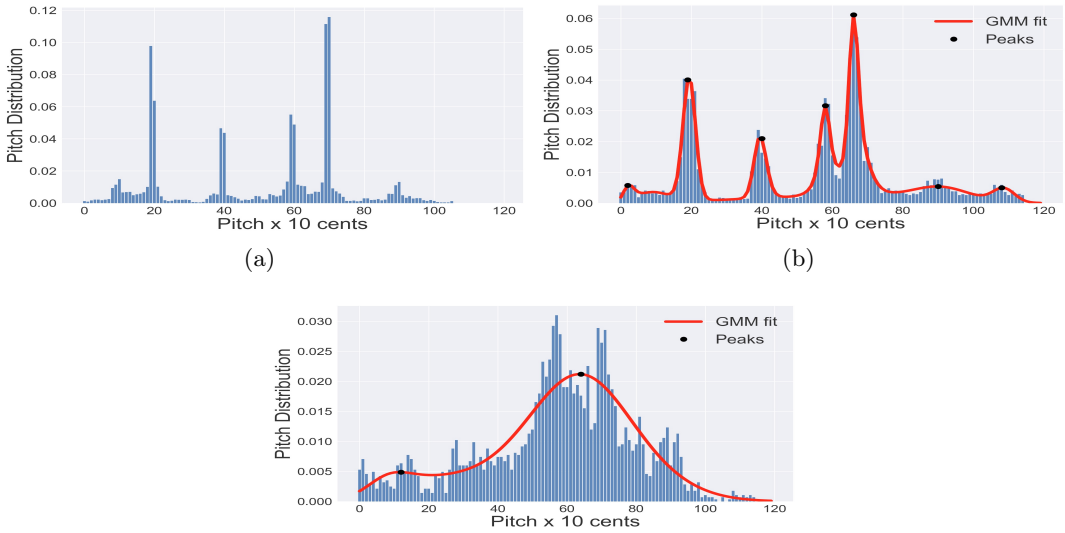


Figure 3.1: Normalized Pitch Histogram for (a) MIDI (b) good singing (c) poor singing of the song "I have a dream" by ABBA. GMM-fit (red line) and detected peaks (black dots) on the normalized pitch histogram for (b) good and (c) poor singing (the y-axis scales are different for better visibility).

then computed the histogram of the differences of the pitch values from the nearest correct frequencies. This resulted in a histogram of values within one semitone. Nakano et al. [NGH06a] used a filterbank method to obtain the correct frequencies grid, but then computed the one semitone histogram in the same way as in [NDAL12]. The statistical measures kurtosis and skew were computed to measure the sharpness of this histogram. However, determining the tuning frequency is a challenging task, that has not been solved yet [KW17, DBS11]. Therefore, in this work, we characterize the musical properties of singing quality with the 12 semitones pitch histogram. We believe that the shape of this histogram, for example, the number of peaks, the height and spread of the peaks, intervals between the peaks etc. contain vital information about the goodness of the sung melody. Although we cannot directly determine the correctness of the notes being sung when the notes of the song are not available, we can measure the consistency of the pitch values being hit, which is an indicator of the singing quality, as will be discussed in the following sub-sections.

In this work, we formulate several statistical measures for singing quality evaluation when the song or melody is unknown. In this section, we systematically discuss the group of musically-motivated pitch histogram-based measures. Moreover, we analyze and understand the musical perspective that each of these measures provides.

3.2.1 From the perspective of overall pitch distribution

This is a group of global statistical measures where we consider the overall distribution of pitch values that the singer sings on a 12 semitone grid. They essentially measure the deviation of the pitch distribution from a normal distribution. As seen in Figure 3.1, the pitch histogram of good singers show multiple sharp peaks, while that of poor singers show a dispersed distribution of pitch values. Therefore, we hypothesize that the histogram of a poor singer will be closer to a normal distribution, than that of a good singer.

Kurtosis

Kurtosis is a statistical measure of whether the data is heavy-tailed or light-tailed relative to a normal distribution. It has been used [NDAL12] to quantify the quality of singing pitch without a reference. Kurtosis is the fourth standardized moment, defined as

$$Kurt = E\left[\left(\frac{\vec{x} - \mu}{\sigma}\right)^4\right] \tag{3.1}$$

where $\vec{x}$ is the data vector, which in our case is the pitch values over time, μ is the mean and σ is the standard deviation of $\vec{x}$. A good singer's pitch histogram is expected to have several sharp spikes, as in Figure 3.1, and thus away from a normal distribution. So a good singer would have a higher kurtosis value than a poor singer.

Skew

Skew is another measure used in the literature for singing quality assessment [NDAL12]. It is a measure of the asymmetry of a distribution with respect to the mean, defined as

$$Skew = E\left[\left(\frac{\vec{x} - \mu}{\sigma}\right)^3\right] \tag{3.2}$$

where $\vec{x}$ is the data vector, μ is the mean and σ is the standard deviation of $\vec{x}$.

The pitch histogram of a good singer has peaks around the notes of the song, whereas that of a poor singer is expected to be more dispersed and spread out evenly. As the pitch histogram of a poor singer is expected to be closer to a normal distribution (see Fig. 3.1), or more symmetric. The skewness is a measure of the asymmetry, that can differentiate good singers from poor ones.

3.2.2 From the perspective of pitch concentration

The previous group of measures considered the overall distribution of the pitch values with respect to a normal distribution. However they do not provide any insight on the accuracy of the musical notes being hit. Next we would like to quantify the precision with which the notes are being hit.

To do this, we would essentially want to measure the concentration of the pitch values in the pitch histogram. Multiple sharp peaks in the histogram indicate precision in hitting the notes. Moreover, the intervals between these peaks contain information about the relative location of these notes in the song indicating the musical scale in which the song was sung. We consider the shape of the histogram in detail, i.e. the shape (spread and height) of the peaks, and the number of peaks in the histogram. These details provide more insights about the quality of singing, for example the number of prominent peaks in the histogram indicates the number of dominant notes sung, the sharper the peaks, the more precisely the notes are sung.

Gaussian mixture model-fit (GMM-fit)

To capture the fine details of the histogram, we fit a mixture of Gaussian distributions to model the pitch histogram. A GMM should be able to fit a histogram with several dominant peaks, as well as a dispersed histogram, thus providing a less noisy approximate representation of the histogram. Figure 3.1(b) and (c) show the GMM-fit for the good and the poor singer respectively.

The idea is to design a measure that characterizes the shape of this GMM-fit of the pitch histogram, that will capture the inherent discerning properties of the pitch histogram. To characterize the peaks in the histogram, we detect the local maximas in the GMM-fit [Bil]. Figure 3.1(b) and (c) show the detected local maximas.

We characterize singing quality on the basis of the detected peaks in the two following ways.

Firstly, we measure the spread around the peaks, that we call the Peak Bandwidth measure. It indicates the consistency of hitting the same notes. The smaller the spread, the higher the consistency, and therefore better the singer.

We define the peak bandwidth (BW) as the one that is 3dB down, or half power, from the peak. The half of the extent between the right and the left edges is termed as the peak standard deviation w. Therefore a measure of poor singing quality is directly proportional to the average w^2 of all the peaks, i.e. the larger the spread, the poorer is the singing quality. Moreover, since a pop song is expected to have more than one or two significant peaks in the pitch histogram, we would like to additionally penalize if there are only a small number of peaks, by dividing by the number of peaks N. Therefore, the peak-BW measure averaged

over the number of peaks becomes inversely proportional to N^2, defined as:

$$PeakBW = \frac{1}{N^2}\sum_{i=1}^{N} w_i^2 \tag{3.3}$$

where w_i^2 is the variance of the i^{th} detected peak.

Secondly, we measure the percentage of pitch values around the peaks, that we call Peak Concentration measure. This indicates the amount of time a singer spends on singing the intended notes. If this percentage is high, it means that most of the pitch values are concentrated around the peaks, indicating that the singer hits the same notes consistently. This measure takes the height of the peaks into consideration, which is also an indicator of the duration of the sustained long notes of the song. We define peak-concentration measure as

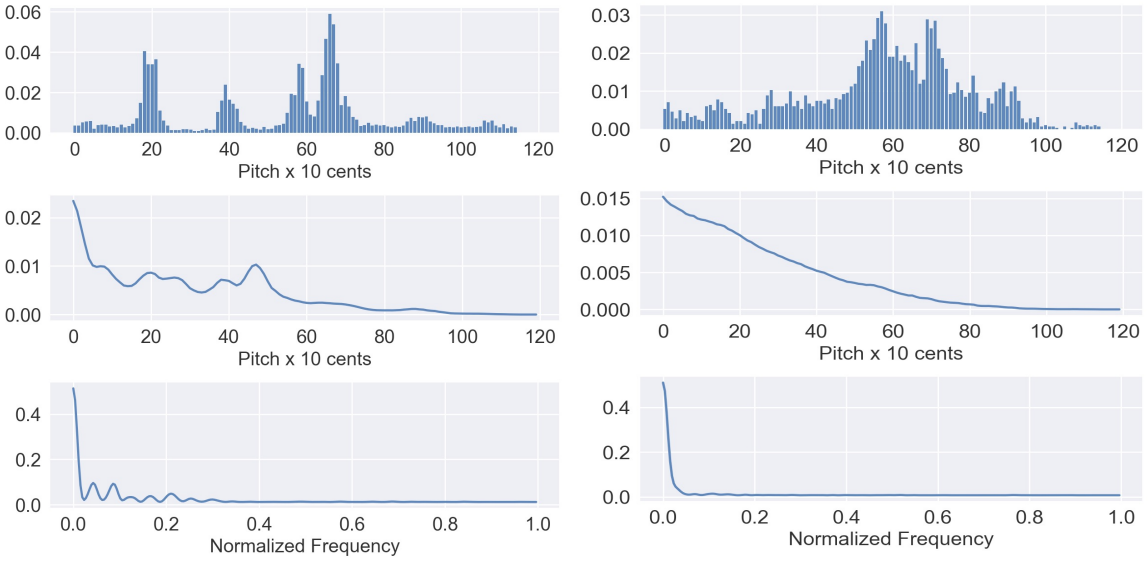
$$PeakConc = \frac{\sum_{j=1}^{N}\sum_{i=bin_j-\Delta}^{bin_j+\Delta} A_i}{\sum_{k=1}^{M} A_k} \tag{3.4}$$

where N is the number of peaks, bin_j is the bin number of the j^{th} peak, A_i is the histogram value of the i^{th} bin, and M is the total number of bins, i.e. 120 here. Human perception to is sensitive to pitch changes, but what is the small perceptible change has been debated upon. Scientists agree that normal adults are able to recognize pitch differences of as small as 25 cents reliably [PH03]. Thus in equation 3.4, Δ is the number of bins on either sides of the peak to be considered for measuring peak concentration. It represents the allowable range of pitch change without being perceived as out-of-tune. We have empirically considered the Δ values of ±5 and ±2 bins, i.e. ±50 cents and ±20 cents respectively, which are a total of 110 cents and 50 cents, respectively, around a peak including the center bin. We term these measures as $PeakConc_{110}$ and $PeakConc_{50}$ respectively.

Autocorrelation

We have 12 semitones to cover the range, singers are supposed to sing at the semitones. The minimum interval is one semitone, and the intervals between the musical notes should be one or multiples of a semitone, that can be observed if we perform autocorrelation on the pitch histogram. If a good singer is hitting the correct notes all the time, we expect to see sharp peaks at semitone and multiples of semitones in the Fourier transform of the autocorrelation of the pitch histograms. This is evident from Figure 3.2, where the magnitude spectrum of the autocorrelation of a good singing sample pitch histogram has energy in the higher frequencies representing the interval pattern of the strengthened peaks in the pitch histogram, while that of the poor singing sample only has the zero frequency component.

We compute the autocorrelation energy ratio measure or *Autocorr* as the ratio of the energy in the higher frequencies to the total energy in the Fourier transform of the autocorrelation



(a) (b)

Figure 3.2: The normalized pitch histogram (top), autocorrelation of the histogram (middle), and the magnitude spectrum of autocorrelation (bottom) for (a) good singing (b) poor singing (the y-axis scales are different for better visibility).

of the pitch histogram,

$$Autocorr = \frac{\sum_{f=4Hz} |Y(f)|^2}{\sum_{f=0Hz} |Y(f)|^2} \tag{3.5}$$

where,

$$Y(f) = F\left(\sum_{n=1}^{120} (y(n)y^*(n-l) \right) \tag{3.6}$$

i.e. the Fourier transform of the autocorrelation of the histogram $y(n)$ where n is the bin number, and total number of bins is 120, and l is the lag. The lower cut-off frequency of 4 Hz in the numerator of equation 3.6 corresponds to the assumption that at least 4 peaks or dominant notes are expected in a good singing rendition.

3.2.3 Clustering based on musical notes

A song typically consists of a set of dominant musical notes. This means that only a subset of the semitones of an octave will be consistently hit while singing a song. Although the melody of the song is unknown, we can imagine that the pitch values, when sung correctly, will be clustered into these dominant notes of the song. Therefore, the dominant set of notes present in any given song serves as a natural reference for evaluation.

k-Means Clustering

The density of pitch values across the histogram bins is an indicator of how well the pitch values are clustered together. Tightly grouped clusters indicate that most of the pitch values are close to the cluster centers which means that the same notes are hit consistently. Keeping this idea in mind, we apply k-Means clustering to the pitch values such that 12 clusters are formed. We chose k=12 for the 12 semitones in an octave. k-Means clustering algorithm optimizes the cluster centroids and boundaries by minimizing the sample distances within the clusters, while maximizing the distances between the clusters [Llo82, For65].

If a centroid is located close to a peak in the histogram, it implies that a large number of samples (or pitch values) have a small distance from the centroid. Moreover, when two centroids are closely spaced, the average distance of the samples from the centroid in each of those clusters will be less. We can see that the centroids around the highest peaks of the good singer's histogram are closely spaced, implying smaller sample distances. Therefore, whether the pitch values are tightly or loosely clustered can be represented by the average distance of each pitch value to its corresponding cluster centroid. So this distance is inversely proportional to the singing quality, i.e. smaller the distance, better the singing quality. We define the average cluster distance as

$$kMeans = \frac{1}{L}\sum_{i=1}^{k} d_i^2 \tag{3.7}$$

where k is the number of clusters (=12), L is the total number of pitch values (or frames with valid pitch values), and d_i is the total distance of the pitch values from the centroid in i^{th} cluster, defined as

$$d_i^2 = \sum_{j=1}^{L_i} \left(p_{ij} - c_i\right)^2 \tag{3.8}$$

where p_{ij} is the j^{th} pitch value in i^{th} cluster, c_i is the i^{th} cluster centroid obtained from the k-Means algorithm, L_i is the number of pitch values in i^{th} cluster, and i ranges from $1, 2, ..., k$ number of clusters.

Binning

Another way to measure the clustering of the pitch values is by simply dividing the 1200 cents (or 120 pitch bins) into 12 equi-spaced semitone bins, and computing the average distance of each pitch value to its corresponding bin centroid. The bin centroid is the average of the pitch values present in that bin. This method is simpler in computation than the k-means clustering method. Equations 3.7 and 3.8 hold true for this method too, the only difference is that the cluster boundaries are fixed in binning method at 100 cents (or 10 pitch bins).

Table 3.1: List of musically-motivated absolute and inter-singer relative features

Feature Group	**Sub-group based on**	**Feature names**
Musically-motivated absolute features	Overall pitch distribution	$Kurt$, $Skew$
	Pitch concentration	$PeakBW$, $PeakConc_{110}$, $PeakConc_{50}$, $Autocorr$
	Musical notes as references	$kMeans$, $Binning$
Inter-singer relative features	Pitch	$pitch\_med\_dist$, $pitch\_med\_L2$, $pitch\_med\_L6\_L2$, $pitchhist12DDistance$, $pitchhist120DDistance$, $pitchhistKLD12$, $pitchhistKLD120$
	Rhythm	$molina\_rhythm\_mfcc\_dist$, $rhythm\_L2$, $rhythm\_L6\_L2$
	Timbre	$timbral\_dist$

In summary, we have eight musically-motivated absolute measures for evaluating singing quality without a reference (Table 3.1): $Kurt$, $Skew$, $PeakBW$, $PeakConc_{110}$, $PeakConc_{50}$, $kMeans$, $Binning$, and $Autocorr$.

3.3 Inter-Singer Features

Another approach for evaluating singing quality without a reference is by leveraging on the general behaviour of a large number of singers singing the same song. This is a novel crowd-sourcing approach that captures the inter-singer statistics and leads us to a self-organizing way of rank-ordering singers. We believe that a song can be sung correctly by many people in one consistent way, but incorrectly in many different and dissimilar ways. That leads us to our hypothesis that good singers sing a song in the same or similar ways, while poor singers sing the same song in many different and dissimilar ways. Therefore, the singers who are similar to each other are good singers, while those who sing differently are the poor ones.

The problem of discovering good singers from a large pool of singers is similar to the problem of finding “true" facts from a large amount of conflicting information provided by various websites [YHP08, PMSW17, LC18]. The truth-finder algorithm utilizes the relationships between websites and their information. A website is trustworthy if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many

Table 3.2: Comparison between our hypothesis and the heuristics in the Truth-Discovery Literature

Our Hypothesis	**Analogy from Truth-Discovery Literature [YHP08]**
Usually there is only one correct way to sing a song	Usually there is only one true fact for a property of an object
This song can be sung correctly in only one or a few similar ways	This true fact appears to be the same or similar on different websites, eg. "Jennifer Widom" versus "J. Widom"
Poor singers will sing the song in many different and dissimilar ways	Different websites often make different mistakes for the same object and thus provide different false "facts".

trustworthy websites. The premise of the truth-finder algorithm is the heuristic that there is only one true version of a fact, and the true fact should appear in the same or similar way on different websites. Moreover, the false "facts" will be different and dissimilar between websites, because there can be many different ways of making a mistake. Our hypothesis about singing quality follows the same heuristics, as discussed in Table 3.2.

The computational model for truth-finding is based on the following two criteria: (a) a website's trustworthiness is proportional to the average of the confidence of the facts stated by it, and (b) the confidence of a fact is proportional to the trustworthiness of the websites providing that fact. On the basis of our hypothesis, we modified these criteria for our task of talented-singer finding. We say that (a) a singer's quality is proportional to the quality of the different perceptual parameters such as pitch, rhythm, and timbre, and (b) the goodness of a perceptual parameter is proportional to the similarity of a singer with the rest of the singers, which relates to the singer's quality.

The next question is how do we test if this hypothesis is correct. Let's first define a feature that represents a perceptual parameter of singing quality, say pitch contour. We compare this feature of each singer with every other singer by a distance metric, where all the singers are singing the same song. According to our hypothesis, a good singer will be similar to the other good singers, therefore they will be close to each other in terms of the distance metric, whereas a bad singer will be far from everyone. Figure 3.3 demonstrates a radial visualization of the Euclidean distance between the pitch contours of 100 singers, where each radial plot shows the distance values of one singer from the 99 other singers. It is evident that the best singers (top-ranked) are similar to other singers, therefore they are clustered together at the center, whereas the worst singers show a large distance from every one else, thus spreading away from the center. This behavior supports our hypothesis that good singers are similar, and poor quality singers are dissimilar. This also points us to a method of ranking singers with the help of other singers in a large pool of singers, instead of relying on one ideal reference singer. Characterizing this behavior will result in singing

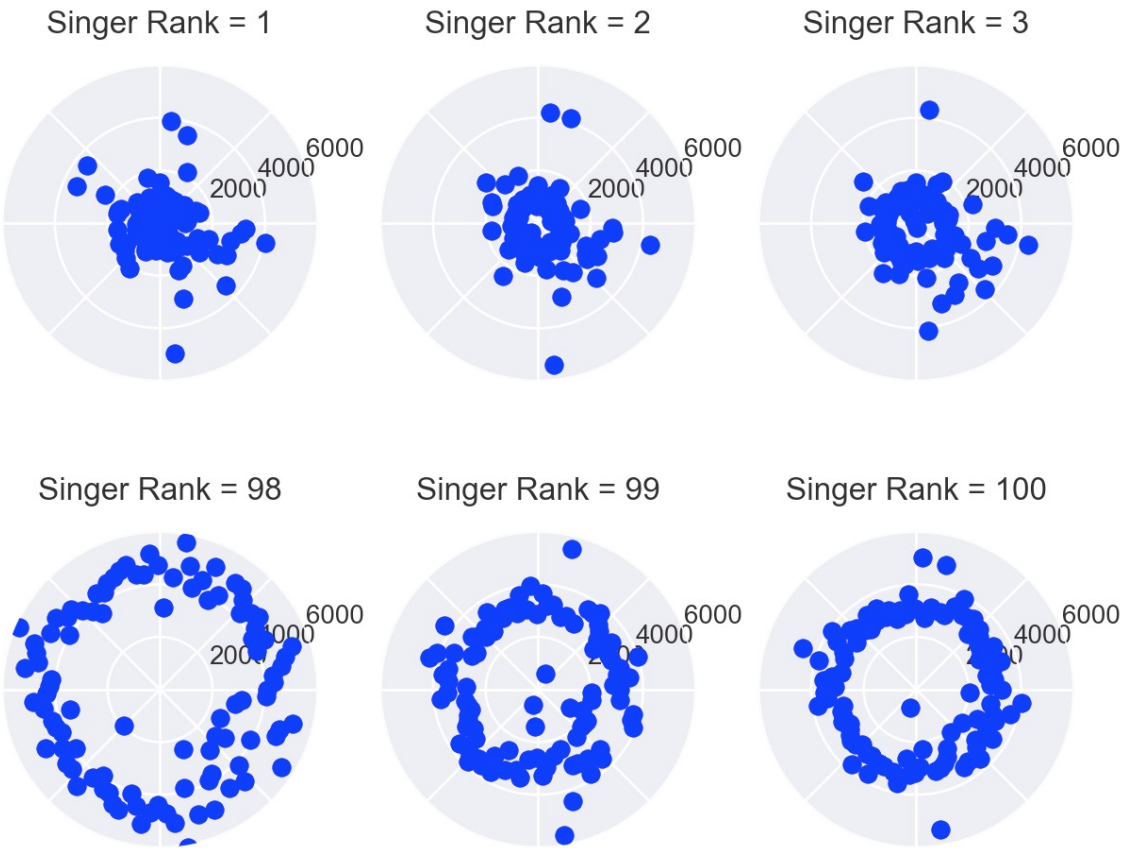


Figure 3.3: Visualization of the pitch-based relative measure distance metric *pitch_med_dist* between each singer and the remaining 99 singers, for the best 3 (top row) and the worst 3 (bottom row) singers among 100 singers singing the song "Let it go".

quality assessment as per a perceptual parameter, and the fusion of these assessments from the different perceptual parameters will result in the singer's overall quality rank.

In the following sub-sections, we discuss the various distance metrics that we have designed to measure the distance between one singer from another, i.e. inter-singer distance, as summarized in Table 3.1. These metrics measure the distance in terms of the perceptual parameters pitch, rhythm, and timbre.

3.3.1 Musically-Motivated Inter-Singer Distance Metrics

In this section, we briefly discuss various relative distance metrics of singing quality evaluation. In Chapter 2, these distance measures were used to compare against a reference ideal singer. But in this study, we utilize these measures to derive inter-singer statistics.

Pitch-based Relative Distance

In Chapter 2, we computed various features that compared a test singer against a reference singer and measured different aspects of pitch accuracy. In this work, we apply them to compare one singer with another, instead of a reference. The distance metrics we use here are the DTW-based distance between the two median-subtracted pitch contours (*pitch_med_dist*), the PESQ-based [RBHH01] cognitive modeling theory [HHG94]-inspired pitch disturbance measures *pitch_med_L6_L2* and *pitch_med_L2*.

Additionally, in this work, we also compute pitch histogram-based relative distance metrics. Pitch histograms provide insights about the statistics of the pitch content of a rendition. As seen in Figure 3.1, there is a clear distinction between the pitch distribution of a good and a poor singer. We compute the Kullback-Liebler (KL) divergence between the normalized pitch histograms to measure the distance between the histograms of singers. Moreover, as the pitch histogram is computed after subtracting the median of the pitch values, not the actual key in which the song is sung, the pitch histograms are expected to be shifted by a few bins. Therefore, we also compute DTW-based distance of the 12-bin and 120-bin histograms between singers as relative measures (*pitchhist12KLdist, pitchhist120KLdist, pitchhist12Ddist, pitchhist120Ddist*).

Rhythm-based Relative Distance

Rhythm is defined as the regular repeated pattern in music, that relates to the timing of the notes sung, and is often referred to as tempo. In karaoke singing, rhythm is determined by the pace of the background music of the song. The lyrics cue on the screen helps the singer to keep up with the pace of the song. Therefore rhythm inconsistencies in karaoke singing only occurs when the singer is unfamiliar with the melody and/or the lyrics of the song.

MFCCs are commonly used in speech recognition as they capture the short-term power spectrum of the speech signal that represents the shape of the vocal tract and thus the phonemes uttered. In Chapter 2, we assumed that the sequences of phonemes and words are uttered correctly by the test singer. So if the words are uttered at the same pace as the reference singer, then the rhythm should be correct. Thus we computed the DTW alignment between the test singer utterance and the reference singer utterance with respect to their MFCC vectors. In this work, we use the three best performing rhythm measures from the previous chapter: the modified version of Molina et al.'s [MBG+13] rhythm deviation measure (*molina_rhythm_mfcc_dist*), PESQ-based *rhythm_L6_L2*, and *rhythm_L2*.

Timbre-based Relative Distance

Perception of timbre often relates to the voice quality [CLLY08, TL12]. Timbre is physically represented by spectral envelope of the sound, which is captured well by MFCC vectors, as shown in [Pra08]. We compute the *timbral_ dist* as the DTW distance between the MFCC vectors between the renditions of two singers.

3.3.2 Inter-singer Distance based Relative Measure Computation

Our goal is to rank a large number of singers singing the same song according to their singing quality, where we do not have a standard reference singing vocal. As discussed earlier, the quality of a singer can be judged from his/her distance from the rest of the singers. As can be observed in Figure 3.3, when a singer is similar to a number of other singers, there is a dense cluster of singers in the center, which is typically seen in singers rated as good singers by humans. Therefore, this similarity cluster at the center for a singer is indicative of good singing. On the other hand, singers who are far away from most other singers is indicative of poor singing. Therefore we base the relative measure computation on inter-singer distance with respect to all the pitch, rhythm, and timbre distance measures discussed in the previous sub-section. We have introduced the way to measure distance between singers for pitch, rhythm, and timbre. We will elaborately discuss the relative measure computation methods in Section 3.6.4.

3.4 Feature Analysis and Fusion

The primary objective of a leaderboard is to give feedback to the singer about where they rank in relation to their contemporaries, or how they progress over time. However, all the musically-motivated absolute features and the inter-singer relative features give numerical scores, where the actual values for a metric may not be universally accepted. For example, as the best-worst scaling (BWS) theory [MFA15] says, humans are known to be able to choose the best and the worst in a small set of choices. However, when humans are asked to absolutely rate singers on a scale of say 1 to 5, they do not reveal discriminatory results. Therefore, it makes sense to convert the numerical values of the features into ranks, and design our algorithm towards a better prediction of the overall rank-order of the singers.

3.4.1 Individual Measures Analysis

We analyze how well can each of the absolute and relative measures individually predict the ranks of the singers. Figure 3.4 shows the Spearman's rank correlation of each of the 8 absolute and the 11 relative measures with the human BWS ranks (explained further in the

Section 3.5.2). We can see that all the derived measures show a positive correlation with humans, although some correlate better than others. The *autocorr* measure shows the best correlation among the absolute measures. The relative measures, in general, perform better than the absolute measures, which means that our method of inter-singer comparison as a way of evaluating singers is closer to how humans evaluate singers. The pitch-based relative measures perform better than the rhythm-based relative measures. This is an expected behavior for karaoke performances, where the background music and the lyrical cues help the singers to maintain their timing. Therefore, the rhythm-based features do not contribute as much in rating the singing quality. Among the relative measures, *pitchhist120DDistance* performs the best, along with the KL-divergence measures, showing that inter-singer pitch histogram similarities is a good indicator of singing quality. The *pitch_ med_ dist* measure follows closely, indicating that the comparison of the actual sequence of pitch values and the duration of each note give valuable information for assessing singing quality. These aspects are not captured by the pitch histogram-based methods. Another interesting observation is the high correlation of the *timbral_ dist* measure. It indicates that voice quality, represented by the timbral distance, is an important parameter when humans compare singers to assess singing quality. This observation supports the timbre-related perceptual evaluation criteria of humans, as discussed in Chapter 1, Section 1.3.1. Music experts judge based on various timbre-related criteria such as *timbre brightness, color/warmth, vocal clarity, strain.* These criteria relate to certain aspects of the spectrum, other than pitch and rhythm. Although not exactly defined in terms of the perceptual criteria, the timbral distance feature does capture the overall spectral characteristics.

3.4.2 Strategies for Feature Fusion

Each of the absolute and relative features can provide a rank-ordering of the singers. To arrive at an overall ranking of the singers, these features need to be combined or fused together. One way to compute an overall ranking is by computing an average of the ranks of all the features. This method of feature fusion does not need any statistical model training, but gives equal importance to all the features. Some features may be more important or a better predictor than others. So we also use a linear regression model that gives different weights to the features according to their importance. Owing to the success of neural networks and the possibility of a non-linear relation between the features and the overall rank, we also explore neural network models to predict the overall ranking from the absolute and the relative features.

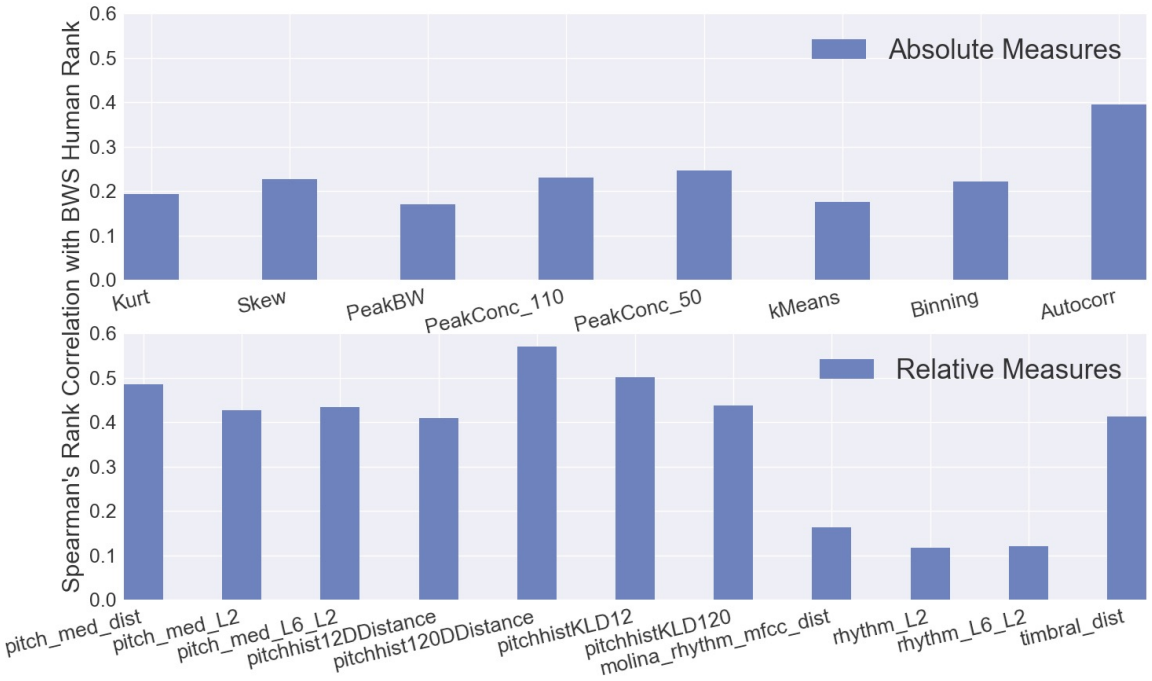


Figure 3.4: Spearman's rank correlation of the individual absolute measures (top) and relative measures (bottom) with human BWS ranks.

3.5 Data Preparation

To test our methods of evaluating singing quality without a reference, we conducted experiments using the musically-motivated absolute features, the inter-singer distance based relative features, and combinations of these features. In this section, we discuss the singing voice dataset and the subjective ground-truths used for these experiments.

3.5.1 Singing Voice Dataset

Our dataset consists of 4 popular Western songs each sung by 100 unique singers (50 male, 50 female) extracted from Smule's DAMP dataset [Smu]. DAMP dataset consists of 35k solo-singing recordings without any background accompaniments. Our selected subset of songs were the most popular four songs in the DAMP dataset with more than 100 unique singers singing them. All the songs are rich in steady notes and rhythm, as summarized in Table 3.3. The dataset consists of a mix of songs with long and sustained as well as short duration notes. They also have a range of different tempi (beats per minute).

We divide every song into 4 snippets, where each snippet is of approximately 20 seconds in duration. Such short duration clips are recommended for the relative feature computation as shorter duration segments are less prone to misalignments during DTW [MJTG98].

Table 3.3: Summary of the singing voice dataset.

#	**Song Name**	**Nature of melody**	**Tempo (bpm)**
1	*Let it go* (Frozen)	Pitch range is more than an octave, mix of short and long duration notes	68
2	*Cups* (Pitch Perfect)	Pitch range is within an octave, mostly short duration notes	130
3	*When I was your man* (Bruno Mars)	Pitch range is more than an octave, mix of short and long duration notes	73
4	*Stay* (Rihanna)	Pitch range is within an octave, mix of short and long duration notes	112

In this work, we use the pitch estimates from the autocorrelation-based pitch estimator PRAAT[B+02, Rab77] with one generic post-processing step to remove unreliable pitch values by detecting and removing the frames with low periodicity, as described in detail in Chapter 2.

3.5.2 Subjective Ground-Truth

We need subjective ratings as ground-truth to validate objective measures for singing evaluation. Reliable subjective ratings for singing quality can be provided by trained or professional music experts. However, obtaining such ratings at a large scale is a challenging task. Music experts may not be easily available, and the process of obtaining these ratings from them is time consuming, and expensive. Previously, crowd-sourcing platforms have been used for labor-intensive large-scale data collection tasks such as speech transcription [MBR10], speech quality assessment tasks [RFZS11, NPW+15, PBTO13], and speech pronunciation quality assessments [WM12]. In another related work that is discussed in detail in Chapter 5, we have validated a method to leverage on crowd-sourcing platforms, such as Amazon mechanical turk (MTurk), to collect reliable human judgments for singing quality in a scalable and cost-effective manner, while filtering noisy data. We have showed that the ratings provided by MTurk users correlated well with the ratings obtained from professional musicians in a laboratory-controlled experiment (please refer Chapter 5). The Pearson's correlation between lab-controlled music-expert ratings and filtered MTurk ratings for various parameters are as follows: overall singing quality: 0.91, pitch: 0.93, rhythm: 0.93, pronunciation: 0.86, voice quality: 0.65, vibrato: 0.88, volume: 0.65, and pitch dynamic range: 0.88.

Although professional musicians are able to rate singing quality on an absolute scale of 5 reliably and consistently, on crowd-sourcing platforms however, we cannot be sure of their absolute ratings. Also, absolute ratings are known to not discriminate between items, and each rating on the scale is not precisely defined [MFA15, LLI+13]. Therefore, we used a

method of relative rating called *best-worst scaling* (BWS) which can handle a long list of options and always generates discriminating results as the respondents are asked to choose the BEST and WORST option in a choice set [LLI+13, ÇLT17]. At the end of this exercise, the items can be rank ordered according to the aggregate BWS scores of each item, given by

$$BWS_{score} = \frac{n_{best} - n_{worst}}{n} \tag{3.9}$$

where n_{best} and n_{worst} are the number of times the item is marked as best and worst respectively, and n is the total number of times the item appears. The Spearman's rank correlation between the MTurk experiment and the lab-controlled experiment reported in [GLW18] was 0.859.

In this work, we conducted a pairwise BWS test on MTurk where a listener was asked to choose the better singer among a pair of singers singing the same song. We presented one excerpt of approximately 20 seconds from every singer of a song (the same 20 seconds for all the singers of a song). There are ${}^{100}C_2$ number of ways to choose 2 singers from 100 singers of a song, i.e. 4950 Human Intelligence Tasks (HITs) per song. This experiment was conducted separately for each of the 4 songs of Table 3.3. Therefore there were in total $4950 \times 4 = 19,800$ HITs.

We applied filters to the MTurk users in the same way as we have done in Chapter 5. We asked the users for their experience in music and asked them to annotate musical notes. We accepted their attempt only if they had some formal training in music, and could write the musical notations successfully. For example, a user whose attempt was accepted had mentioned in his/her music skill description, "I am a classical voice teacher, and double bass teacher. I also play the piano and sing in public quite often." We made exceptions for the music notations if they mentioned that they were trained in music, but haven't learnt the Western music notation style. We also applied a filter on the time spent in performing the task to remove the less serious attempts where they may not have spent time listening to the tracks. Empirically we set the time threshold as 10 seconds, i.e. an attempt is accepted only if it took more than 10 seconds to complete.

3.6 Experiments

In Sections 3.2, and 3.3, we designed various objective measures that, we believe, can assess the inherent properties of singing quality that are independent of a reference. This will result in automatic generation of a leaderboard of singers ranked in the order of their singing ability. To validate our hypothesis, we conduct several experiments. We investigate the role of the absolute and the relative features individually in predicting the overall human judgment, and the methods of combining these features. We also observe the influence of the duration of a song excerpt for computational singing quality analysis. Moreover, we compare

the ability of our machine-based features and humans in predicting the performance of the underlying perceptual parameters.

In this section, we first discuss the baseline performance from the literature, and the achievable upper limit of performance in the form of the human judges' consistency in evaluating singing quality. Then we describe our system and its configurations followed by experiments and result analysis.

3.6.1 Baseline

As discussed earlier, Nakano et al. [NGH06a] and Bohm et al. [BES+17] attempted to evaluate singing quality without a reference. They measured the consistency of the offset of pitch values within a musical semitone grid. They used the global statistics kurtosis and skew to measure the consistency of pitch values. These are two of our eight absolute measures. Moreover, [BES+17] used the Interspeech ComParE 2013 (Computational Paralinguistics Challenge) feature set as baseline. It comprises of a total of 6,373 acoustic features per audio segment or snippet, such as loudness, pitch, MFCCs, and their 1st and 2nd order derivatives [SSB+13]. We extract this same set of features using OpenSmile toolbox [EWGS13] to create our baseline for comparison. We conducted a 10-fold cross-validation experiment using the snippet 1 from all the songs to train a linear regression model with these features. The Spearman's rank correlation between the human BWS rank and the output of this model is 0.39.

3.6.2 Upper limit of performance

In the experiments in Chapter 2, we recruited 5 professional musicians to provide singing quality ratings for 10 singers singing a song. These judges were trained in vocal and/or musical instruments in different genres of music such as jazz, contemporary, and Chinese orchestra, and all of them were stage performers and/or music teachers. The subjective ratings obtained from them showed high inter-judge correlation of 0.82. This shows that humans do not always agree with each other, and there is, in general, an upper limit of the achievable performance of any machine-based singing quality evaluation. Thus, the goal of our singing evaluation algorithm is to achieve this upper limit of correlation with human judges.

3.6.3 System Configuration

Our singing evaluation algorithm provides a rank-ordering of the singers singing a particular song using the absolute and relative features, as shown in Figure 3.5. As discussed in Section 3.4.2, the overall rank-order of the singers can be computed as an average of the ranks of all

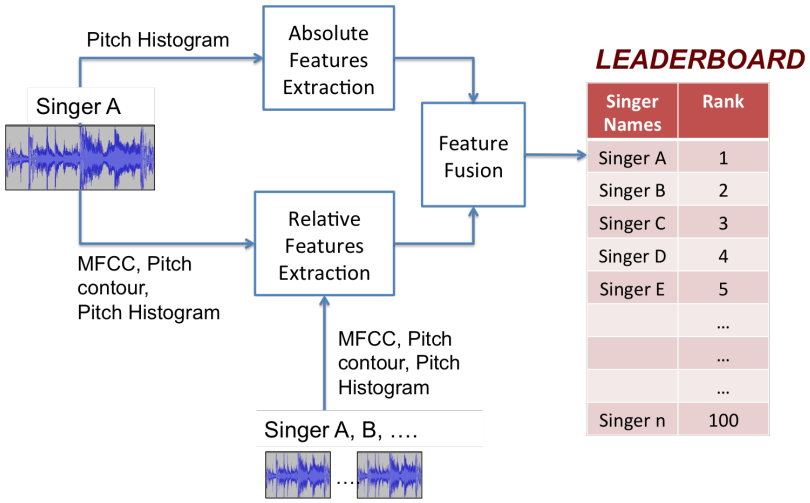


Figure 3.5: Overview of the framework for automatic singing quality leaderboard generation, consisting of fusion of musically-motivated absolute features and inter-singer statistics based relative features.

the features. Moreover, we train a linear regression model, and four different neural network models in 10-fold cross-validation to account for the different importance of the different features. We ensure that in every fold, equal number of singers are present from every song, both in train and test data. Two of the neural network models consist of no hidden layers, but a non-linear activation function, sigmoid or ReLU. The other two neural network models consist of one hidden layer with 5 nodes, with sigmoid or ReLU as activation functions for both the input and the hidden layers. All computations are done using scikit-learn [PVG+11], and the models are summarized in Table 3.4.

3.6.4 Experiment 1: Deciding the best relative feature computation method

As discussed in Section 3.3.1, the distance between one singer, say singer A from the each of the other singers (singers B, C...) singing the same song is indicative of singer A's quality. If more number of singers are similar to A, i.e. distances are small, indicates that singer A is good, and vice versa. Based on this hypothesis, we explored three ways of computing a relative feature value for each singer.

1. **Number of singers within a fixed distance threshold:** We can set a constant threshold on the distance value across all singer clusters and count the number of singers within the set threshold as the relative measure. It measures the number of singers in close proximity to that particular singer. If a large number of singers are similar to that singer, then the number of points within the threshold circle will be high.

Table 3.4: Summary of the feature fusion models. ($\mathbf{r}_i$ = rank-ordering of singers according to i^{th} feature; N = # of features; $\mathbf{x}$ = feature vector; $\mathbf{w}^i$ = weight vector of i^{th} layer; $\mathbf{b}$ = bias; $S(.)$ = sigmoid activation function; $R(.)$ = ReLU activation function)

#	**Model Name**	**Description**	**Equation**
1	Avg. rank	The individual features are first rank-ordered, then the ranks are linearly combined with equal weights to all the features	$y = \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i$
2	Linear Regression	Weighted sum of features	$y = \mathbf{b} + \mathbf{w}^T\mathbf{x}$
3	NN-1	MLP with sigmoid activation, no hidden layer	$y = S(\mathbf{b} + \mathbf{w}^T\mathbf{x})$
4	NN-2	MLP with ReLU activation, no hidden layer	$y = R(\mathbf{b} + \mathbf{w}^T\mathbf{x})$
5	NN-3	MLP with sigmoid activation, 1 hidden layer with 5 nodes	$y = S(\mathbf{b^{(2)}} + \mathbf{w^{(2)}}S(\mathbf{b^{(1)}} + \mathbf{w^{(1)}}^T\mathbf{x}))$
6	NN-4	MLP with ReLU activation, 1 hidden layer with 5 nodes	$y = R(\mathbf{b^{(2)}} + \mathbf{w^{(2)}}R(\mathbf{b^{(1)}} + \mathbf{w^{(1)}}^T\mathbf{x}))$

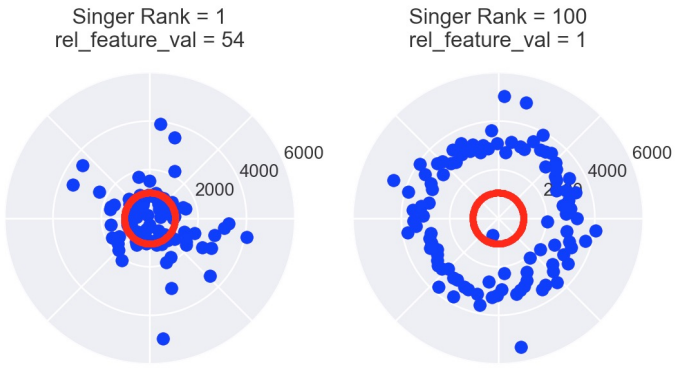
2. **Distance of x^{th} nearest singer:** We can fix the number of singers as the threshold, say x singers and consider the distance of the x^{th} nearest singer as the relative measure. If this distance is small for a singer, the singer is likely to be good.

3. **Median of distances of a singer from all other singers:** The median of all the singer distances in a singer cluster can be assigned as the relative measure. This median distance feature of a singer represents his/her overall distance from the rest of the singers. Median is taken instead of mean to avoid the effect of outliers. If this distance is small for a singer, the singer is likely to be good.

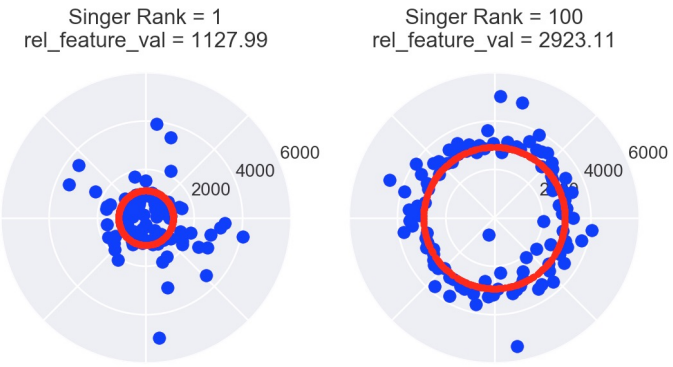
Figure 3.6 demonstrates the relative feature computation from the *pitch_median_dist* measure with the three methods for the best and the worst singer out of 100 singers of a song.

To investigate the performance of each of these methods, we obtained the relative feature values from these methods for each of the 11 intonation and rhythm distance measures. The distance threshold for method 1 is optimized for each feature. The number of singers threshold for method 2 is empirically set as 10 singers, assuming that roughly at least ten percent of singers in a large pool of singers would be good. Therefore, if the distance of a particular singer from the 10th nearest singer is small, it means that at least 10 singers are singing similar to that singer, so the singer is good.

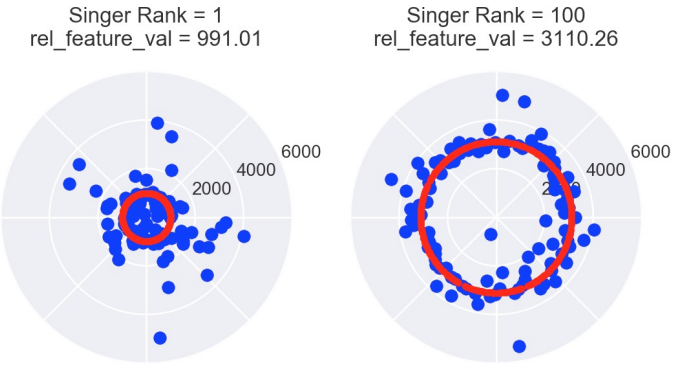
Figure 3.7 shows the Spearman's rank correlation of the human BWS ranks with ranks from these relative features used with the six models of Table 3.4, over the snippet 1 of all the 4 songs for the three methods. We observe that method 2 performs better than the other two methods for all the six models. It implies that our assumption that at least ten percent of



(a)



(b)



(c)

Figure 3.6: Demonstration of relative feature computation methods from the *pitch_ med_ dist* measure for the best (Rank 1) and the worst (Rank 100) singer of an example song (Song 1, snippet 1), along with the respective relative feature values using the three methods: (a) Method 1: Number of singers within a fixed distance threshold (b) Method 2: Distance of x^{th} nearest singer, x=10 (c) Method 3: Median of distances of a singer from all other singers. The red circle in (a) and (b) are the thresholds, while for (c) it is the median value.

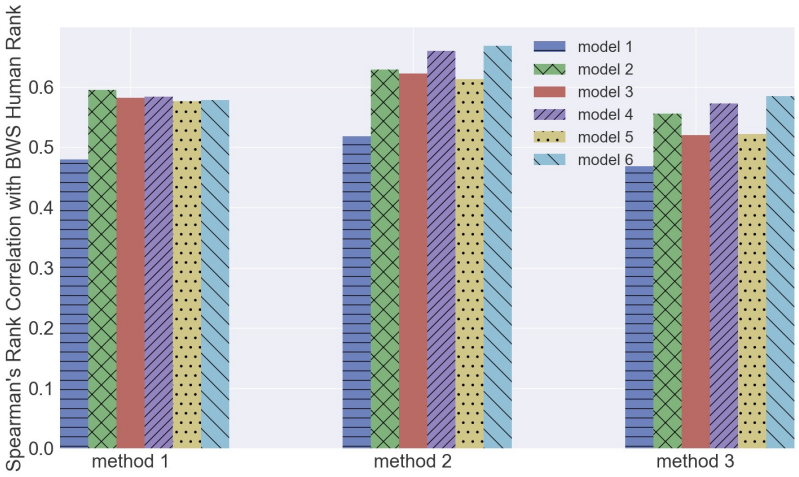


Figure 3.7: Spearman's rank correlation performance of the three methods for inter-singer distance measurement (Section 3.3.2): Method 1=# of singers within a fixed distance threshold; Method 2=Distance of the 10^{th} nearest singer; Method 3=Median of distances of a singer from all other singers. Models are as listed in Table 3.4.

singers in a pool of singers would be good works in favor of the outcome. Method 3, i.e. the median of the distances of a particular singer from the rest of the singers assumes that half of the pool of singers would be good singers, which is not a reliable assumption, therefore this method performs the worst.

In the following experiments, all the relative features are computed using method 2.

3.6.5 Experiment 2: Evaluating fusion of the absolute measures

In this experiment, we evaluate the performance of the musically-motivated pitch histogram-based absolute measures that were introduced in Section 3.2 in ranking the singers as per their singing quality. Table 3.5 shows the Spearman's rank correlation of the human BWS ranks with the ranks predicted by absolute measures with different feature fusion models. We evaluate four different snippets from each song and average the ranks over multiple snippets. The last column shows the performance of the absolute measures extracted from the full song (more than 2 minutes duration) (*AbsFull*) combined with the individual snippet ranks.

Influence of duration:

The pitch histogram for the full song is expected to show a better representation than the histogram of a snippet of the song because more data results in better statistics. As seen in Table 3.5, with increase in the number of snippets, i.e. increase in the duration of the song being evaluated, the predictions improve, with the one with the full song performing the

best. This indicates that the machine needs longer duration statistics or more data (∼80 seconds) to give better predictions, while humans can judge reliably by a shorter duration clip of ∼20 seconds.

Influence of the feature fusion models:

As some absolute features are more important than the others, the weighted combination of the features with non-linear activation functions (Models 5 and 6) show better performance than the equally weighted average of ranks (Model 1). One hidden layer in the neural network models performs better than the ones without a hidden layer. Interestingly, the average of ranks (Model 1) performs comparably with the others, showing that the features are good enough to rank the singers without needing to train a statistical model. It also indicates that although the features individually may not have performed equally well (Figure 3.4), each of them captures a different aspect of the pitch histogram quality, therefore, combining them with equal weights results in a comparable performance.

It is important to note that there are specific conditions when the absolute measures fail to perform [GLW18]. By converting a pitch contour into a histogram, information about timing or rhythm is lost. The correctness of the note order also cannot be evaluated through the pitch histogram. Moreover, the relative positions of the peaks in the histogram cannot be modeled without a reference, i.e. incorrect location of peaks goes undetected. For example, if a song consists of five notes, and a singer sings five notes precisely but they are not the same notes as that present in the song, then the absolute measures would not be able to detect it. Pitch histogram also loses the information about localized error or error that occurs for a short duration. According to cognitive psychology and PESnQ measures [HHG94, RBHH01, GLW17], localized errors have greater subjective impact than distributed errors. Therefore if a singer sings incorrectly for a short duration, and then corrects himself/herself, the absolute measures are unable to capture it.

3.6.6 Experiment 3: Evaluating combination of the relative measures

In this experiment, we investigate the performance of the combination of the inter-singer relative measures computed from method 2 in Section 3.6.4. Table 3.6, third column shows the Spearman's rank correlation of the human BWS ranks with the ranks predicted by the relative measures with the different feature fusion models, averaged over the four snippets.

The combinations of the relative features result in a better performance than the combinations of the absolute features. This follows from the observation in Section 3.4.1 that the relative features individually perform better than the absolute features. Like the absolute features, some relative features are also more important than others, therefore a weighted combination

Table 3.5: Absolute features performance evaluation. The values in the table are Spearman's rank correlation between Human BWS ranks and the machine generated ranks. (All p-values<0.05)

Model #	**Snippet 1**	**Snippets 1+2**	**Snippets 1+2+3**	**Snippets 1+2+3+4**	**Snippets 1+2+3+4 +AbsFull**
1	0.3556	0.4134	0.4702	0.4796	0.4796
2	0.3695	0.3879	0.4143	0.4205	0.4558
3	0.3329	0.3567	0.3917	0.3975	0.4331
4	0.3073	0.3372	0.3866	0.3838	0.4228
5	0.3924	0.4589	0.4781	0.4711	**0.4942**
6	0.3860	0.4475	0.4650	0.4603	0.4887

Table 3.6: Summary of the performance of absolute and relative measures, and their combinations. The values in the table are Spearman's rank correlation between Human BWS ranks and the machine generated ranks averaged over 4 snippets.(All p-values<0.05)

Model #	**Absolute Features**	**Relative Features**	**Early-fusion**	**Late-fusion**
1	**0.4796**	**0.6396**	**0.6877**	**0.7059**
2	0.4205	0.5737	0.6413	0.6426
3	0.3975	0.5799	0.6385	0.6407
4	0.3838	0.5688	0.6222	0.6274
5	0.4711	0.6153	0.6636	0.6692
6	0.4603	0.6020	0.6623	0.6678

of the relative features with a non-linear activation function (Models 5 and 6) perform better than the other feature fusion models.

3.6.7 Experiment 4: Combining absolute and relative measures

In this experiment, we investigated combinations of the 8 absolute and 11 relative features. One method of combination was early-fusion where we concatenated the features to get a 19 dimensional feature vector for each snippet. The rank correlation of the BWS ranks with the ranks obtained from early-fusion method averaged over four snippets is reported in Table 3.6 column 4.

The second method of combining the features is late-fusion. Instead of fusing the feature vectors, we computed the average of the ranks predicted separately from the absolute and

the relative features. Since in experiment 3, we observed that the absolute features computed for the full song gives a better prediction, therefore we use the full song absolute features in this experiment. Table 3.6 column 5 shows the rank correlation results from the late fusion method.

The results indicate that the late-fusion of the features show a better correlation with humans than early-fusion. This means that predictions given separately from the absolute and the relative features provide different and equally important information. Therefore equal weighting to both shows better correlation with humans. Moreover the results from Models 1, 5, and 6 are comparable, i.e. a simple rank average shows performance similar to the complex neural network models. This shows that the individual features, although show different amounts of correlation with the humans, they individually capture different information about singing quality. Therefore, equal weighting of the ranks provided by them individually shows comparable results.

3.6.8 Experiment 5: Humans versus Machines

An important advantage of objective methods for singing evaluation is that each underlying perceptual parameter is objectively evaluated independent of the other parameters, i.e. the computed features are uncorrelated amongst each other. On the other hand, the individual parameter scores from humans tend to be biased by their overall judgment of the rendition. For example, a singer who is bad in pitch, may or may not be bad in rhythm. However, humans tend to rate their rhythm poorly due to bias towards their overall judgment. In this experiment, we used the data from the previous chapter 2.4.2, where music experts were asked to rate each singer on a scale of 1 to 5 with respect to the three perceptual parameters pitch, rhythm, and timbre individually. Figure 3.8(a) shows that human ratings for the three perceptual parameters are highly correlated amongst each other. On the same data, machine scores for the three parameters show significantly less correlation (Figure 3.8(b)). We also verified this observation on the data used for the other experiments in this work (Figure 3.8(c)). Therefore, machine scores are better than humans in giving unbiased objective feedback to a singer on the underlying perceptual details of their rendition. This is useful to a learner for understanding how they can improve upon the individual parameters.

3.7 Conclusions and Future Work

In this work, we successfully evaluated singing quality without a reference singing sample or musical score by leveraging on musically-motivated absolute measures and *truth-finding* based inter-singer relative measures. The baseline methods show a correlation of 0.39 with human assessment using linear regression, while the linear regression model with our proposed measures shows a correlation of 0.64, and the best performing method shows a correlation of

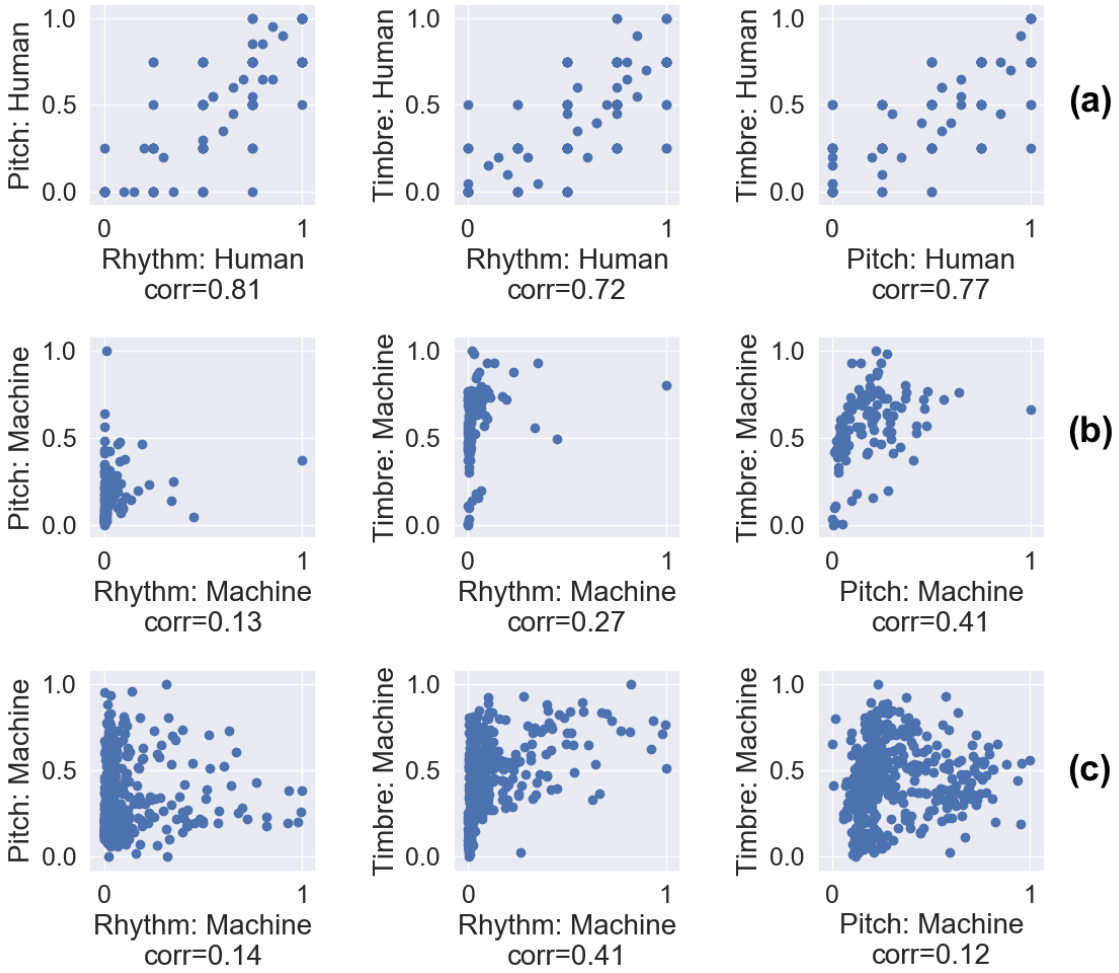


Figure 3.8: Humans vs. Machines: Correlation between scores given individually for pitch, rhythm, and timbre by (a) human experts on the data in Section 2.4.2,(b) machine, on the same data as in (a), and (c) machine, on the data used in this chapter, i.e. Table 3.3.

0.71, which is an improvement of 82.1% over the baseline. This improvement is attributed to:

- the musically-motivated absolute measures, that quantify various singing quality discerning properties of the pitch histogram, and
- the novel *truth-finding* based musically-informed relative measures that leverages on inter-singer statistics and overcomes the drawbacks of using only absolute measures

We find that the two kinds of measures provide distinct information about singing quality, therefore a combination of them boosts the performance. Our proposed method has led to a self-organizing way of rank-ordering a large pool of singers according to their singing quality. Moreover we show that our proposed measures have the ability to provide objective scores for the individual perceptual parameters such as pitch, rhythm, and timbre independently, whereas human assessments for the individual parameters get biased by their overall judgment of the rendition.

Human experts, in general, are more consistent amongst themselves than the machine scores, with a correlation of 0.82 (Section 3.6.2). Thus, the machine scores still have some scope of improvement. Extension of the proposed methods to music genres other than Western pop also needs to be investigated in future.

Part II

Pronunciation aspect of singing voice

Summary

In this part, we discuss the pronunciation aspect of singing voice. We present our ideas on acoustic modeling of sung phonemes, that has various applications in Music Information Retrieval (MIR), such as lyrics recognition, lyrics alignment, pronunciation evaluation, etc. In Chapter 4, we formulate a method to build the acoustic models with limited singing data. In Chapters 5 and 6, we observe the usability of these models for two applications of MIR: pronunciation evaluation in singing and audio-to-lyrics alignment, respectively.

The research findings of Chapter 4 have appeared in our publication:

- Chitralekha Gupta, Rong Tong, Haizhou Li, and Ye Wang, "Semi-supervised lyrics and solo-singing alignment", *In Proceedings of International Society of Music Information Retrieval (ISMIR)*, Paris, 2018.

Parts of Chapter 5 were reflected in the following publications:

- Chitralekha Gupta, David Grunberg, Preeti Rao, and Ye Wang, "Towards Automatic Mispronunciation Detection in Singing", *In Proceedings of International Society of Music Information Retrieval (ISMIR)*, Suzhou, 2017.
- Chitralekha Gupta, Haizhou Li, and Ye Wang, "Automatic Pronunciation Evaluation of Singing", *In Proceedings of Interspeech*, Hyderabad, 2018.

Chapter 6 consists of the content from our submitted paper:

- Chitralekha Gupta*, Bidisha Sharma*, Haizhou Li, and Ye Wang, "Automatic Lyrics-to-Audio Alignment on Polyphonic Music Using Singing-Adapted Acoustic Models", *Submitted to ICASSP*, 2019. (*equal contributors)

Chapter 4

Phonetic modeling of singing voice

Lyrics serve as an important component of music, that often defines the mood of the song [AP06, BAB+11], affects the opinion of a listener about the song [ABD+81], and even improves the vocabulary and pronunciation of a foreign language learner [NS11, GRS15b]. Research in MIR in the past has explored tasks involving lyrics such as automatic lyrics recognition [MEG14, MV10, Kru16a, Han12] and automatic lyrics alignment [FGOO11, MV08, CWJ16] for various applications such as karaoke singing, song subtitling, query-by-singing as well as acoustic modeling for singing voice. In spite of huge advances in speech technology, automatic lyrics transcription and alignment in singing face challenges due to the differences between sung and spoken voices [FGOO11, MEG14], and a lack of transcribed singing data to train phonetic models for singing [FGOO11, MEG14, MV10, MV08, Han12].

As singing and speech differ in many ways such as pitch dynamics, duration of phonemes, and vibrato [FGOO11, MEG14], the direct use of automatic speech recognition (ASR) systems for lyrics alignment or transcription of singing voice will result in erroneous output. Therefore, speech acoustic models need to be adapted to singing voice [MV08]. For training singing-adapted acoustic models, lyrics-aligned singing dataset is necessary. Lack of annotated singing datasets has been a bottleneck for research in this field. Duan et al. [DFL+13] published a small singing dataset (1.92 hours) with phone-level annotations, which were done manually that requires a lot of time and effort, and is not scalable. One way of getting data for training is to force-align the lyrics with singing using speech models, and use this aligned singing data for model training and adaptation. But due to the differences in speech and singing acoustic characteristics, alignment of lyrics with speech acoustic models will be prone to errors, that will result in badly adapted singing acoustic models.

With the increase in popularity of mobile phone karaoke applications, singing data collected from such apps are being made available for research. Smule's *Sing!* karaoke dataset, called Digital Archive of Mobile Performances (DAMP) [Smu], is one such dataset that contains more than 34K a capella (solo) singing recordings of 301 songs. But it does not have time-aligned lyrics, although the textual lyrics are available on Smule's website. The data also contains inconsistencies in recording conditions, out-of-vocabulary words, and incorrectly pronounced words because of unfamiliar lyrics or non-native language speakers. Although

the presence of such datasets is a huge boon to MIR research, we need tools to further clean up such data to make them more usable. There is a need for aligned lyrics transcriptions for singing vocals while also eliminating inconsistent or noisy recordings. To address this need, we propose a simple yet effective solution to produce clean audio segments with aligned transcriptions.

In this work, we study a strategy to obtain time-aligned sung-lyrics dataset with the help of the state-of-the-art ASR as well as an external resource, i.e. published lyrics. We use the speech acoustic models to transcribe solo-singing audio segments, and then align this imperfect transcription with the published lyrics of the song to obtain a better transcription of the sung segments. We hypothesize that this strategy will help in correcting the imperfect transcriptions from the ASR module and in cleaning up bad audio recordings. We validate our hypothesis by a human listening experiment. Moreover we show that a semi-supervised adaptation of speech acoustic models with this cleaned-up annotated dataset results in further improvement in alignment as well as transcription, iteratively. Hence, such an algorithm will potentially automate the labor-intensive process of time aligning lyrics such as in karaoke or MTV. Furthermore, it will enable large-scale singing transcription generation, thus increasing the scope of research in music information retrieval. We have applied our algorithm on a subset of the DAMP dataset, and have published the resulting dataset and code[1].

Our contributions:

- We design a strategy for obtaining sentence-level lyrics alignment with the help of ASR and published lyrics
- With the help of this algorithm, we automatically obtain large-scale singing-lyrics annotated dataset to train singing-adapted speech acoustic models that show promising performance in lyrics recognition

This chapter is organized in the following manner: we discuss the related work in Section 4.1, our proposed alignment algorithm in Section 4.2, experiments for validation of the algorithm in Section 4.3, and conclusions in Section 4.4.

4.1 Background studies

One of the traditional methods of aligning lyrics to music is with the help of the timing information from the musical structure such as chords [MFG10, WKN+04, KWI+08, MFG12], and chorus [LC08], but such methods are more suitable for singing in the presence of background accompaniments. Another study uses musical score to align lyrics [GCOC15], but such methods would be applicable for professional singing where the notes are correctly

[1]Dataset: https://github.com/chitralekha18/lyrics-aligned-solo-singing-dataset.git; Code: https://github.com/chitralekha18/AutomaticSungLyricsAnnotation_ISMIR2018.git

sung. In karaoke applications, as addressed in this work, correctness of notes is less likely.

One of the pioneering studies of applying speech recognition for lyric alignment was by Mesaros and Virtanen [MV08], who used 49 fragments of songs, 20-30 seconds long, along with their manually acquired transcriptions to adapt Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) speech models for singing in the same way as speaker adaptation is done. They then used these singing-adapted speech models to align vocal sections of songs with their manually paired lyrics lines using the Viterbi algorithm. In [MV10], the authors used the same alignment method to automatically obtain the singing-to-lyrics aligned lines, and then explored multiple model adaptation techniques, to report the best phoneme error rate (PER) of 80%. This work has provided a direction for solving the problem of lyrics alignment and recognition in singing, but it suffers from manual post-processing and the models are based on a small number of annotated singing samples.

Recently, with the availability of more singing data, a subset of the DAMP solo-singing dataset was used for the task of sung phoneme recognition by Kruspe [Kru16b, Kru16a]. In her work, the author built new phonetic models trained only on singing data (DAMP data subset) and compared it with a pitch-shifted, time-stretched, and vibrato-applied version of a speech dataset called *songified* speech data TimitM [Kru15]. Their best reported PER was 80%, and weighted PER (that gives 0.5 weights to deletions and insertions) was 56%, using the DAMP data subset, which outperformed the songified dataset. This work shows an effective use of the available (unannotated) singing data to build improved singing phonetic models. But there is still room for improvement.

The first step in Kruspe's work was to obtain aligned lyrics annotations of every song, for which the whole lyrics of a song was force-aligned with the audio using speech-trained models. These force-aligned sung phonemes were then used to build the new acoustic phonetic models for singing. This approach of forced-alignment of singing using speech acoustic models has also been applied in the earlier attempts of automatic lyrics alignment in a capella singing as well as in singing with background music [FGOO11, WKN+04, IWKL06, KWI+08]. But, as noted by Kruspe [Kru16a], forced-alignment of singing with speech models causes unavoidable errors, because of the mismatch between speech and singing acoustic characteristics [FG12, LCB99], as well as the mismatch between the actual lyrics and what the singer sings. Thus, the lack of appropriate lyrics-aligned song dataset and the eventual use of forced-alignment with speech models to obtain this annotation is a source of errors.

4.2 Semi-supervised lyrics and singing vocals alignment algorithm

We propose an algorithm to align lyrics to singing vocals, that consists of two main steps: dividing the singing vocals into shorter segments (Segmentation), and obtaining the aligned

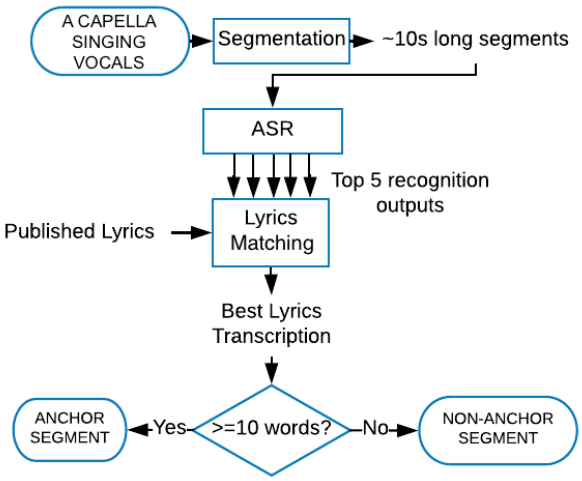


Figure 4.1: The diagram of lyrics to singing vocal alignment algorithm.

lyrics for each segment (Lyrics Matching). Figure 4.1 shows the overview of our algorithm.

4.2.1 Segmentation

One way to automatically align the published lyrics with a solo-singing audio is to force-align the lyrics with the full rendition audio (2 to 4 minutes long) using speech trained acoustic models, as discussed in [Kru16a]. However, the Viterbi alignment algorithm used in forced-alignment, fails to scale well for long audio segments leading to accumulated alignment errors[MJTG98]. In our singing-lyrics transcription and alignment algorithm, we propose to first divide the audio into shorter segments such that the ASR is less prone to the alignment errors. We find silent regions in the rendition by imposing constraints on the magnitude of the short time energy and the silence duration (Algorithm 1). The center of these silent regions are marked as boundaries of non-silent sub-segments. Such non-silent sub-segments are of varying lengths. So we stitch consecutive sub-segments together to make segments of ∼10 seconds duration. We also add silence samples before and after every such segment so that the ASR has some time to adapt to the utterance and start recognition in the beginning of the utterance, and to avoid abrupt termination at the end of the utterance.

Algorithm 1 Segmentation algorithm

1: Calculate short time energy E for 32 ms window with 16 ms hop $E > 0.1 \times mean(E)$ is true
2: non-silent region
3: silent region silent region duration>=200 ms
4: valid silence region
5: center of this region marks the boundary
6: invalid silent region
7: sub-segment = boundary-to-boundary region
8: segment = stitch together such sub-segments for ∼10s duration
9: add 2s silence before and after every segment, to improve ASR performance

4.2.2 Lyrics Matching

We would like to obtain the best possible lyrics transcription for these short singing segments. Moreover, to obtain a clean transcribed dataset of singing vocals, we would also like to reject the noisy audio segments that contain out-of-vocabulary, incorrectly pronounced words, and background noise. We use ASR to decode these segments because such ASR transcription ideally suggests words that are actually sung and different from the published lyrics. The ASR transcription also help detect erroneous pronunciations, reject noise segments. We understand that the the state-of-the-art ASR is not perfect, and for singing it is even more unreliable, as the ASR is trained on speech while singing is acoustically different from speech. So we designed an algorithm to overcome these imperfections of the ASR. This algorithm produces time-aligned transcriptions of clean audio segments with the help of the published lyrics.

Algorithm 2 Lyrics Matching algorithm

1: $X_{N\times 5}$ s.t. $x_{i,j} = e$ where, X = error matrix, N = number of words in published lyrics, e = ratio of number of errors obtained from Levenshtein distance between ASR output and published lyrics window, to the total number of words in the lyrics window
2: $i_{min}, j_{min} = X$ where i_{min} = minimum distance transcription start index in lyrics, where j_{min} = minimum distance transcription slack window size
3: transcription = lyrics$[i_{min} : i_{min} + M + j_{min}]$ where, M is the number of words in ASR transcription

ASR Transcription of Lyrics

To obtain the transcription of each of the audio segments, we use the Google speech-to-text API package in python [Zha17] that transcribes a given audio segment into a string of words, and gives a set of best possible transcriptions. We compare the top five of these transcriptions with the published lyrics of the song, and select the one that matches the most, as described in Algorithm 2. The idea is that the ASR provides a hypothesis of the aligned lyrics although imperfect, and the published lyrics helps in checking these hypothesized lyrics, and retrieving the correct lyrics. Also, we use the Google ASR to bootstrap, with a plan to improve our own ASR (as discussed further in Section 4.3.2). Different ASR systems have different error patterns, therefore we expect that the Google ASR would boost the performance of our ASR. We use the Google ASR only for bootstrapping, the rest of the experiments use our own ASR. Below is the description of the lyrics-matching algorithm.

For an ASR output of length M words, we took a lyrics window of size M, and also empirically decided to provide a slack of 0 to 4 words, i.e. the lyrics window size could be of length M to M+4. This slack provides room for accommodating insertions and deletions in the ASR output, thus allowing improvement in the alignment. So, starting from the first word of the published lyrics, we calculate the Levenshtein distance [Lev66] between the ASR output

and the lyrics window of different slack sizes, iterated through the entire lyrics by one word shifts. This distance represents the number of errors (substitutions, deletions, insertions) occurred in ASR output with respect to the actual lyrics.

For the lyrics of a song containing a total of N words, we obtain an error matrix X of dimensions N×5, where 5 is the number of slack lyric window sizes ranging from M to M+4. Each element e of the matrix is the ratio of the number of errors obtained from Levenshtein distance between the ASR output and the lyrics window, to the total number of words in that lyrics window. If (i_{min}, j_{min}) is the coordinate of the minimum error element of this matrix, then i_{min} is the starting index of the minimum distance lyrics transcription, j_{min} is the slack lyric window size. Amongst the top five ASR outputs, we choose the one that gives minimum error e, and select the corresponding lyrics window from the error matrix to obtain the best lyrics transcription for that audio segment. We illustrate this with the help of the following example.

Let's assume that the ASR transcription of an audio segment is *"the snow glows on the mountain"*, therefore M=6. The slack window size will range from 6 to 10 words. The lyrics of this song contains a total of N words, where a word sub-sequence is *"the snow glows white on the mountain tonight not a footprint to be seen..."*. The corresponding error matrix X is shown in Figure 4.2. The error element $e_{1,2}$ is the distance between the ASR transcription and the slack lyric window "the snow glows white on the mountain" which is 1. The error element $e_{2,1}$ is the distance between the ASR transcription and the slack lyric window "snow glows white on the mountain" which is 2, and so on. So in this example, (i_{min}, j_{min}) is (1,2), i.e. the best lyrics transcription is "the snow glows white on the mountain".

Anchor and Non-Anchor Segments

From our preliminary study, we found that many of the ASR transcriptions had missing words because either the audio contained background noise or there were incorrectly pronounced words or deviation of singing acoustics from speech. For example, a 10 seconds long non-silent segment from a popular English song would rarely ever have as few as four or five words. In order to retrieve more reliable transcripts, we added a constraint on the number of words, as described below.

To check the reliability of the lyrics transcriptions, we marked the best lyrics transcriptions of a small subset of 360 singing segments (extracted across 27 singers singing) as correct or incorrect, depending on whether the transcription matched with the audio. We found that all those segments for which the best lyrics transcription had less than 10 words were more likely to be incorrect matches, as shown in Figure 4.4. The segment transcriptions were 94.0% times incorrect (235 incorrect out of 250 total number of segments) when they contained less than 10 words, while they were 57.3% times incorrect (63 out of 110) when they contained more than or equal to 10 words. So we empirically set 10 words as the threshold for selecting

$e_{i_{min},j_{min}}$	1 (M=6)	2 (M=7)	3 (M=8)	4 (M=9)	5 (M=10)
Word 1	2	1	2	3	4
Word 2	2	3	4	5	6
.....	....	....	....	....	....
Word N	....	....	....	....	....

ASR transcript (ASR): *"the snow glows on the mountain"*
Published lyrics sub-sequence (Pub):
"the snow glows white on the mountain tonight not a footprint...."

M=6; Start Word #1
ASR: *the snow glows on the* ***mountain***
Pub: *the snow glows* ***white*** *on the* Error = 2

M=7; Start Word #1
ASR: *the snow glows on the mountain*
Pub: *the snow glows* ***white*** *on the mountain* Error = 1

Figure 4.2: Illustration of how the error matrix X is computed for an example where the ASR transcript is *"the snow glows on the mountain"*, and the published lyrics of this song has N words where a word sub-sequence is *"the snow glows white on the mountain tonight not a footprint to be seen..."*.

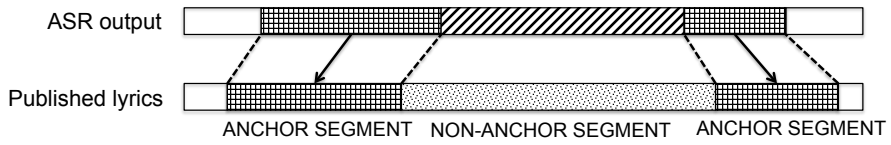


Figure 4.3: Anchor and non-anchor segments of a song based on sung-lyrics alignment algorithm. Anchor segments: ASR output and lyrics reliably match; Non-Anchor segments: ASR output and lyrics do not match.

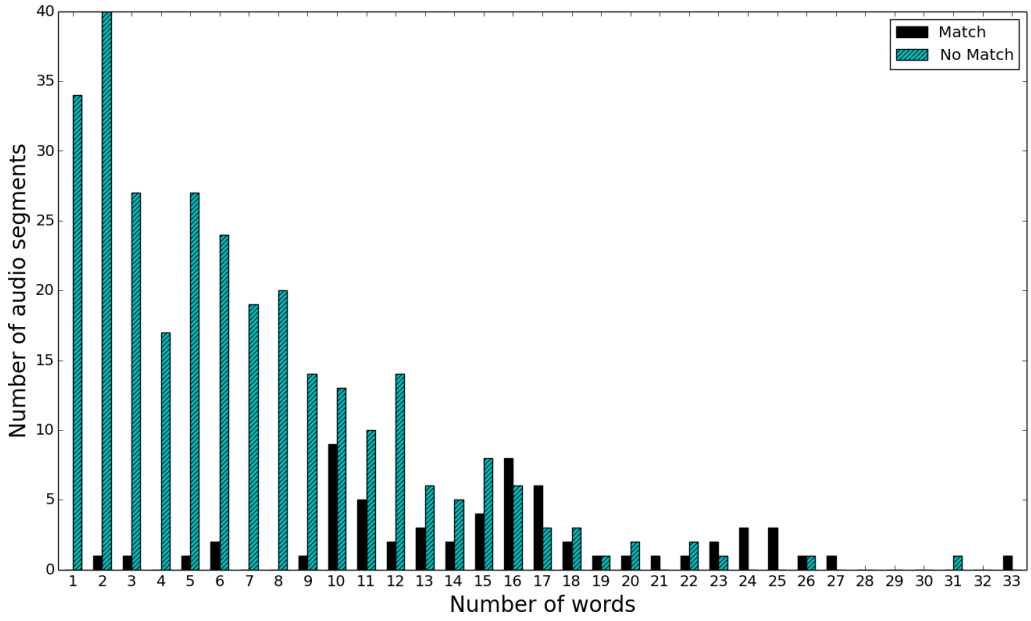


Figure 4.4: The number of audio segments with correct transcription (blue) or incorrect transcription (cyan) according to human judgment on y-axis versus the number of words in the transcription of an audio segment on x-axis. We set 10 words as the minimum threshold for a transcription to be valid for an approximately 10 seconds long segment.

reliable audio segments and transcriptions. By applying this constraint, we reject those audio segments that are noisy, or have wrongly pronounced words, or cause errors in transcription because of model mismatch, thus deriving a clean transcribed singing dataset.

The audio segments with reliable transcription are labeled as *Anchor segments*, and the audio segment(s) between two anchor segments that have unreliable transcription, are strung together and labeled as *Non-Anchor segments*, as illustrated in Figure 4.3.

One may argue that we could have used the error score e to evaluate the reliability of a segment. However, if the ASR output itself is wrong, then this lyrics-matching error score will be misleading. For example, if only 4 words get detected by the ASR, out of 12 words in the audio segment, and all the 4 words are correct according to the published lyrics, then e will be zero for this transcription, which is incorrect, and also undetectable. Thus we set a threshold on the number of detected words (i.e. 10 words) as a way to measure the reliability of the segment and its transcription.

4.3 Experiments and Results

In order to validate our hypothesis that our algorithm can retrieve good quality aligned transcriptions, we conducted three experiments: A) Human verification of the quality of the aligned transcriptions through a listening test, B) Semi-supervised adaptation of speech models to singing using our aligned sung-lyrics transcriptions for assessing the performance

of automatic lyrics recognition, and C) Second iteration of alignment, and re-training of acoustic models, to check for further improvement in lyrics recognition.

Our experiments are conducted on 6,000 audio recordings from the DAMP dataset that was used by Kruspe [Kru16a]. The list of recordings used by Kruspe is here [Mir17], however the training and test subsets are not clearly marked. Therefore we have defined our training and test datasets, and they are subsets of Kruspe's dataset, as discussed in the following subsections. This data set contains recordings of amateur singing of English language pop songs with no background music but different recording conditions, which were obtained from the Smule Sing! karaoke app. Each performance is labeled with metadata such as the gender of the singer, the region of origin, the song title, etc. We obtained the textual lyrics of the songs from Smule Sing! website [Smu]. Since the songs in DAMP dataset were sung on Smule Sing! Karaoke app that uses these lyrics, it is safe to assume that these were the intended lyrics.

4.3.1 Experiment 1: Human Verification of the Quality of the Aligned-Transcriptions

In this experiment, we evaluate the quality of our aligned transcriptions (*segment transcriptions*), by asking participants to listen to the audio segments and verify if the given transcription for the segment is correct or not. As opposed to word intelligibility evaluation tasks [CH08] where participants are asked to transcribe after listening to the stimuli once, in this task the participants were provided with the transcription and were free to listen to the audio as many times as they needed. Also the songs were popular English songs, that are less prone to perception errors [CSH15].

Dataset

By applying our lyrics transcription and alignment algorithm (see Section 4.2), we obtained 19,873 of anchor segments (∼58 hours) each ∼10 seconds long, out of which we asked humans to validate 5,400 (15 hours) anchor segment transcriptions through a listening test. The only criterion to qualify for the listening test was to be proficient in English language. 15 university graduate students were the human listeners. Every listener was given one hour of audio segments containing 360 anchor segments along with the obtained lyrics transcription. The task was to listen to each of the audio segments and compare the given transcription with the audio. If at least 90% of the words in the transcription match with that in the audio, then the audio segment was marked as correctly transcribed. If not, then it was marked as incorrectly transcribed.

Similarly, we also tested the quality of the non-anchor segments. Non-anchor segments could be of varying durations, greater than or equal to 10 seconds. We conducted the same human

Table 4.1: A summary of correct and error transcriptions by the proposed algorithm. Google ASR is used for singing transcription. Total # anchor segments = 5,400 (15 hours).

Segment Transcriptions	#	%	**Total %**
Correct transcriptions - fully matching ASR	838	15.52	**73.32**
Correct transcriptions - partially matching ASR	3,121	57.80	
Error transcriptions due to imperfect ASR or incorrect singing	1,441	26.68	**26.68**

validation task for 2 hours (1,262 segments) of the non-anchor segments of different durations.

Results and Discussion

There were two types of successful segment transcriptions, one was verified by humans as correct and also matched perfectly with the ASR output, and was labeled as *correct transcriptions fully matching ASR*. Another was verified as correct by humans but did not match with the ASR output due to ASR errors, but our algorithm could successfully retrieve the correct transcriptions, that we call *correct transcriptions partially matching ASR*. And the ones that were verified as wrong by humans are labeled as *error transcriptions due to imperfect ASR or incorrect singing*.

Anchor Segments: Table 4.1 shows the validation results for the anchor segments. We found that a total of 73.32% of the segments were correctly transcribed by our algorithm, where 57.80% of the segments were *partially matching ASR*. This means that our algorithm could successfully rectify many incorrect ASR transcriptions, which validates our hypothesis that the extra information provided by the published lyrics coupled with ASR decoding produces good aligned transcriptions. We also found that incorrect singing of lyrics and imperfect ASR output resulted in 26.68% erroneous transcriptions. A common error reported by the listeners was many missing words at the trailing end of the incorrectly aligned transcriptions, although the correct words were clearly audible, which is possibly a result of model mismatch between singing and speech.

Non-Anchor Segments: From the human validation of the non-anchor segments, we find that 62.07% of the total of 1,262 non-anchor segments transcriptions are correct. This suggests that these segments are less reliable. Moreover, these audio segments could be long in duration (even more than a minute) that would cause errors in the Viterbi alignments. Thus in the subsequent experiments, we only use the anchor segments.

4.3.2 Experiment 2: Lyrics Transcription with Singing-Adapted Acoustic Models

In this experiment, we use our automatically generated aligned-transcriptions of sung audio segments in a semi-supervised adaptation of the speech models for singing. We use these singing-adapted models in an open test to validate our hypothesis that better aligned transcriptions for training singing acoustic models will result in improvement in automatic lyrics recognition compared to the best known baseline from the literature.

Adaptation of speech models for singing was previously attempted by Mesaros et al. [MV08, MV10] who applied the speaker adaptation techniques to transform speech recognizer to singing voice. Speaker adaptation techniques use a small amount of target speaker samples to reduce the mismatch between the target speaker and the trained models. The same idea was used to adapt speech models to singing voice. To reduce the mismatch between singing and speech, they used constrained maximum likelihood linear regression (CMLLR) to compute a set of transformations that shifts the GMM means and variances of the speech models so that the resulting models are more likely to generate the singing data. In our work, we use CMLLR, also called feature-space maximum likelihood linear regression (fMLLR) [PS06] and our lyrics-aligned anchor segments to compute transformations for a semi-supervised adaptation of the speech models to singing. In MLLR, mean vectors and covariance matrices of the GMM-HMMs are affinely transformed. In fMLLR, the transformation applied to the variance must correspond to the transform applied to the mean, which effectively results in feature vector transformation [Gal98]. Adaptation can be done with the test dataset only, or the adaptation transformations can be applied at the time of training, called speaker adaptive training (SAT). Instead of applying the test singing data adaptation transforms to a speech acoustic model, they are applied to a model set trained using that adaptation scheme. Literature shows that the use of SAT with fMLLR transform requires minimum alteration to the standard code for training [Gal98], and thus is a popular tool for speaker adaptation that we have used for singing adaptation here.

In another configuration, we have also trained a DNN model [ea12] on top of the SAT model with the same set of training data. The DNN is a stack of non-linear hidden layers, whose output neurons are trained to estimate the posterior probability of the states, i.e. re-estimate the state parameters, given the MFCC vectors as the acoustic observation input vector. During DNN training, temporal splicing is applied on each frame with left and right context window of 4. The SAT+DNN model has 3 hidden layers and 2,976 output targets.

Dataset

The singing train set consists of 18,176 singing anchor segments from 2,395 singers while the singing test set consists of 1,697 singing anchor segments of 331 singers. The training set consists of both human verified and non-verified anchor segments, while the test set consists

of only those anchor segment transcriptions that are verified as correct by humans. All of these anchor segments (training and test) are of ∼10 seconds duration. There is no speaker overlap between the acoustic model training and test sets. A language model is obtained by interpolating a speech language model trained from Librispeech [PCPK15] text and a lyric language model trained from lyrics of the 301 songs of the DAMP dataset. The same language model is used in all the recognition experiments.

Results and Discussion

Table 4.2 reports the automatic lyrics recognition results on the singing test set using different acoustic models to observe the effect of adapting speech models for singing using our sung segments with aligned transcriptions.

The baseline speech acoustic model is a tri-phone HMM model trained on Librispeech corpus using MFCC features. Due to the mismatch between speech and singing acoustic characteristics, the WER and PER are high (Table 4.2 (1)). Adapting the baseline model with the singing test data results in a significant improvement in the error rates (Table 4.2 (2)). Speaker adaptive training (SAT) further improves the recognition accuracy (Table 4.2 (3)). With SAT+DNN training, the WER is reduced by about 7.7% relative to the SAT model (Table 4.2 (4)) and PER is 31.20%.

Mesaros et al. [MV08] reported the best PER to be 80% with speech models adapted to singing, while Kruspe [Kru16a] reported the best PER to be 80% and weighted PER to be 56% with pure singing phonetic models trained on a subset of the DAMP dataset. Compared to [Kru16a] and [MV08], our results show a significant improvement, which is attributed to three factors. First, leveraging on ASR along with the published lyrics as an external resource to validate and clean-up the transcriptions has led to better aligned transcriptions for training. Second, our automatic method for generating aligned transcriptions for singing provides us with a much larger training dataset. And third, the segment-wise alignment is more accurate than the whole-song forced-aligned with the speech acoustic models.

4.3.3 Experiment 3: Alignment with Singing-Adapted Acoustic Models and Re-training

We would like to test if the singing-adapted acoustic models can provide more number of reliably aligned transcriptions in a second round of alignment. Moreover whether re-training the models with this second round of transcriptions lead to better lyrics recognition.

Dataset

We used the singing-adapted models obtained in Experiment 2 to decode 12,162 segments, and then applied our lyrics-alignment algorithm to obtain new anchor and non-anchor

Table 4.2: The sung word and phone error rate (WER and PER) in the lyrics recognition experiments with the speech acoustic models (baseline) and the singing-adapted acoustic models, on 1,697 correctly transcribed test singing anchor segments.

Models Adapted by Singing Data	**%WER**	**%PER**
(1) Baseline (speech acoustic models)	72.08	57.52
(2) Adapted with test data	47.42	39.34
(3) Adapted (SAT) with training data	40.25	33.18
(4) Adapted (SAT+DNN) with training data	37.15	31.20
(5) Repeat (3) and (4) for 2nd round	**36.32**	**28.49**

Table 4.3: Comparing the number of anchor segments obtained from the proposed transcription and alignment algorithm using Google ASR and the singing-adapted models.

Model	**# anchor**	**total # segments**	**% anchor**
Google ASR	5,400	12,162	44.40
Adapted (DNN) with training data	11,184	12,162	**91.96**

segments. For comparison, we obtained the same from the Google ASR on the same dataset.

Results and Discussion

Table 4.3 shows that the number of anchor segments with the new models have increased from 44.40% with Google ASR to 91.96% with the singing-adapted models, which means that the number of reliable segment transcriptions have increased significantly. With the new anchor segments, we re-train our singing-adapted acoustic models. Table 4.2 (5) shows the free-decoding results after this second round of training. The WER and PER have dropped further to 36.32% and 28.49% respectively .

The results of this experiment are promising as they show iterative improvement in the quality of our alignment and transcription. This means that we can apply the following strategy: use only the reliably aligned segments from the Google ASR to adapt acoustic models for singing, and use these models to improve the quality of alignment and transcription, and then again use the reliable segments from the improved alignments for further adaptations.

4.4 Conclusions

In this chapter, we proposed an algorithm to automatically obtain time-aligned transcriptions for singing by using the imperfect transcriptions from the state-of-the-art ASR along with

the non-aligned published lyrics. Through a human listening test, we showed that the extra information provided by the published lyrics helps to correct many incorrect ASR transcriptions. Furthermore, using the time-aligned lyrics transcriptions for iterative semi-supervised adaptation of speech acoustic models for singing shows significant improvement in automatic lyrics transcription performance. Thus our strategy to obtain time-aligned transcriptions for large-scale singing dataset is useful to train improved acoustic models for singing.

Our contribution provides an automatic way to obtain reliable lyrics transcription for singing, that results in an annotated singing dataset. Lack of such datasets has been a bottleneck in the field of singing voice research in MIR. This will not only generate lyrics transcription and alignment for karaoke and subtitling applications, but also provide reliable data to improve acoustic models for singing, thus widening the scope of research in MIR.

Chapter 5

Applications of phonetic modeling of singing: (1) Pronunciation evaluation in singing voice

Automatic pronunciation evaluation of singing is an essential technology in a wide-range of applications. First, lyrics play an important role in music, serving as a cue for detecting a song's identity, or its mood or genre [AP06, BAB+11]. Therefore, correctly pronouncing the lyrics of a song becomes an important component of a singing performance. In addition, singing is shown to be helpful in improving pronunciation in foreign language learning classes [NS11, GRS15b] (Section 1.3.2). Furthermore, music and speech therapists apply a therapeutic process called Melodic Intonation Therapy (MIT) for speech rehabilitation of patients with speech disorders, such as non-fluent aphasia [NZMS09].

Computer-aided pronunciation training (CAPT) for speech has been an active area of research [NFDW00, ZSG+05]. But automatic pronunciation evaluation of singing is still a relatively unexplored area. The state-of-the-art ASR technology cannot be directly applied for singing pronunciation evaluation because of the mismatch between speech and singing. The acoustic characteristics of singing and speech differ in many ways, such as pitch range, vibrato, and phoneme durations [FG12, LCB99]. Thus to build a pronunciation evaluation algorithm for singing, the ASR needs to be adapted to singing voice. Moreover, the applicability of the traditional speech pronunciation scoring methods for evaluating singing pronunciation needs to be investigated.

In this work, we propose a novel singing-specific lexicon modification method to overcome the vowel duration differences between singing and speech. We hypothesize that this lexicon modification method for obtaining singing-adapted models leads to better word boundary alignment, which is necessary for a reliable scoring. Next, we investigate methods for scoring pronunciation of sung-utterances at word- and song-levels. We believe that incorporating singing-specific characteristics in scoring would yield improved results. Finally, we validate our scores with human judgments. We also verify that crowd-sourcing platforms can be used to obtain reliable human scores for singing evaluation. We report the encouraging experimental results.

Our contributions:

- We design a strategy of duration-based lexicon modification to improve the performance of the singing-adapted speech acoustic models
- We show that incorporating these modified acoustic models results in better word boundary alignment, recognition, and pronunciation assessment in singing voice

5.1 Related Work

Phonetic errors in non-native (L2) speech are attributed to the influence of the native language (L1) that results in phone substitutions, deletions or insertions [CL16]. L1 influence also results in phonetic errors in singing of non-native speakers, as reported in [GGRW17]. In karaoke-singing, incorrectly pronounced words often occur due to unfamiliar lyrics or song, that results in substitution, deletion, and insertion of words. In this study, we focus on detecting word pronunciation errors which may be due to the influence of L1 or unfamiliarity of the lyrics of the song.

Only a few studies have addressed the problem of evaluating pronunciation of singing. Jha et al. [JR12] attempted to develop a method for evaluating pronunciation of singing based on vowels. They compared MFCC and pitch-based features to classify sung vowels, and found that there is no significant difference in performance between the two feature sets. However, their work involved manually extracting vowel segments, and also did not extend to consonants. Recently, we studied the difference in pronunciation between speech and singing in South-East Asian English accents, and found that in singing vocals, the consonant errors are more prominent than the vowel errors [GGRW17]. We also incorporated the common pronunciation error patterns for a given L1-L2 pair in a dictionary to automatically detect the mispronounced words. But this work is limited by the need of developing an L1-L2 pair specific dictionary, hence cannot be easily generalized.

In traditional CAPT systems, an L1-independent method of scoring a phoneme is the Goodness of Pronunciation (GOP) score, which is the difference between the log-likelihood score from forced alignment and that from open phone loop decoding, where the phone boundaries are obtained from forced-alignment [WY00, KFN97]. This is a template (or reference) independent method of scoring. Another method of scoring is template dependent [ZG09, LG12, LZG13], where deep neural net (DNN) phone posteriorgrams (PPG) are used in dynamic time warping (DTW) between a reference utterance and a test utterance to detect word-level mispronunciations. In this work, we investigate how such methods work for singing pronunciation evaluation.

As noted in earlier chapters, a large corpus of solo-singing karaoke data DAMP [Smu] was recently made available for research purposes. However, annotating such data for qualitative tasks such as singing quality assessment or pronunciation quality evaluation, is still a challenging task. We note that crowd-sourcing platforms have been used for

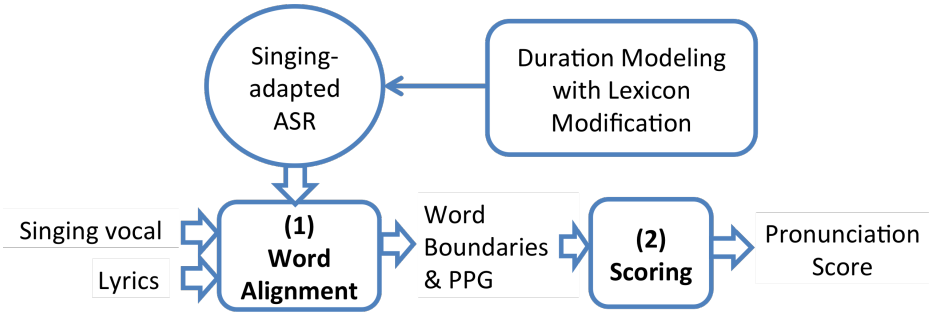


Figure 5.1: Two-stage approach for pronunciation evaluation: word alignment, and scoring (PPG: phonetic posteriorgram).

labor-intensive tasks such as speech transcription [MBR10], speech quality assessment tasks [RFZS11, NPW+15, PBTO13], and speech pronunciation quality assessments [WM12]. Researchers have found methods to overcome the noisy nature of the data from such platforms, using gold standard questions, and trapping questions [NPW+15]. Encouraged by the findings, here we would like to study how to obtain reliable human judgments of singing pronunciation from the crowd-sourcing platform. We validate the crowd-sourced data against a laboratory-controlled listening experiment data.

With the scarcity of large-scale lyrics-aligned singing data, acoustic models for singing pronunciation evaluation can be built by adapting the speech phonetic models to singing. In our previous work (Chapter 4), we used fMLLR (feature-space maximum likelihood linear regression) and lyrics-aligned transcriptions of a subset of the DAMP dataset for semi-supervised speaker adaptive training (SAT) of the speech models to singing voice. In this work, we investigate the performance of word-alignment with our proposed duration-based lexicon modification method along with these singing-adapted models.

5.2 Singing Pronunciation Evaluation

Speech and singing have many similarities because they share the underlying physiological mechanisms for production. This involves similar articulatory movements to produce words in speech and lyrics in singing [ZB12, ZRS14], thus resulting in similar spectral characteristics for the place and manner of articulation of phonemes. Therefore, we adopt the speech pronunciation evaluation methodology for evaluating pronunciation in singing. We evaluate singing pronunciation in a two-stage approach: word alignment, and scoring, as shown in Figure 5.1. We apply the singing-adapted speech acoustic models to force-align the lyrics to the test singing utterance, which results in automatic word-boundary segmentation; and then, we use phonetic posteriorgram (PPG)-based template-dependent and template-independent methods to evaluate the segments between the obtained word boundaries. In the following sub-sections, we discuss the two stages in detail.

5.2.1 Word Alignment

Evaluation of pronunciation is based on phonetic segments, therefore accuracy of alignment is important as it is going to affect the scoring accuracy. Force-aligning the lyrical words to singing with a speech acoustic model does not provide good alignment due to the mismatch between speech and singing signals. One main difference between singing and speech is the duration of the vowels. In singing, the vowels are stretched in time as dictated by the musical score. Previously, musical score-informed duration modeling of vowels has been used in speech-to-singing voice conversion [SGUA07], singing-to-speech conversion [NDC+10], and singing syllable segmentation [PGS17]. Rong Gong et al. [GCOC15] have incorporated pitch and vowel spectral distribution templates to align audio to the musical score. However in karaoke singing, many amateur singers may not be able to follow the musical scores correctly. Therefore a score-informed method of lyrics-to-audio alignment may not be accurate.

In this work, we propose a novel singing-specific lexicon modification strategy to improve the forced-alignment word boundaries in singing.

Lexicon Modification

The vowels in singing could be longer in duration than spoken vowels, because they are dictated by the melodic and rhythmic attributes of the song. Longer duration of vowels can be viewed as a type of pronunciation variation. One method of evaluating pronunciation in speech, called the *extended recognition network (ERN)* [HLQM09, CL16], enhances the lexicon with the possible and expected pronunciation error patterns in the specific L1-L2 pair, such that the ASR selects the closest matching variant at the time of forced-alignment. In this work, we propose to modify the lexicon to model the duration dynamics of vowels in singing. We modify the lexicon for singing such that there are multiple pronunciation variants of every word that represent different vowel durations. We adopt the strategy of optional repetition (up to 4 times, set empirically) of the vowels so as to allow longer duration of the vowels. For example, the word *sleep* will have the following lexicon variants: [S L IY IY IY IY P], [S L IY IY IY P], [S L IY IY P], [S L IY P]. Such variants are created with respect to every vowel in the word. We expect that this method will result in improvement in force-aligned boundaries and thus the pronunciation scores.

5.2.2 Scoring

We evaluate how close an uttered segment is to an expected phone, while considering the differences between speech and singing phonemes. We define the template independent [WY00, KFN97] and dependent [ZG09, LG12, LZG13] scores, called Pronunciation Evaluation Metric (PEM) scores, based on the Phonetic Posteriorgram (PPG). PPG contains the normalized posterior probability of every phone per frame, obtained from decoding a sung utterance with the singing-adapted acoustic models, as illustrated in Figure 5.2.

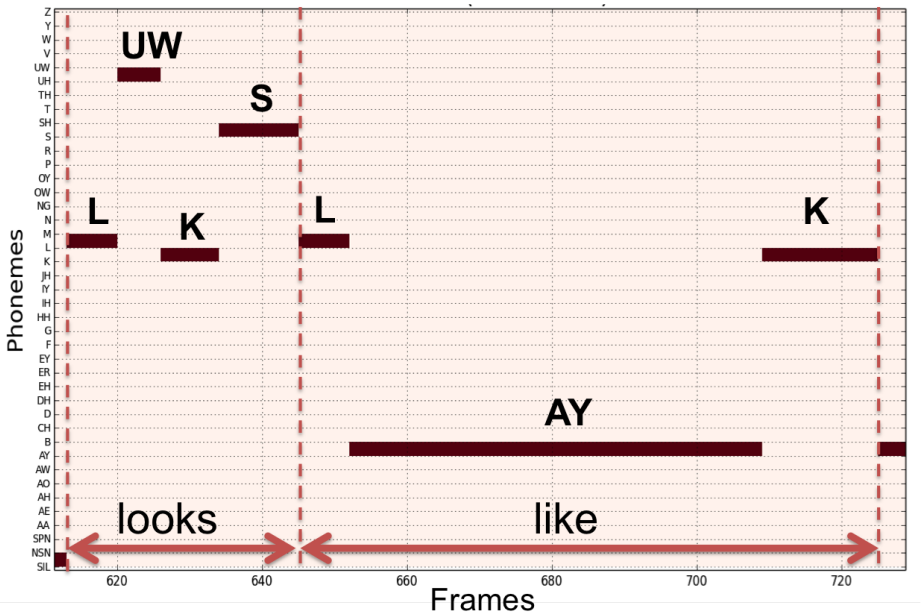


Figure 5.2: Phonetic Posteriorgram (PPG)

Template Independent PEM

Template independent PEM score (PEM_{ind}) indicates how close the pronunciation of a test sung utterance is to the target lyrics, similar to the GOP scores of the CAPT systems [WY00, KFN97]. PEM_{ind} is defined as the ratio of the probability of the target phoneme to the sum of probabilities of the rest of the phonemes, averaged over all the frames within the phoneme:

$$PEM_{ind} = \frac{1}{N}\sum_{i=1}^{N}\frac{P_i(T_p)}{\sum_{\forall k \neq p} P_i(T_k)} \tag{5.1}$$

where $P_i(T_p)$ is posterior probability of the target phone T_p from PPG for a frame i. And N is the number of frames within the phone boundaries obtained from forced-alignment with the target transcription. A high PEM_{ind} score means the uttered phone is close to the target phone.

Template Dependent PEM

Template dependent PEM score (PEM_{dep}) indicates how close the pronunciation of the test sung utterance is to a reference sung utterance [ZG09, LG12, LZG13]. It is computed as the dot product of the reference PPG vector P_r and the test PPG vector P_t.

$$PEM_{dep} = -\log(P_r \cdot P_t) \tag{5.2}$$

If the reference and the test probabilities match, then the dot product of the vectors P_r and P_t will be high, therefore, PEM will be small.

During the transition period between phones, the phone identity is ambiguous, resulting in unreliable PPG values at the phone boundaries. This ambiguity is even more in singing compared to speech, because all phones are not always prominently articulated in singing,

such as the word-end consonants [GGRW17], causing unclear boundaries. So to avoid the unreliable boundary values, we compute the phone-level scores by using only the center frames. Empirically, we found that 58% of the center frames results in the lowest detection error rate.

In singing, the vowels are often stretched in time, thus occupying a larger proportion of the word than that in speech [DFL+13]. We consider this characteristic of singing in computing the word-level scores, by either giving equal weights to all the phone-level scores of the word, or by giving weights to them according to the percentage of frames occupied by the phone in the word. We observe that frame-weighting the phone-scores shows higher correlation with human judgment compared to equally weighting them, which is intuitively justified as the long duration vowels have more time to make an impression on the listener. We also found that the PPG values for the short duration consonants tend to have more errors than the long duration vowel segments. Thus frame-weighting also reduces the sensitivity of the score to PPG errors.

5.3 Experiment

We now conduct experiments to validate the proposed pronunciation evaluation strategy for singing. We test our hypotheses that lexicon modification leads to better word boundary alignment, investigate the performance of the speech pronunciation evaluation methods for singing with this modification, and validate with human scores collected via crowd-sourcing platform.

5.3.1 Dataset

From the DAMP dataset, we selected 24 singers (13 female, 11 male) each singing one of 6 unique English popular songs: *Let it go, Lovefool, I dreamed a dream, When I was your man, and Stay.* According to the metadata provided in DAMP, the singers belonged to different language speaking zones of the world: 4 from JA (Japanese), 1 from ZH (Chinese), 2 from ES (Spanish), 1 from FR (French), and the rest from EN (English). Pronunciation errors were caused by L1-influence, or unfamiliar lyrics, or both, in both native English and non-native English singer renditions, but more in non-native singers. For more details, please refer to our published dataset[1].

Since the songs were sung on Smule's karaoke app Sing!, all the renditions of each song were time-aligned. Forced-alignment using ASR is known to work well for short utterances. So we split the renditions into shorter utterances of 5-10 seconds by marking the line boundaries

[1]Dataset available here: https://github.com/chitralekha18/Dataset-for-pronunciation-evaluation-in-singing.git

of one good pronunciation rendition of each song, and aligning the rest with DTW. This resulted in a total of 666 short sung utterances.

5.3.2 Human Annotations

We obtained three types of human annotations for a subset or the whole of this dataset to validate our automated word alignment and pronunciation scoring: word boundary time markings, pronunciation judgments at song-level, and word-level.

Word boundary markings

For the word-alignment validation experiment (Section 5.3.3), we manually marked the word boundaries of 100 utterances from the well-sung, i.e. correct pronunciation renditions of 5 singers who belonged to the EN zone (20 utterances per singer).

Word-level pronunciation judgments

For validating word-level automatic scoring, we asked two university students fluent in English to listen to 10 sung utterances from 10 singers, (5 from EN zone, and 5 from non-EN zone) i.e. 100 utterances, and marked the words in the lyrics that are mispronounced, i.e. substituted, deleted, or new words inserted. This is a binary judgment per word where the marking '1' for a word is to indicate incorrect pronunciation, and '0' is for correct pronunciation. In this way, we obtained word-level ground-truth pronunciation judgments for 990 words.

For the template-dependent method of scoring, we obtained the word-level evaluation for 10 utterances from 5 more EN zone singers with good pronunciation, who were considered as the reference templates for this experiment.

Song-level pronunciation judgments

To validate song-level pronunciation scores, we wanted to collect reliable human song-level pronunciation judgments in a scalable way, by leveraging on a crowd-sourcing platform, Amazon mechanical turk (MTurk). A method of proving reliability of the MTurk data is to observe the correlation between the MTurk data and that from a laboratory-controlled experiment [NPW+15].

MTurk data reliability test: In Chapter 2, our task was to build an algorithm for automatic singing quality evaluation. We asked 5 professionally trained musicians to give singing quality assessment for various singing parameters including pronunciation for 20 singers on a likert scale of 5. So we obtained lab-controlled average pronunciation scores from this experiment.

Here, we conducted the same experiment on MTurk, where Human Intelligence Tasks (HIT) consisted of a song audio file, along with its lyrics, followed by a questionnaire. The

questionnaire now included five additional questions to inquire about the judge's music experience and English speaking proficiency. The music experience related questions asked about how many years of vocal training/musical instrument training/stage performance experience have they got, a short description about their music experience, and asking them to transcribe randomly chosen four musical notes. The English speaking proficiency questions asked if they were native English speakers, and to rate their own English speaking proficiency. Each of the 20 singers' songs were rated by at least 7 human judges.

A human rating was rejected if two out of the three music-related questions showed that they did not have any music experience, and if they were non-native English speakers with English speaking fluency below 4. We also rejected a judgment that marked the exact same rating for all the questions. This shows that the rater was not serious about the task. After this questionnaire-based data clean-up procedure, we had at least 5 ratings per song, from which we computed the average ratings for each of the singing parameters. The average Pearson's correlation between these ratings and that from the controlled-lab experiment done by professional musician was 0.86. Thus the questionnaire-based data clean-up results in high correlation between the lab-controlled experiment and the MTurk experiment validating our hypothesis that we can get reliable subjective ratings for singing evaluation parameters, including pronunciation, from crowd-sourcing platforms.

We then implemented the same MTurk experiment for our new dataset of 24 songs (see Section 5.3.1). We have focused only on the pronunciation ratings in this study. After the questionnaire-based data clean-up, the average inter-rating correlation for pronunciation between the 5 selected judges is 0.60. The average of the 5 ratings for each song is considered as the human ground-truth pronunciation score of the song.

5.3.3 Singing Pronunciation Evaluation Validation

In Chapter 4, we adapted baseline speech acoustic models (tri-phone HMM model trained on Librispeech corpus [PCPK15] using MFCC features) to singing with sung utterances from the DAMP dataset that resulted in WER of 36.32% from SAT+DNN models in an open loop decoding experiment. To account for the long duration vowels and obtain better singing-adapted models for alignment, we repeat the same experiment by using the singing-specific modified lexicon discussed in Section 5.2.1, which reduces the WER to 29.65% with SAT+DNN models. We also verified that our lexicon modification helps in modeling the long duration vowels (Table 5.1), i.e. longer duration vowels are modeled by more vowel repetitions in the lexicon. We use these singing-adapted models in the two-stage approach of pronunciation evaluation, as discussed in Section 5.2.

Word Alignment Validation

We validate the quality of word alignment from the forced-alignment of the singing-adapted SAT models with the sung utterance by comparing with the human annotations for word

Table 5.1: Effect of lexicon modification: # of vowels modeled by the different optional repetition variants in the lexicon, and the avg. duration of those vowels. (across the 666 sung utterances)

repetition times in lexicon→	**0**	**1**	**2**	**3**
# of vowels	5804	3886	1299	402
avg. dur. of vowels (seconds)	0.218	0.380	0.674	1.518

Table 5.2: Word alignment validation of well-sung renditions: the number of words within a range of absolute deviation of the automatic boundaries from the ground-truth. Total number of words=896. LEX: lexicon modification.

Singing-adapted SAT Models	**<20ms**	**20-50ms**	**50-100ms**	**100-200ms**	**>200ms**
w/o LEX	635	115	82	24	40
LEX	748	74	33	9	32

boundaries. We also compare the word alignment performance of the lexicon-modified singing-adapted models with the one with the baseline singing-adapted models in Chapter 4. We expect our model to detect the word boundaries accurately in the sung utterances with good pronunciation, and that the word boundary detection should improve with the lexicon-modification. We compute the sum of absolute deviation of the start and the end boundaries from the ground-truth markings for every word as a measure of boundary deviation. Table 5.2 shows the number of words within different ranges of boundary deviations using SAT models with and without the lexicon modification, for sung utterances with good pronunciation. We see that lexicon modification improves the boundary alignment performance from 83.7% to 91.7% within 50ms of absolute boundary deviation, which is an 8% improvement. An example of alignment is shown in Figure 5.3. The last tier in the figure shows that the transcription aligned with lexicon modification is "G AA AA D", which is clearly an improvement over the one without the lexicon modification. This verifies our hypothesis that duration-based lexicon modification for obtaining singing-adapted acoustic models leads to better word alignment.

Scoring Validation

We performed two experiments for validating our pronunciation scores for singing: word-level and song-level. In word-level evaluation, we compared the PPG-based template dependent and independent methods of scoring the aligned words, with the human judgments (see Section 5.2.2 and Section 5.3.2). In song-level evaluation, we compared the overall pronunciation score for a song rendition, with the human judgments obtained from MTurk (Section 5.3.2).

(1) Word-level scoring validation:

We wanted to see whether our scoring algorithms are able to correctly detect mispronounced words in a sung utterance. Table 5.3 compares the template dependent and independent methods of scoring with human word-level scores. It also shows the effect of lexicon

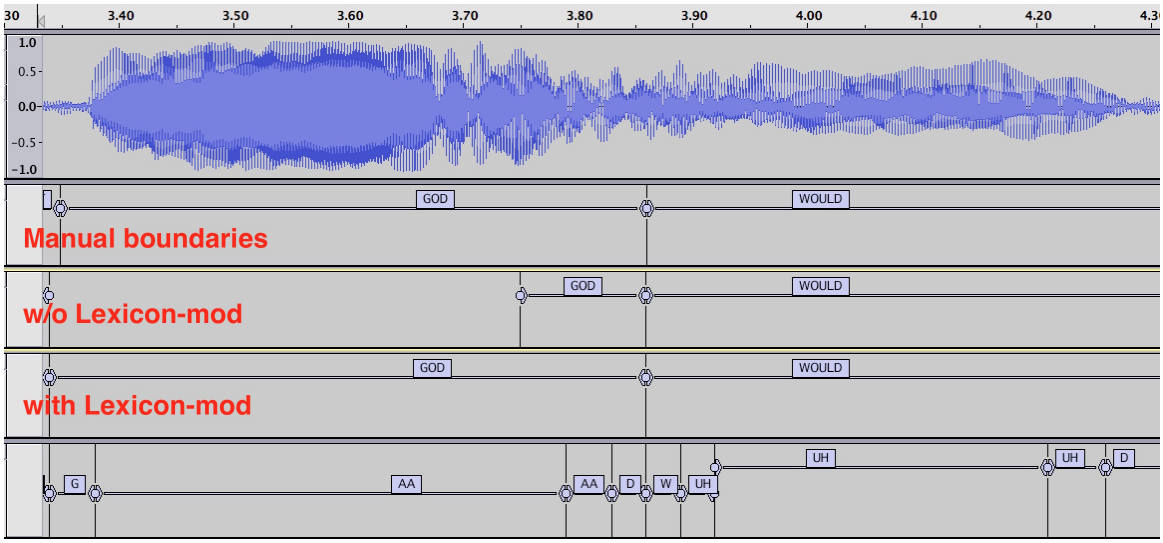


Figure 5.3: Automatic word alignment example for the sung-utterance "god would" with SAT models trained without and with lexicon modification. Tier 1: Singing waveform; Tier 2: Manual or ground-truth word boundaries; Tier 3: Automatic word boundaries with SAT models without lexicon modification; Tier 4: Automatic word boundaries with SAT models with lexicon modification; Tier 5: Automatic phone boundaries with SAT models with lexicon modification.

modification on scoring. To evaluate the performance of the methods, we compute the metrics precision (Pre), recall (Rec), and F-score (F), where TP (True Positive) is the # of mispronounced words detected as mispronounced, TN (True Negative) is the # of correctly pronounced words detected as correctly pronounced, FP (False Positive) is the # of correctly pronounced words detected as mispronounced, and FN (False Negative) is the # of mispronounced words detected as correctly pronounced.

Template independent outperforms template-dependent method, with an equal error rate (EER) of 0.28 and accuracy of 0.72, compared to 0.47 and 0.52 respectively from template-dependent method. The main reason for high error rate is the high false positives in both the methods, but more in the template-dependent method. The template-independent method depends on the test utterance PPG and the lyrics, whereas the template-dependent method relies on the PPG of the test as well as the reference utterances. However the imperfect singing acoustic models that estimate the PPGs may cause errors in the PPGs. This results in false positives in both the methods, but more so in the template-dependent method because of errors in the reference template PPG.

Also, with lexicon modification, template independent method performs better than without the modification. Mc Nemar's test P [GC89] is 0.038, implying the observed difference between the performance of the two algorithms would arise by chance on only 3.8% of occasions. So there is evidence of a statistically significant improvement in pronunciation

Table 5.3: Word-level scoring: Performance of automatic mispronunciation detection for singing and speech. P: Precision = TP/(TP+FP); R: Recall = TP/(TP+FN); F: F-score = 2.P.R/(P+R); FPR: False Positive Rate = FP/(FP+TN); FNR: False Negative Rate = FN/(FN+TP); Total number of words=990. LEX: lexicon modification.

Method: Template-	TP	TN	FP	FN	Pre	Rec	F	Acc	FPR	FNR
Dependent (LEX)	71	445	410	64	0.15	0.53	0.23	0.52	0.48	0.47
Independent (LEX)	97	613	242	38	0.29	0.72	0.41	**0.72**	**0.28**	**0.28**
Independent (w/o LEX)	95	598	257	40	0.27	0.70	0.39	0.70	0.30	0.30

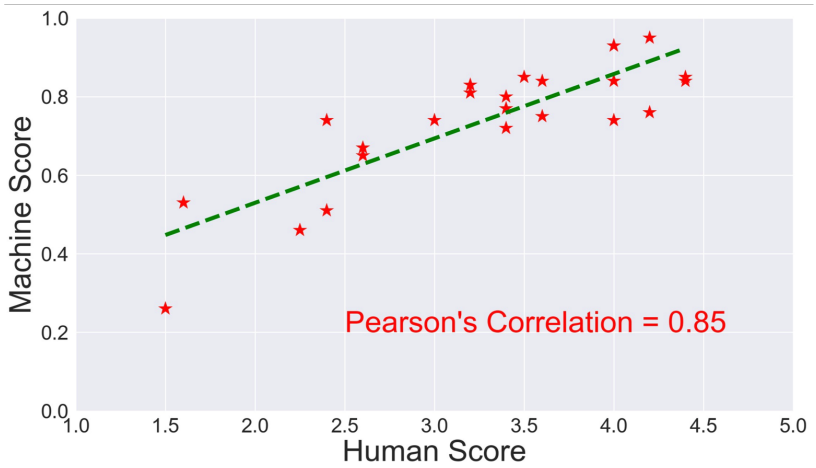


Figure 5.4: Song-level score comparison: machine vs. humans. Pearson's correlation is 0.85.

evaluation performance based on the lexicon-modification method.

This experiment verified our hypothesis that the speech pronunciation evaluation methods are applicable for singing with singing-specific modifications.

(2) Song-level scoring validation:

We wanted to see if the word-level scores across all the utterances of a song can give an overall song-level pronunciation score that correlates with the human judgments. We computed the percentage of words detected as incorrectly pronounced ($\%error$) by template-independent method across all the utterances of a song by a singer, thus $1 - \%error$ is the measure for song-level pronunciation accuracy of a singer. The Pearson's correlation between the automatic and the average human annotated song-level pronunciation scores for the 24 songs is 0.85 (Figure 5.4). This verifies that the evaluation of pronunciation of singing based on template-independent pronunciation evaluation method gives reliable song-level assessment. We also found that computing the song-level scores with only the well-aligned words (aligned within 50 ms of the ground-truth boundaries) results in an even better correlation of 0.91 with the human scores. This means that better alignment leads to better evaluation performance.

5.4 Conclusions

We developed a strategy to compute reliable pronunciation evaluation scores for singing. We showed that duration-based lexicon modification for singing acoustic model adaptation results in improvement in word alignment as well as scoring accuracy. We also found that the template independent method of scoring with singing-specific modifications shows high correlation with human judgments both at word- and song-levels. Additionally, we verified that the subjective pronunciation scores for singing, that is needed for algorithm validation, can be reliably obtained through crowd-sourcing. Future work will involve analyzing the relationships between singer geographical origin, song difficulty level, and evaluation accuracy.

Chapter 6

Applications of phonetic modeling of singing: (2) Lyrics-to-audio alignment

Automatic lyrics-to-audio alignment has various applications such as the automatic generation of karaoke scores, song-browsing by lyrics, and the generation of audio thumbnails. Given the audio signal of singing voice and the corresponding textual lyrics as input data, lyrics-to-audio alignment can be defined as a problem of estimating the temporal relationship between them.

Many studies on lyrics-to-audio alignment exploit the knowledge of the musical structure of the song to align the lyrics [LC08, WKN+04]. One of the earliest studies [WKN+04] (LyricAlly) uses the structural information of popular songs to align the chorus and the verse sections of lyrics to the music audio. Mauch et al. [MFG12] incorporated time-aligned chord information along with the lyrics to improve the alignment. The limitation of these methods is that the music structure may vary with genre, and they need manually transcribed chord labels with reliable temporal information corresponding to the lyrics.

In automatic speech recognition (ASR) tasks, word or phone level segmentation is obtained by forced-aligning the transcription to the speech using acoustic models trained with speech data. The same idea has been applied to align lyrics to music audio [FGO+06, FGOO11, MFG12, MEG14, MV10]. In [FGO+06], singing vocal is separated from polyphonic music, and maximum likelihood linear regression (MLLR) is used for adapting the speech phone models to the singing vocal. These adapted phone models achieved a low word alignment accuracy of 46.4%. In another work, Mesaros et al. [MV10] used 49 fragments of songs, 20-30 seconds long, along with their manually acquired transcriptions to adapt Gaussian mixture model-hidden Markov model (GMM-HMM) speech models for singing. Using these singing-adapted speech models, they reported a phoneme error rate of 80%. These works have provided a direction for solving the problem of lyrics alignment in music, but they suffer from manual post-processing and the models are based on a small number of annotated singing samples.

A major problem in building a lyrics alignment system is the lack of availability of annotated dataset. In Chapter 4, we designed an algorithm to automatically obtain lyrics annotations for solo-singing data by leveraging on speech models to force-align ∼50 hours of solo-singing

audio from the karaoke singing dataset DAMP [Smu] with the lyrics. We iteratively adapted the speech models to singing voice, while automatically refining the training data by removing the bad quality audio and lyrics, and improving the lyrics annotations for the songs. Kruspe [Kru16a] and Dzhambazov [DS15] also attempted the alignment task in MIREX 2017, but did not account for the bad audio recordings and refinement of the training data. Our singing-adapted models showed 36% word error rate (WER) in a free-decoding experiment on solo-singing (Chapter 4). However, these models are not expected to perform well in the presence of background accompaniments.

Recently, in MIREX 2018 [Mir18], the systems submitted by KKBOX Inc. achieved a mean average absolute error (ASE) of 2.7 seconds for Hansen's polyphonic music dataset [Han12] and 4.12 seconds for Mauch's polyphonic dataset [MFG12]. They have used 7,300 annotated English songs (more than 300 hours) from KKBOX's music library to train HMM based models. As a pre-processing step, they segmented the audio files according to the position of blank lines in lyrics and performed vocal detection. Although this work achieved a good performance, they have used a large amount of annotated polyphonic data to train the models. Such copyrighted large polyphonic music audio dataset is not available to the research community. Moreover, obtaining human annotations of polyphonic music is a tedious task and extending such a system to an under-resource language will be challenging.

In this work, we propose to use singing-adapted speech acoustic models trained on a relatively small solo-singing dataset in conjunction with audio source separation to obtain word-level lyrics alignment boundaries for polyphonic audio. The acoustic models are trained to handle the differences between speech and singing, such as long duration of vowels, and the pitch dynamics. Furthermore, we incorporate audio source separation as a pre-processing step to extract the singing vocals, and conduct a comparative study of the effect of different audio source separation methods on the performance of our lyrics-to-audio alignment system. We also study the impact of reliable vocal detection in the task of lyrics alignment.

Our contribution:

- We demonstrate that with the help of our singing-adapted speech acoustic models trained on a small monophonic singing dataset, along with a reliable audio-source separation, we can develop a high performing lyrics-to-audio alignment system for polyphonic audio

The rest of the chapter is organized as follows: we describe the framework of the proposed system in Section 6.1. The experiments performed to establish our framework is discussed in detail in Section 6.2. We have summarized our results in Section 6.3.

6.1 Framework for lyrics-to-audio alignment

In this work, we would like to build a framework for automatically aligning lyrics to the polyphonic music audio, as shown in Figure 6.1. The idea is to use our trained solo-singing-adapted speech acoustic models to force-align the lyrics with the music audio. But there is a training and test data mismatch. We can bridge this gap in two ways: (a) by improving the acoustic models further by training on polyphonic data, and (b) by making the test data closer to the trained solo-singing acoustic models. As a large polyphonic music dataset with aligned lyrics is not publicly available for training, in this work we consider the latter. Our framework consists of three main components: singing-adapted speech acoustic models, singing vocal separation, and vocal end-point detection.

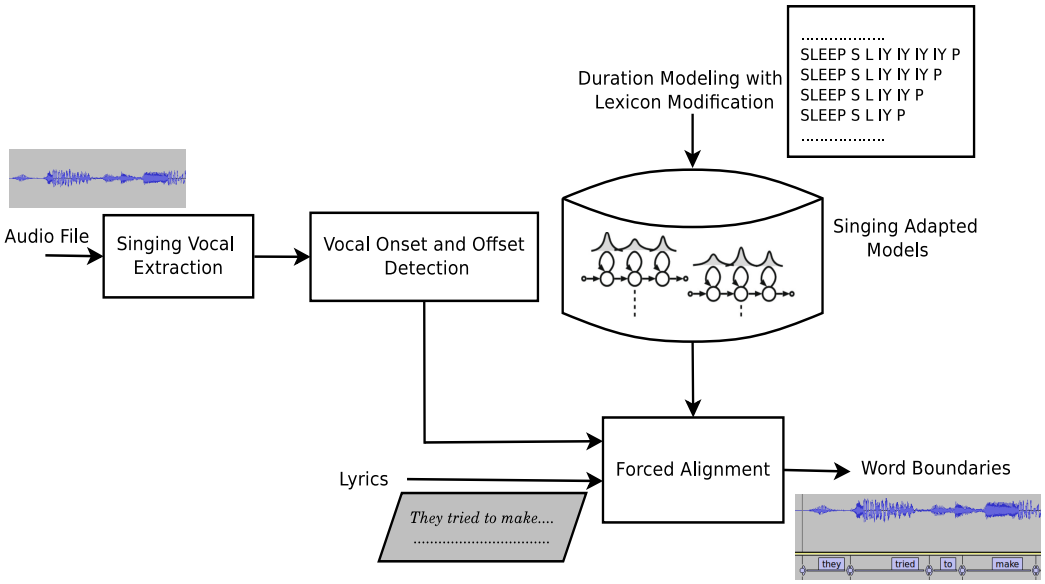


Figure 6.1: Framework for automatic lyrics-to-audio alignment.

6.1.1 Singing-adapted acoustic models

To reduce the mismatch between singing and speech, speech acoustic models are adapted to singing voice using speaker adaptation methods [MV10]. In Chapter 4, we applied a semi-supervised speaker adaptive training (SAT) method, with lyrics-aligned solo-singing dataset to adapt speech models to singing voice. A feature-space maximum likelihood linear regression (fMLLR) was applied for the adaptation of the speech models. One major difference between speech and singing voice is the duration of vowels. The vowels in singing could be longer in duration than spoken vowels. Therefore we introduce pronunciation variants in the lexicon to model the longer duration of vowels in singing. This modification reduces the WER from 36% to 29.65% (Chapter 5). However, these singing-adapted models were not evaluated for lyrics-to-audio alignment. We expect that introducing these modified models will contribute to achieve good performance in this task. However, these models may

not be well-suited for polyphonic music because the presence of background music introduces noisy components that results in mismatch between our models and the test data.

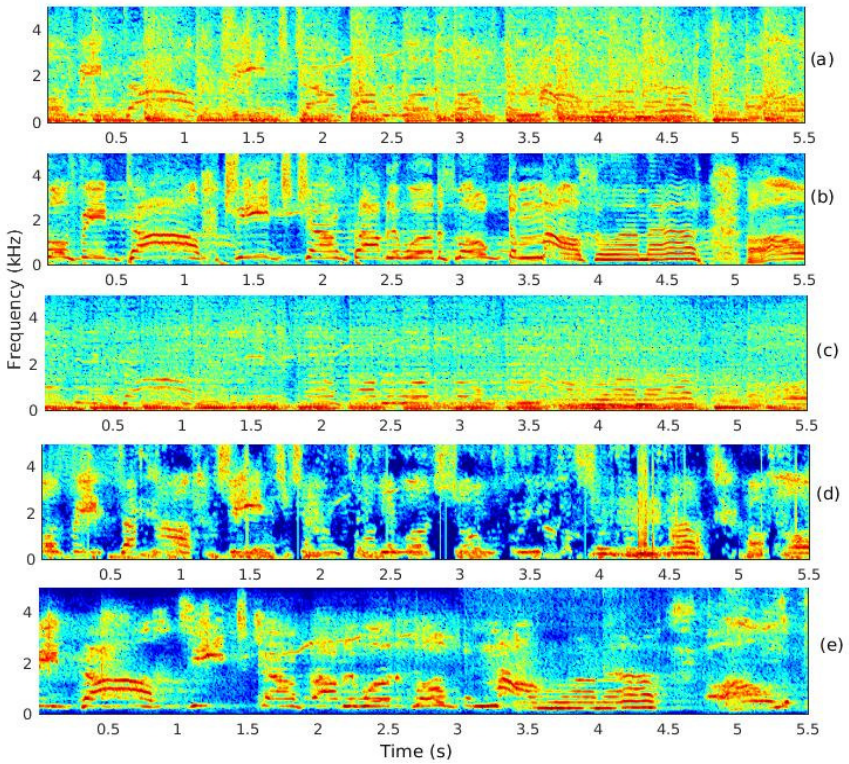


Figure 6.2: Comparison of spectrograms for different audio source separation methods for "*this afternoon*" song from Hansen's dataset, (a) original mixed audio, (b) original clean audio, extracted vocal using (c) harmonic/percussive, (d) CNN based, (e) U-Net based audio source separation method.

6.1.2 Singing vocal separation

To accommodate the deviation between the trained models and the test data, we incorporate a source separation module to extract the singing vocals from polyphonic songs. We study the effect of three different audio source separation methods on our lyrics alignment algorithm: harmonic/percussive, convolutional neural network (CNN) based, and U-Net based. Percussion component in the background accompaniment introduces vertical lines in the spectrogram, which makes it noisy. Therefore, we first attempted to remove these using the traditional harmonic/percussive method [Fit10], which is reported to be simple and effective. This method uses median filters individually in the horizontal and vertical directions to separate the harmonic and the percussive events. This separation method is integrated in the widely used audio and music analysis library Librosa [MRL+15].

As an alternate to the traditional methods and owing to the advantages and success of CNN, we are also interested to apply the audio source separation method proposed in [CMJG17], which achieves same performance as that of multilayer perceptron based audio source separation with less time complexity and compact representation. In this case we have used

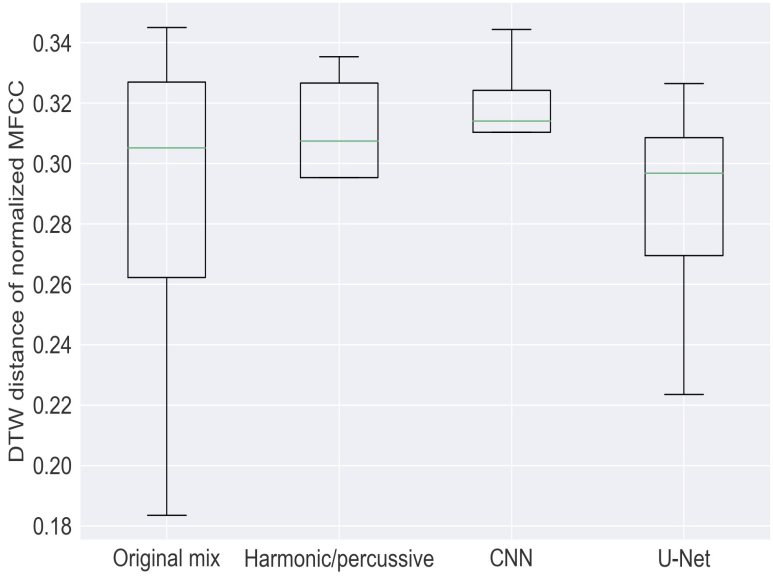


Figure 6.3: Boxplot representing the distribution of DTW distances between normalized MFCCs extracted from solo singing audio and corresponding original mixed audio, extracted vocals using harmonic/percussive, CNN and U-Net based audio source separation.

the model trained on iKala dataset, for voice, bass, and drums separation. Although CNN based audio source separation is undoubtedly successful, it would be interesting for us to investigate another method recently proposed by Jansson et al. [JHM+17], that uses U-Net architecture (initially developed for medical imaging) which has the capacity for recreating the fine, low-level detail required for high-quality audio reproduction.

In Figure 6.2, we show a comparison between the vocals obtained from the three audio source separation methods. Figure 6.2(a) and Figure 6.2(b) show the spectrograms (with 20 ms frame-size, 10 ms frame-shift, sampling rate 10 kHz) corresponding to original mixed audio and solo-singing audio for a 5.5 s segment of the song "*this afternoon*" from Hansen's dataset [Han12]. The extracted vocals for the same audio segment using harmonic/percussive, CNN and U-Net based audio source separation are shown in Figure 6.2(c),(d),(e) respectively. If we compare each of these with Figure 6.2(b) we can observe that, using harmonic/percussive method, the percussive component is removed, however the other components are preserved in the spectrogram, similar to the original mixed audio shown in Figure 6.2(a). After applying CNN based source separation, although the vocal specific characteristics are preserved in the spectrogram, as shown in Figure 6.2(c), there are some glitches present in the boundaries of the phonemes. This distortion is also evident for the songs with high intensity background accompaniment during informal listening. Figure 6.2(e) shows that the extracted vocal from U-Net based source separation has highest similarity with that of clean speech and is least distorted.

To further analyze the deviation of the extracted vocals from clean solo-singing audio, we

perform DTW between the normalized Mel-frequency cepstral coefficients (MFFCs) (13-dimensional) of the solo-singing, and the extracted vocals from the three source separation methods shown in Figure 6.3. It can be observed that the mean distance with solo-singing audio is significantly less in case of U-Net based source separation method, with reasonable standard deviation. This gives us the intuition that the U-Net method will reduce the train and test data deviation the most.

Table 6.1: Average absolute error/deviation (ASE) (seconds) and percentage of correct segments (PCS) for lyrics-to-audio alignment systems using GMM-HMM (SAT) model after applying different audio source separation methods.

Database	Source separation method							
	No source separation		Harmonic/percussive		CNN		U-Net	
Metric	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS
Hansen's [Han12]	33.81	0.12	24.56	0.08	10.99	0.23	**9.48**	0.36
Mauch's [MFG12]	26.94	0.13	24.83	0.04	12.28	0.13	**6.43**	0.25

Table 6.2: Average absolute error/deviation (ASE) (seconds) and percentage of correct segments (PCS) for lyrics-to-audio alignment systems using GMM-HMM (SAT) model after applying different audio source separation methods and removal of beginning and ending silences.

Database	Source separation method								MIREX 2017 best system		KKBOX System (MIREX 2018 best system)	
	No source separation		Harmonic/percussive		CNN		U-Net					
Metric	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS
Hansen's	18.78	0.19	20.69	0.09	3.14	0.29	**1.39**	0.13	**7.34**	0.25	**2.07**	0.45
Mauch's	25.42	0.08	25.83	0.03	17.74	0.06	**6.34**	0.07	**9.03**	0.15	**4.13**	0.35

6.1.3 Intro and outro non-vocal suppression

In the extracted sung vocals, the intro and outro sections, that contain only instrumental accompaniments, are suppressed resulting in silence segments. To detect these silence segments, we divided the spectrum of each frame (framesize 25 ms, frameshift 5 ms, sampling frequency 44.1 kHz) into four equal sub-bands. The energy corresponding to the 2^{nd} sub-band shows a prominent difference between the segments with vocals and without vocals. A threshold based on the average 2^{nd} sub-band energy is set to classify the frames into vocal and non-vocal. The non-vocal segments with very long duration corresponding to intro and outro are removed from the audio as a pre-processing step. Based on the above discussion, we propose the framework for lyrics-to-audio alignment, as shown in Figure 6.1.

6.2 Experimental Evaluation

We develop the lyrics-to-audio alignment framework at different stages to observe the efficiency and significance of each component. We have used Hansen's and Mauch's datasets for evaluation of the alignment systems. Hansen's dataset contains 9 pop music songs in

English with annotations of both begin- and end-timestamps of each word [Han12]. The audio has two versions: the original with instrumental accompaniment and a *capella* singing voice only. Mauch's dataset contains 20 pop music songs in English with annotations of begin-timestamps of each word [MFG12]. We have used two different metrics for the evaluation, which are average absolute error/deviation (ASE) and percentage of correct segments from total audio duration (PCS) [Dzh17].

As discussed in Section 6.1.1, we use singing-adapted speech acoustic models trained on solo-singing dataset to force-align lyrics with the audio. The baseline speech acoustic model is a tri-phone Gaussian mixture model (GMM)-HMM trained on Librispeech corpus [PCPK15] using MFCC features on Kaldi toolkit [PGB+11]. To make the Viterbi alignment algorithm operate over the long duration of songs (4–5 minutes), we set the alignment retry-beamwidth to a high value of 4000. Also the flag for optional silence was on to handle the possibility of pauses. To avoid misalignment due to the presence of long duration musical intro, we apply an energy-based algorithm over the extracted vocals to detect and remove the non-vocal part over the first few seconds of the song.

6.2.1 Effect of singing vocal separation

We first tested the system with solo-singing versions of the songs from Hansen's data, which gives average ASE of 0.4 second. Using the polyphonic version of the songs from the same dataset, we obtain an average ASE of 33.81 seconds as shown in Table 6.1. This shows that our singing-adapted acoustic models are well-suited for solo-singing data. However, as expected, they do not perform as well on polyphonic music. Therefore, we use different source separation methods to extract the singing vocal for which the performance of the system is depicted in Table 6.1. We observe that the average ASE values are best for the system with U-Net based source separation, which is 9.48 seconds for Hansen's data and 6.43 seconds for Mauch's data. Moreover, we observe that the harmonic/percussive method gives a relatively poor performance, which implies that the presence of other non-percussive instruments have an impact on the performance of singing voice models.

In order to reduce the alignment error further, we have detected and removed the silence segments at the beginning and ending of the extracted vocal. After obtaining the alignment, we have taken care of duration of these removed silence segments. The average ASE and PCS values for the systems with different audio source separation methods and after applying silence removal are depicted in Table 6.2. Analogous to the previous observation the U-Net source separation performs the best. We achieve ASE of 1.39 and 6.34 seconds for Hansen's and Mauch's data respectively. This is a significant improvement over MIREX 2017 [Mir17] best system. We also note that our system performance is comparable to MIREX 2018 [Mir18] best system which are 2.07 and 4.13 seconds respectively for the two datasets. Moreover, the performance of all the systems have significantly improved after removal of beginning and ending silences. To summarize, the average ASE value

obtained from our best system is 3.87 seconds. A demo of the presented results is given in https://www.comp.nus.edu.sg/~chitrale/LyricsAlignmentDemo.html.

Table 6.3: ASE and PCS for lyrics-to-audio alignment systems using SAT+DNN model after applying different audio source separation methods and removal of beginning and ending silences.

Database	Source separation method							
	No source separation		Harmonic/percussive		CNN		U-Net	
Metric	ASE	PCS	ASE	PCS	ASE	PCS	ASE	PCS
Hansen's	23.51	0.06	27.49	0.01	9.26	0.14	2.71	0.12
Mauch's	30.27	0.01	26.79	0.01	23.64	0.03	6.99	0.04

6.2.2 DNN-SAT singing-adapted models

We reported in Chapter 4 that the use of deep neural network (DNN) model achieved a significant reduction in the WER. A DNN model [HDY+12] is trained on top of the SAT model with the same set of training data. In this work, apart from the GMM-HMM (SAT) models, we have also applied the SAT+DNN model for lyrics-to-audio alignment for which the ASE and PCS values are shown in Table 6.3. Although the SAT+DNN model is reported to improve the lyrics recognition performance (Chapter 4), it does not show an improvement for the alignment task. The DNN models are not good for alignment since the objective function they are trained with does not force them to produce good alignments. For forced-alignment, a GMM-based model is generally recommended. So we expect that the alignment using GMM-HMM (SAT) models should perform better than that of SAT+DNN model. To show the results obtained for all the songs, we have plotted distributions in terms

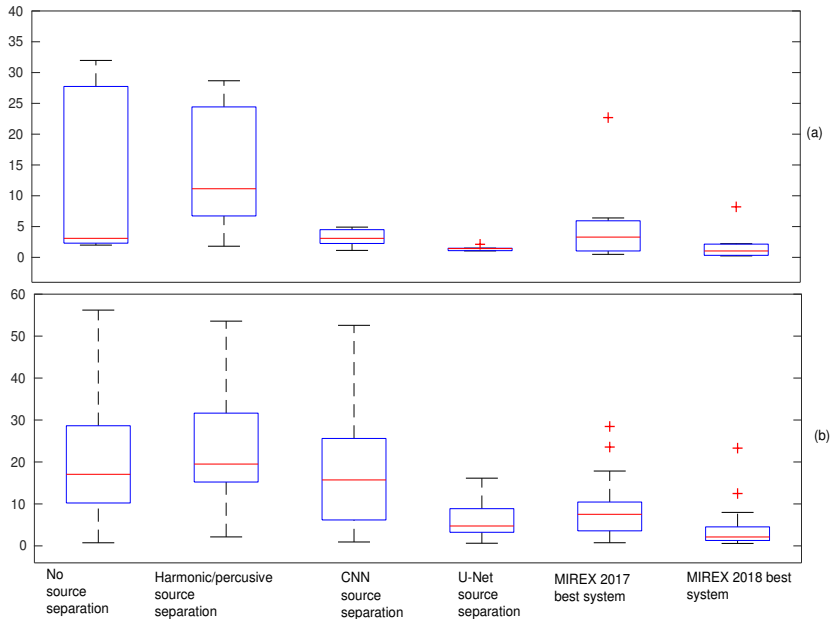


Figure 6.4: Boxplot showing the distribution of ASE values for all the songs from (a) Hansen's data (b) Mauch's data, using different systems shown in Table 6.2.

of boxplot shown in Figure 6.4 corresponding to all the songs for each system. We observe that the performance of the system using U-Net source separation has better performance compared to best system of MIREX 2017 and comparable performance with MIREX 2018 best system. In this case the system with CNN based source separation method also achieves a comparable result. For the Mauch's dataset similar distributions are shown in Figure 6.4 where the system with CNN based source separation performs poorer than U-Net source separation. In this case also it is evident that the proposed lyrics-to-audio alignment system has comparable performance with the best system of MIREX 2018.

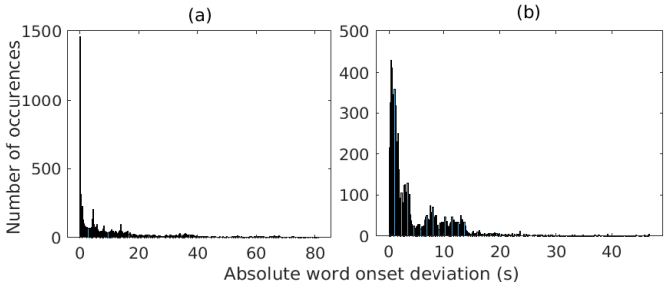


Figure 6.5: Histogram showing absolute word onset deviation for the alignment obtained using (a) CNN, (b) U-Net based vocal extraction.

To analyze the above mentioned results further we have shown the histogram distributions of absolute word onset deviation between the boundaries obtained from our algorithm and ground truth in Figure 6.5 for both the datasets. From Figure 6.5 in case of the system using CNN based source separation the range of error is very high compared to that of the system using U-Net source separation. We can observe that in lower bins the number of instances are prominently high in case of Figure 6.5. For our final lyrics-to-audio alignment system, using U-Net source separation higher number instances are accumulated towards the lower range of absolute word onset deviation as shown in Figure 6.5.

6.3 Summary

In this work, we present lyrics-to-audio alignment systems using singing adapted acoustic models. The speech acoustic models are adapted to singing voice using a relatively small set of solo-singing data. We use different audio source separation methods to extract the singing voice and obtain alignment for polyphonic songs. Three different audio source separation methods are compared and observed that the efficacy of singing vocal extraction has a high impact on alignment accuracy. The U-Net based audio source separation performs best for our system. After removal of the begin and end non-vocal sections, the system performance improves further. Our best system has lower ASE value compared to MIREX 2017 best system and comparable to that of MIREX 2018 best system, which is trained on a large polyphonic database. This study demonstrates that by using a relatively small solo-singing database for adaptation of speech models to singing voice, along with a reliable audio-source separation, we can develop a high performing lyrics-to-audio alignment system.

Chapter 7

Epilogue

In this thesis, we have proposed and validated various algorithms to objectively characterize and evaluate two broad aspects of singing voice quality - prosody and pronunciation. In Chapters 2 and 3, we discuss prosody-based singing quality evaluation. In Chapters 4, 5, and 6, we discuss the phonetic aspect of singing voice.

7.1 Summary

In Chapter 2, we presented a framework for automatic perceptual evaluation of singing quality (PESnQ). We evaluate test singing with respect to ideal or a standard reference singing. We analyze and compute various objective measures to evaluate perceptually relevant parameters of singing quality such as intonation, and rhythm, while incorporating the cognitive modeling theory for audio perception used in the PESQ standard. We inferred that humans may follow a two stage mental process to judge singing quality - convert the perceived singing audio signal into a weighted representation of the identified perceptual parameters as a symbolic representation, and then map them to an overall singing quality judgment through a human judgment parametric model. We applied the same idea to predict machine scores for singing quality judgment, i.e. predict the individual parameters first and linearly combine them as per the human judgment parametric model to compute the overall quality score. This late-fusion method showed a promising improvement over the early-fusion method.

In Chapter 3, we presented methods to evaluate singing quality in absence of a standard reference. Instead of relying on a reference singer, we leveraged on music theoretic rules, and inter-singer relative comparisons to rank large number of singers according to their singing quality. Our proposed method showed a high correlation of 0.71 with human judgments.

To address the problem of the lack of a lyrics-annotated singing voice dataset in the field of acoustic modeling of sung phonemes, in Chapter 4, we designed an algorithm that can automatically generate reliable time-aligned lyrics transcription for singing by using imperfect transcriptions from ASR along with the non-aligned published lyrics. An ASR, adapted to singing voice using this lyrics aligned singing voice dataset, showed a low recognition WER. In Chapter 5, we further improved these singing-adapted acoustic models by incorporating

a duration-based lexicon modeling for the vowels, that reduced the WER further to 30% approximately.

In Chapters 5, and 6, we observed the usability of this singing-adapted ASR for two applications of MIR: pronunciation evaluation in singing and audio-to-lyrics alignment, respectively. In Chapter 5, we developed a strategy to reliably evaluate pronunciation quality in singing. We found that our strategy of duration-based lexicon modeling of vowels has a positive impact on word alignment as well as evaluation accuracy.

In Chapter 6, we explored the application of these singing-adapted acoustic models for the task of audio-to-lyrics alignment. We applied a source separation algorithm as a pre-processing step to extract the singing voice from the polyphonic music. We demonstrated that our singing-adapted models in conjunction with a source separation algorithm could provide better word-level lyrics alignment than baseline systems.

7.1.1 Summary of the novel contributions of this work

- We improved our understanding about the importance of the perceptual parameters of singing skill evaluation and their representation in human mind to predict the singing quality scores (See Chapter 2.3.1, 2.4.4, and 2.4.7)
- We provided a comprehensive and musically relevant objective feedback, called PESnQ, to aspiring singers to help them improve their singing skills (See Chapter 2.3.1, and 2.4.5)
- We designed a self-organizing method to rank-order large number of singing renditions based on singing quality without relying on a reference singer (See Chapter 3.3, and 3.6.3)
- We proposed and analyzed various inter-singer relative distance measures and musically-motivated pitch histogram-based measures to characterize the inherent properties of singing quality, that provides a way to evaluate singing quality without a reference singing sample (See Chapter 3.4.1)
- We provided evidence that indicates that machines can provide a more unbiased assessment of the underlying parameters of singing quality compared to humans (See Chapter 3.6.8)
- We developed an algorithm to automatically obtain large-scale singing-lyrics annotated dataset to train singing-adapted speech acoustic models (See Chapter 4.2)
- We incorporated the strategy of duration-based lexicon modification to improve the singing-adapted speech acoustic models, that results in improvement in word alignment, recognition as well as pronunciation scoring accuracy (See Chapter 5.2.1 and 5.3.3)

- We designed a method to automatically assess pronunciation in singing voice, which has applications in language learning and speech therapy (See Chapter 5.2 and 5.3.3)
- We demonstrated that with the help of our singing-adapted speech acoustic models, along with a reliable audio-source separation, we can develop a high performing lyrics-to-audio alignment system (See Chapter 6.1 and 6.2.1)

7.2 Future Work

From the phonetic aspect, the adaptation of singing-adapted phonetic models for language learning and speech therapy application needs to be further explored. For example, analyzing the relationships between singer geographical origin, song difficulty level, and evaluation accuracy. The acoustic models can be further improved by data-specific information, such as accent, music genre, and type of background music.

This work has opened doors to the possibility of a large-scale mining of talented singers with the help of machines. In the future, it will be interesting to analyze features that can be learned by a neural network when assessing singing quality. Whether such features provide any further insights to our understanding of singing voice evaluation should be investigated, as it will be beneficial for further MIR-related research. Furthermore, generalizability of our singing quality evaluation measures to other music genres and styles needs to be investigated in the future.

References

[ABD+81] Bruce Anderson, David G Berger, R Serge Denisoff, K Peter Etzkorn, and Peter Hesbacher. Love negative lyrics: Some shifts in stature and alterations in song. *Communications*, 7(1):3–20, 1981.

[AEGF09] Noam Amir, Tom Erlich, Nitzan Grabstein, and Jacob Fainguelernt. Automated evaluation of singers' vibrato through time and frequency analysis of the pitch contour using the dsk6713. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–5. IEEE, 2009.

[AP06] S Omar Ali and Zehra F Peynircioğlu. Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music*, 34(4):511–534, 2006.

[ASH73] Martin L Albert, Robert W Sparks, and Nancy A Helm. Melodic intonation therapy for aphasia. *Archives of neurology*, 29(2):130–131, 1973.

[B+02] Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

[BAB+11] Elvira Brattico, Vinoo Alluri, Brigitte Bogert, Thomas Jacobsen, Nuutti Vartiainen, Sirke Nieminen, and Mari Tervaniemi. A functional mri study of happy and sad emotions in music with and without lyrics. *Frontiers in psychology*, 2, 2011.

[BDd+13] Onur Babacan, Thomas Drugman, Nicolas d'Alessandro, Nathalie Henrich, and Thierry Dutoit. A comparative study of pitch extraction algorithms on a large variety of singing sounds. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7815–7819. IEEE, 2013.

[BES+17] Johanna Böhm, Florian Eyben, Maximilian Schmitt, Harald Kosch, and Björn Schuller. Seeking the superstar: Automatic assessment of perceived singing quality. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 1560–1569. IEEE, 2017.

[Bil] E. Billauer. function PeakDet, MATLAB (Converted to python). https://gist.github.com/endolith/250860. [Online; accessed 20-May-2018].

[Boz08] Barış Bozkurt. An automatic pitch analysis method for turkish maqam music. *Journal of New Music Research*, 37(1):1–13, 2008.

[CH08] Lauren B Collister and David Huron. Comparison of word intelligibility in spoken and sung phrases. 2008.

[Cha07] Pei-Chen Chang. Method and apparatus for karaoke scoring, December 4 2007. US Patent 7,304,229.

[CL16] Nancy F Chen and Haizhou Li. Computer-assisted pronunciation training: From pronunciation scoring towards spoken language learning. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, pages 1–7. IEEE, 2016.

[CLLY08] Chuan Cao, Ming Li, Jian Liu, and Yonghong Yan. A study on singing performance evaluation criteria for untrained singers. In *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pages 1475–1478. IEEE, 2008.

[ÇLT17] Berrak Çişman, Haizhou Li, and Kay Chen Tan. Sparse representation of phonetic features for voice conversion with and without parallel data. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 677–684. IEEE, 2017.

[CMJG17] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 258–266. Springer, 2017.

[CSH15] Nathaniel Condit-Schultz and David Huron. Catching the lyrics: intelligibility in twelve song genres. *Music Perception: An Interdisciplinary Journal*, 32(5):470–483, 2015.

[CWJ16] Yu-Ren Chien, Hsin-Min Wang, and Shyh-Kang Jeng. Alignment of lyrics with accompanied singing audio based on acoustic-phonetic vowel likelihood modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1998–2008, 2016.

[Dar88] Charles Darwin. *The descent of man and selection in relation to sex*, volume 1. Murray, 1888.

[DBS11] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *12th International Society for Music Information Retrieval Conference (ISMIR-2011)*, pages 669–674. University of Miami, 2011.

[DFL+13] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *Signal and Information Processing Association Annual*

Summit and Conference (APSIPA), 2013 Asia-Pacific, pages 1–9. IEEE, 2013.

[DS11] F. Dege and G. Schwarzer. The effect of a music program on phonological awareness in preschoolers. *Frontiers in Psychology*, 2(124):7–13, 2011.

[DS15] Georgi Bogomilov Dzhambazov and Xavier Serra. Modeling of phoneme durations for alignment between polyphonic audio and lyrics. In *12th Sound and Music Computing Conference*, pages 281–286, 2015.

[Dzh17] Georgi Dzhambazov. *Knowledge-based Probabilistic Modeling for Tracking Lyrics in Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, 2017.

[ea12] G. Hinton et al. Deep neural networks for acoustic modeling in speech recognition. In *IEEE Signal Processing Magazine*, volume 29, pages 82–97, 2012.

[Eco08] The Economist. Why music? https://www.economist.com/christmas-specials/2008/12/18/why-music, 2008.

[EWGS13] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[FG12] Hiromasa Fujihara and Masataka Goto. Lyrics-to-audio alignment and its application. In *Dagstuhl Follow-Ups*, volume 3. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

[FGO+06] Hiromasa Fujihara, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G Okuno. Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals. In *Eighth IEEE International Symposium on Multimedia*, pages 257–264, 2006.

[FGOO11] Hiromasa Fujihara, Masataka Goto, Jun Ogata, and Hiroshi G Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1252–1261, 2011.

[Fit10] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *3th International Conference on Digital Audio Effects (DAFX10), Graz, Austria*. Dublin Institute of Technology, 2010.

[For65] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

[Gal98] Mark JF Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[GC89] Laurence Gillick and Stephen J Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*, pages 532–535. IEEE, 1989.

[GCOC15] Rong Gong, Philippe Cuvillier, Nicolas Obin, and Arshia Cont. Real-time audio-to-score alignment of singing voice based on melody and lyric information. In *Interspeech*, 2015.

[GGRW17] Chitralekha Gupta, David Grunberg, Preeti Rao, and Ye Wang. Towards automatic mispronunciation detection in singing. In *Proceedings of International Society for Music Information Retrieval (ISMIR), Suzhou, China, 2017*, 2017.

[GLW17] Chitralekha Gupta, Haizhou Li, and Ye Wang. Perceptual evaluation of singing quality. In *Proceedings of APSIPA Annual Summit and Conference*, volume 2017, pages 12–15, 2017.

[GLW18] Chitralekha Gupta, Haizhou Li, and Ye Wang. Automatic evaluation of singing quality without a reference. In *To appear in Proceedings of APSIPA Annual Summit and Conference*, 2018.

[GRS15a] A. Good, F. Russo, and J. Sullivan. The efficacy of singing in foreign-language learning. *Psychology of Music*, 43(5):627–640, 2015.

[GRS15b] Arla J Good, Frank A Russo, and Jennifer Sullivan. The efficacy of singing in foreign-language learning. *Psychology of Music*, 43(5):627–640, 2015.

[GS18] Rong Gong and Xavier Serra. Singing voice phoneme segmentation by hierarchically inferring syllable and phoneme onset positions. In *Interspeech 2018, India*, 2018.

[Han12] Jens Kofod Hansen. Recognition of phonemes in a-cappella recordings using temporal patterns and mel frequency cepstral coefficients. In *9th Sound and Music Computing Conference (SMC)*, pages 494–499, 2012.

[HDY+12] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[HFH+09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[HH09] Edward H Hagen and Peter Hammerstein. Did neanderthals and other early humans sing? seeking the biological roots of music in the territorial advertisements of primates, lions, hyenas, and wolves. *Musicae Scientiae*, 13(2_suppl):291–320, 2009.

[HHG94] Michael P Hollier, MO Hawksford, and DR Guard. Error activity and error entropy as a measure of psychoacoustic significance in the perceptual domain. *IEE Proceedings-Vision, Image and Signal Processing*, 141(3):203–208, 1994.

[HLQM09] Alissa M Harrison, Wai-Kit Lo, Xiao-jun Qian, and Helen Meng. Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *International Workshop on Speech and Language Technology in Education*, 2009.

[HM13] Sean Hutchins and Sylvain Moreno. The linked dual representation model of vocal perception and production. *Frontiers in psychology*, 4:825, 2013.

[Hof60] Paul J Hoffman. The paramorphic representation of clinical judgment. *Psychological bulletin*, 57(2):116, 1960.

[HP95] Reid Hastie and Nancy Pennington. Cognitive approaches to judgment and decision making. In *Psychology of learning and motivation*, volume 32, pages 1–31. Elsevier, 1995.

[HSD06] Dave Hoppe, Makiko Sadakata, and Peter Desain. Development of real-time visual feedback assistance in singing training: a review. *Journal of computer assisted learning*, 22(4):308–316, 2006.

[HW89] David M Howard and Graham F Welch. Microcomputer-based singing ability assessment and development. *Applied Acoustics*, 27(2):89–102, 1989.

[IFP18] IFPI. Global music report 2018: state of the industry. https://www.ifpi.org/downloads/GMR2018.pdf, 2018.

[IWKL06] Denny Iskandar, Ye Wang, Min-Yen Kan, and Haizhou Li. Syllabic level automatic synchronization of music signals and text lyrics. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 659–662. ACM, 2006.

[JH05] Timothy Justus and Jeffrey J Hutsler. Fundamental issues in the evolutionary psychology of music: Assessing innateness and domain specificity. *Music Perception: An Interdisciplinary Journal*, 23(1):1–27, 2005.

[JHM+17] Andreas Jansson, Eric J. Humphrey, Nicola Montecchio, Rachel M. Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In *ISMIR*, 2017.

[JR12] P. Jha and P. Rao. Assessing vowel quality for singing evaluation. In *National Conference on Communications (NCC) 2012, IEEE*, pages 1–5, 2012.

[KEF01] Hideki Kawahara, Jo Estill, and Osamu Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[KFN97] Yoon Kim, Horacio Franco, and Leonardo Neumeyer. Automatic pronunciation scoring of specific phone segments for language instruction. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[Kon10] Kevin W Kondik. *A Critical Review of Three Theories for MusicâĂŹs Origin*. PhD thesis, Ohio University, 2010.

[Koo94] John Koopman. *A brief history of singing*. John Koopman, 1994.

[Kru15] Anna M Kruspe. Training phoneme models for singing with "songified" speech data. In *ISMIR*, pages 336–342, 2015.

[Kru16a] Anna M Kruspe. Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing. In *ISMIR*, pages 358–364, 2016.

[Kru16b] Anna M Kruspe. Retrieval of textual song lyrics from sung inputs. In *INTERSPEECH*, pages 2140–2144, 2016.

[KW17] Filip Korzeniowski and Gerhard Widmer. End-to-end musical key estimation using a convolutional neural network. *arXiv preprint arXiv:1706.02921*, 2017.

[KWI+08] Min-Yen Kan, Ye Wang, Denny Iskandar, Tin Lay Nwe, and Arun Shenoy. Lyrically: Automatic synchronization of textual lyrics to acoustic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):338–349, 2008.

[Lal06] Partha Lal. A comparison of singing evaluation algorithms. In *Ninth International Conference on Spoken Language Processing*, 2006.

[LC08] Kyogu Lee and Markus Cremer. Segmentation-based lyrics-audio alignment using dynamic programming. In *ISMIR*, pages 395–400, 2008.

[LC18] Xueling Lin and Lei Chen. Domain-aware multi-truth discovery from conflicting sources. *Proceedings of the VLDB Endowment*, 11(5):635–647, 2018.

[LCB99] Alex Loscos, Pedro Cano, and Jordi Bonada. Low-delay singing voice alignment to text. In *ICMC*, 1999.

[Leh04] I. Lehiste. Prosody in speech and singing. In *Speech Prosody 2004, International Conference*, 2004.

[Lev66] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[LG12] Ann Lee and James Glass. A comparison-based approach to mispronunciation detection. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 382–387. IEEE, 2012.

[LLCW14] Chang-Hung Lin, Yuan-Shan Lee, Ming-Yen Chen, and Jia-Ching Wang. Automatic singing evaluating system based on acoustic features and rhythm. In *Orange Technologies (ICOT), 2014 IEEE International Conference on*, pages 165–168. IEEE, 2014.

[LLI+13] Jordan Louviere, Ian Lings, Towhidul Islam, Siegfried Gudergan, and Terry Flynn. An introduction to the application of (case 1) best–worst scaling in marketing research. *International Journal of Research in Marketing*, 30(3):292–303, 2013.

[Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[LZG13] Ann Lee, Yaodong Zhang, and James Glass. Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8227–8231. IEEE, 2013.

[MBG+13] Emilio Molina, Isabel Barbancho, Emilia Gómez, Ana Maria Barbancho, and Lorenzo J Tardón. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 744–748. IEEE, 2013.

[MBR10] Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273. IEEE, 2010.

[MEG14] Matt McVicar, Daniel PW Ellis, and Masataka Goto. Leveraging repetition for improved automatic lyric transcription in popular music. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3117–3121. IEEE, 2014.

[MFA15] AAJ Marley, Terry N Flynn, and Victoria Australia. Best worst scaling: theory and practice. *International Encyclopedia of the Social & Behavioral Sciences*, 2(2):548–552, 2015.

[MFG10] Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Lyrics-to-audio alignment and phrase-level segmentation using incomplete internet-style chord annotations. In *Proceedings of the 7th Sound and Music Computing Conference (SMC 2010)*, pages 9–16, 2010.

[MFG12] Matthias Mauch, Hiromasa Fujihara, and Masataka Goto. Integrating additional chord information into hmm-based lyrics-to-audio alignment. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):200–210, 2012.

[Mir17] Mirex 2017. https://www.music-ir.org/mirex/wiki/2017:Automatic_Lyrics-to-Audio_Alignment_Results, 2017. [Online; accessed 28-October-2018].

[Mir18] Mirex 2018. https://www.music-ir.org/mirex/wiki/2018:Automatic_Lyrics-to-Audio_Alignment_Results, 2018. [Online; accessed 28-October-2018].

[MJTG98] Pedro J Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. A recursive algorithm for the forced alignment of very long audio segments. In *Fifth International Conference on Spoken Language Processing*, 1998.

[MMW+06] Steven Mithen, Iain Morley, Alison Wray, Maggie Tallerman, and Clive Gamble. The singing neanderthals: the origins of music, language, mind and body, by steven mithen. london: Weidenfeld & nicholson, 2005. isbn 0-297-64317-7. *Cambridge Archaeological Journal*, 16(1):97–112, 2006.

[Mon17] Jeremy Montagu. How music and instruments began: a brief overview of the origin and entire development of music, from its earliest stages. *Frontiers in Sociology*, 2:8, 2017.

[MPTE10] R. Milovanov, P. PietilÃď, M. Tervaniemi, and P. Esquef. Foreign language pronunciation skills and musical aptitude:a study of Finnish adults with higher education. *Learning and Individual Differences*, 20(1):56–60, 2010.

[MR13] C. Markus and S.M. Reiterer. Song and speech: examining the link between singing talent and speech imitation ability. *Frontiers in psychology*, 4:874, 2013.

[MRL+15] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015.

[MV08] Annamaria Mesaros and Tuomas Virtanen. Automatic alignment of music audio and lyrics. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, 2008.

[MV10] Annamaria Mesaros and Tuomas Virtanen. Automatic recognition of lyrics in singing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):546047, 2010.

[NDAL12] Eric Nichols, Charles DuHadway, Hrishikesh Aradhye, and Richard F Lyon. Automatically discovering talented musicians with acoustic analysis of youtube videos. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 559–565. IEEE, 2012.

[NDC+10] Tin Lay New, Minghui Dong, Paul Chan, Xi Wang, Bin Ma, and Haizhou Li. Voice conversion: From spoken vowels to singing vowels. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1421–1426. IEEE, 2010.

[NFDW00] L. Neumeyer, H. Franco, V Digalakis, and M. Weintraub. Automatic scoring of pronunciation quality. *Speech Communication*, 30(2):83–93, 2000.

[NGH06a] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Ninth International Conference on Spoken Language Processing*, 2006.

[NGH06b] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. Subjective evaluation of common singing skills using the rank ordering method. In *Ninth International Conference on Music Perception and Cognition*. Citeseer, 2006.

[NPW+15] Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, and Sebastian Möller. Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[NS11] Hitomi Nakata and Linda Shockey. The effect of singing on improving syllabic pronunciation–vowel epenthesis in japanese. In *International Conference of Phonetic Sciences*, 2011.

[NZMS09] Andrea Norton, Lauryn Zipse, Sarah Marchina, and Gottfried Schlaug. Melodic intonation therapy. *Annals of the New York Academy of Sciences*, 1169(1):431–436, 2009.

[OBD+06] Jennifer M Oates, Belinda Bain, Pamela Davis, Janice Chapman, and Dianna Kenny. Development of an auditory-perceptual rating instrument for the operatic singing voice. *Journal of Voice*, 20(1):71–81, 2006.

[OKC+96] Koichi Omori, Ashutosh Kacker, Linda M Carroll, William D Riley, and Stanley M Blaugrund. Singing power ratio: quantitative evaluation of singing voice quality. *Journal of voice*, 10(3):228–235, 1996.

[PBTO13] Jeanne Parson, Daniela Braga, Michael Tjalve, and Jieun Oh. Evaluating voice quality and speech synthesis using crowdsourcing. In *International Conference on Text, Speech and Dialogue*, pages 233–240. Springer, 2013.

[PCPK15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[PGB+11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[PGS17] Jordi Pons, Rong Gong, and Xavier Serra. Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks. In *ISMIR 2017, Suzhou, China*, 2017.

[PH03] Isabelle Peretz and Krista L Hyde. What is specific to music processing? insights from congenital amusia. *Trends in cognitive sciences*, 7(8):362–367, 2003.

[Pin97] Steven Pinker. How the mind works. new york: W. w, 1997.

[PMSW17] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012. International World Wide Web Conferences Steering Committee, 2017.

[POF06] Christopher J Plack, Andrew J Oxenham, and Richard R Fay. *Pitch: neural coding and perception*, volume 24. Springer Science & Business Media, 2006.

[Pra08] P Prasertvithyakarn. An automatic singing voice evaluation method for voice training. *é§şåŋęèňŻèńŰ, 2008-03*, pages 911–912, 2008.

[PS06] Daniel Povey and George Saon. Feature and model space speaker adaptation with full covariance gaussians. In *Ninth International Conference on Spoken Language Processing*, 2006.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Rab77] Lawrence Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE transactions on acoustics, speech, and signal processing*, 25(1):24–33, 1977.

[RBHH01] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, volume 2, pages 749–752. IEEE, 2001.

[RFZS11] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 2416–2419. IEEE, 2011.

[RS78] Lawrence R Rabiner and Ronald W Schafer. *Digital processing of speech signals*, volume 100. Prentice-hall Englewood Cliffs, NJ, 1978.

[SBN90] Mary Kay Stevenson, Jerome R Busemeyer, and James C Naylor. Judgment and decision-making theory. 1990.

[SGUA07] Takeshi Saitou, Masataka Goto, Masashi Unoki, and Masato Akagi. Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices. In *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, pages 215–218. IEEE, 2007.

[Smu] Smule. Digital Archive Mobile Performances (DAMP). https://ccrma.stanford.edu/damp/. [Online; accessed 15-March-2018].

[Smu08] Smule. Sing! Karaoke app. https://www.smule.com, 2008.

[SR90] Johan Sundberg and Thomas D Rossing. The science of singing voice, 1990.

[SSB+13] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[Sta10] Starmaker. Starmaker karaoke app. https://www.starmakerstudios.com, 2010.

[Tan99] Takahiro Tanaka. Karaoke scoring apparatus analyzing singing voice relative to melody data, March 30 1999. US Patent 5,889,224.

[Tea] Developing pronunciation through songs. https://www.teachingenglish.org.uk/article/developing-pronunciation-through-songs. Accessed: 2017-04-27.

[TEC03] George Tzanetakis, Andrey Ermolinskyi, and Perry Cook. Pitch histograms in audio and symbolic music information retrieval. *Journal of New Music Research*, 32(2):143–152, 2003.

[TL12] Wei-Ho Tsai and Hsin-Chieh Lee. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1233–1243, 2012.

[VDMDSKM+16] Ineke Van Der Meulen, Van De Sandt-Koenderman, WME Mieke, Majanka H Heijenbrok, Evy Visch-Brink, and Gerard M Ribbers. Melodic intonation therapy in chronic aphasia: evidence from a pilot randomized controlled trial. *Frontiers in human neuroscience*, 10:533, 2016.

[WE97] Joel Wapnick and Elizabeth Ekholm. Expert consensus in solo voice performance evaluation. *Journal of Voice*, 11(4):429–436, 1997.

[Wel94] Graham F Welch. The assessment of singing. *Psychology of Music*, 22(1):3–19, 1994.

[WHR89] Graham F Welch, David M Howard, and Christine Rush. Real-time visual feedback in the development of vocal pitch accuracy in singing. *Psychology of Music*, 17(2):146–157, 1989.

[WKN+04] Ye Wang, Min-Yen Kan, Tin Lay Nwe, Arun Shenoy, and Jun Yin. Lyrically: automatic synchronization of acoustic musical signals and textual lyrics. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 212–219. ACM, 2004.

[WM12] Hao Wang and Helen Meng. Deriving perceptual gradation of l2 english mispronunciations using crowdsourcing and the workerrank algorithm. In *Speech Database and Assessments (Oriental COCOSDA), 2012 International Conference on*, pages 145–150. IEEE, 2012.

[WY00] Silke M Witt and Steve J Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, 30(2-3):95–108, 2000.

[YHP08] Xiaoxin Yin, Jiawei Han, and S Yu Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[ZB12] Robert J Zatorre and Shari R Baum. Musical melody and speech intonation: Singing a different tune. *PLoS biology*, 10(7):e1001372, 2012.

[ZG09] Yaodong Zhang and James R Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 398–403. IEEE, 2009.

[Zha17] A Zhang. Speech Recognition (Version 3.7) [Software]. https://github.com/Uberi/speech_recognition#readme, 2017. [Online; accessed 14-Oct-2017].

[ZRS14] Shuo Zhang, Rafael Caro Repetto, and Xavier Serra. Study of the similarity between linguistic tones and melodic pitch contours in beijing opera singing. In *ISMIR*, pages 343–348, 2014.

[ZSG+05] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S. Yoon. Accent detection and speech recognition for shanghai-accented mandarin. In *Interspeech*, pages 217–220. Citeseer, 2005.