TAMPERE UNIVERSITY OF TECHNOLOGY

Annamaria Mesaroş

Singing Voice Recognition for Music Information Retrieval

Julkaisu 1064 • Publication 1064

Tampere 2012

Tampereen teknillinen yliopisto. Julkaisu 1064
Tampere University of Technology. Publication 1064

Annamaria Mesaroş

Singing Voice Recognition for Music Information Retrieval

Thesis for the degree of Doctor of Science in Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB103, at Tampere University of Technology, on the 4th of September 2012, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2012

Supervisor:

Tuomas Virtanen, Docent
Department of Signal Processing
Tampere University of Technology
Tampere, Finland

Pre-examiners:

Petri Toiviainen, Professor
Department of Music
University of Jyväskylä
Jyväskylä, Finland

Geoffroy Peeters, Ph.D.
Sound Analysis & Synthesis team
IRCAM
Paris, France

Opponents:

Gaël Richard, Professor
Signal and Image Processing Department
TELECOM ParisTech
Paris, France

Olivier Lartillot, Ph.D.
Department of Music
University of Jyväskylä
Jyväskylä, Finland

ISBN 978-952-15-2893-4 (printed)
ISBN 978-952-15-3043-2 (PDF)
ISSN 1459-2045

Abstract

This thesis proposes signal processing methods for analysis of singing voice audio signals, with the objectives of obtaining information about the identity and lyrics content of the singing. Two main topics are presented, singer identification in monophonic and polyphonic music, and lyrics transcription and alignment. The information automatically extracted from the singing voice is meant to be used for applications such as music classification, sorting and organizing music databases, music information retrieval, etc.

For singer identification, the thesis introduces methods from general audio classification and specific methods for dealing with the presence of accompaniment. The emphasis is on singer identification in polyphonic audio, where the singing voice is present along with musical accompaniment. The presence of instruments is detrimental to voice identification performance, and eliminating the effect of instrumental accompaniment is an important aspect of the problem. The study of singer identification is centered around the degradation of classification performance in presence of instruments, and separation of the vocal line for improving performance. For the study, monophonic singing was mixed with instrumental accompaniment at different signal-to-noise (singing-to-accompaniment) ratios and the classification process was performed on the polyphonic mixture and on the vocal line separated from the polyphonic mixture. The method for classification including the step for separating the vocals is improving significantly the performance compared to classification of the polyphonic mixtures, but not close to the performance in classifying the monophonic singing itself. Nevertheless, the results show that classification of singing voices can be done robustly in polyphonic music when using source separation.

In the problem of lyrics transcription, the thesis introduces the general speech recognition framework and various adjustments that can be done before applying the methods on singing voice. The variability of

phonation in singing poses a significant challenge to the speech recognition approach. The thesis proposes using phoneme models trained on speech data and adapted to singing voice characteristics for the recognition of phonemes and words from a singing voice signal. Language models and adaptation techniques are an important aspect of the recognition process. There are two different ways of recognizing the phonemes in the audio: one is alignment, when the true transcription is known and the phonemes have to be located, other one is recognition, when both transcription and location of phonemes have to be found. The alignment is, obviously, a simplified form of the recognition task.

Alignment of textual lyrics to music audio is performed by aligning the phonetic transcription of the lyrics with the vocal line separated from the polyphonic mixture, using a collection of commercial songs. The word recognition is tested for transcription of lyrics from monophonic singing. The performance of the proposed system for automatic alignment of lyrics and audio is sufficient for facilitating applications such as automatic karaoke annotation or song browsing. The word recognition accuracy of the lyrics transcription from singing is quite low, but it is shown to be useful in a query-by-singing application, for performing a textual search based on the words recognized from the query. When some key words in the query are recognized, the song can be reliably identified.

Preface

This work has been carried out at the Department of Signal Processing, Tampere University of Technology, during 2006–2011. I wish to express my gratitude to my supervisor Dr. Tuomas Virtanen for guidance and collaboration throughout this work. I also wish to thank my former supervisor Dr. Anssi Klapuri who guided me for the first years of the thesis work. Their knowledgeable advice and leadership constitute the foundation of this thesis, and all the other pieces were built during the time we worked together.

I am grateful to Prof. Corneliu Rusu from Technical University of Cluj Napoca for encouraging me to experience working in an international research institute, and to Prof. Jaakko Astola who provided that opportunity and invited me to the Tampere International Center for Signal Processing (TICSP), where it all began.

I owe thanks to the pre-examiners of my thesis, Prof. Petri Toiviainen and Dr. Geoffroy Peeters, for the thorough review of the manuscript, and also to Prof. Gaël Richard and Dr. Olivier Lartillot for agreeing to be the opponents in the public defence of this thesis.

The Audio Research Group has provided a friendly working environment, and I am grateful for being part of it. Special thanks to Jouni Paulus for being also a close friend. I also thank all the other past and present ARG members that I had the privilege of knowing, including, but not limited to Toni Heittola, Elina Helander, Marko Helén, Antti Hurmalainen, Teemu Karjalainen, Konsta Koppinen, Teemu Korhonen, Katariina Mahkonen, Joonas Nikunen, Pasi Pertilä, Matti Ryynänen, Hanna Silén, and Sakari Tervo.

The funding and financial support of the Graduate School in Electronics, Telecommunications and Automation (GETA) and Nokia Foundation is gratefully acknowledged. This work was partly supported by the Academy of Finland (Finnish Centre of Excellence 2006–2011).

I wish to thank all my friends for providing me with activities and entertainment for my free time. There is not enough space to name

them all, but special thanks go to Nasko and Stane for all the good times, the chit chat and the serious discussions from the long autumn evenings. I also wish to thank to my sister Erika and my friend Cristina for being my online companions throughout these years, and making me feel like I am not away from home.

My deepest gratitude goes to my parents Rozalia and Ștefan for encouraging me all the way. Last, but not least, I thank Toni for being part of my life.

Annamaria Mesaroș
Tampere, 2012

Contents

List of Included Publications

This thesis consists of the following publications, preceded by an introduction to the research field and a summary of the publications. Parts of this thesis have been previously published and the original publications are reprinted, by permission, from the respective copyright holders. The publications are referred to in the text by notation [P1], [P2], and so forth.

P1 A. Mesaros, T. Virtanen and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proceedings of the 8th International Conference on Music Information Retrieval*, (Vienna, Austria), pp. 375–378, 2007

P2 A. Mesaros and T. Virtanen, "Automatic alignment of music audio and lyrics," in *Proceedings of the 11th International Conference on Digital Audio Effects*, (Espoo, Finland), pp. 321–324, 2008

P3 T. Virtanen, A. Mesaros and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, (Brisbane, Australia), pp. 17–22, 2008

P4 A. Mesaros and T. Virtanen, "Adaptation of a speech recognizer for singing voice," in *Proceedings of the 17th European Signal Processing Conference*, (Glasgow, Scotland), pp. 1779–1783, 2009

P5 A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing, " in *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing*, (Dallas, Texas, USA), pp. 2146 – 2149, 2010

P6 A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing", *EURASIP Journal on Audio, Speech and Music Processing*, Volume 2010, 11 pages, 2010

Publication [P1] was done in collaboration with Tuomas Virtanen and Anssi Klapuri. The signal separation algorithm was developed and implemented by Tuomas Virtanen. Anssi Klapuri assisted with the formulation of the evaluation. The rest of the implementations, evaluations, and most of the writing work were carried out by the author.

For publication [P3], the author provided the application for evaluating the proposed singing voice separation method, and writing of the corresponding section.

Publications [P2] and [P4] – [P6] were done in collaboration with Tuomas Virtanen who provided the singing voice separation algorithm and assisted with the development of the methods. The implementation of the speech recognition system, adaptation methods, evaluations, applications and most of the writing work were carried out by the author.

List of Abbreviations

DFT	Discrete Fourier transform
DTW	Dynamic time warping
GMM	Gaussian mixture model
HMM	Hidden Markov model
kNN	k-nearest neighbor
LPC	Linear prediction coefficient
LPCC	Linear prediction cepstral coefficient
LVCSR	Large vocabulary continuous speech recognition
MFCC	Mel-frequency cepstral coefficient
MIR	Music information retrieval
MIREX	Music Information Retrieval Evaluation eXchange
NMF	Non-negative matrix factorization
OOV	Out of vocabulary
PLP	Perceptual linear prediction
QBH	Query by humming
SVM	Support vector machine

Chapter 1

Introduction

Singing voice is considered by some to be the oldest and most expressive musical instrument. Singing plays a central role in songs and it attracts the attention of the listener. Like speech, the singing voice carries information about one's identity and the language information that can be represented as written text. Being able to automatically extract such information directly from the audio can be useful for music information retrieval purposes. Unfortunately, while people generally recognize a familiar voice without any effort, and are usually capable of writing down the lyrics in a song, automatic methods are still quite far from similar performance.

Areas of signal processing that study the singing voice comprise a large variety of topics: singing detection and recognition, singer identification, alignment of singing and textual lyrics or score sheet music, recognition of lyrics content, voice separation, singing voice analysis and synthesis, voice quality and timbre classifications. The work presented in this thesis is concerned with extracting the information about singer identity and lyrics content from the audio signal.

1.1 Terminology

The terminology used in the music information retrieval (MIR) community is not always unambiguous. The main two topics referred in this thesis are *singer identification* and *lyrics transcription*. The choice of terms is presented in the following, along with the descriptions of closely related terms from different areas of research.

The term *artist* associated with a song represents the singer, the band, or the stage name of the person creating or performing the song.

Artist identification is a task of MIREX (Music Information Retrieval Evaluation eXchange) [18] described as identification of the artist or group from musical audio. To avoid any ambiguity, the term *singer identification* [102] can be used to specify the task of recognizing the person who is performing the vocal part of the song, being equivalent to lead singer in case of a band.

Speech recognition in the speech community refers to providing a transcription of speech from audio into written text. The text is a symbolic representation of the spoken words. Research topics revolving around the singing voice were related to recognizing singing in the classification sense, for discrimination between speech and singing [11] or to *singing detection* [6, 20, 44, 60, 77] – a binary classification of a segment of music depending whether it contains singing or not, the segment being usually polyphonic music.

The task of providing a written symbolic representation of the music is referred to as *music transcription*. The symbolic representation provided as an output in music transcription is generally a musical score [12]. In a similar way, *singing transcription* is used to mean the melody transcription of the singing voice [81, 99], because the singing usually represents the (main) melody in a song.

Singing recognition was used to mean also the transcription of lyrics from singing [82]. We chose to use the term *lyrics transcription*, in order to avoid ambiguity. We can easily assume that people associate the term ”lyrics” with the words of a song and therefore it seems easy to understand that lyrics transcription means providing the textual transcription of the words from a song.

The material for study consists of audio containing only singing voice and audio containing singing and instrumental accompaniment. *Monophonic* music refers to music where only one sound source is present – for the purpose of this work it is the singing voice. In singing voice related literature, the terms *monophonic singing*, *clean singing* and *acapella singing* usually all refer to single voice audio material; when the material contains multiple singers, for example duets, it is specifically emphasized. *Polyphonic* audio is music containing multiple sound sources that are overlapping. To the purpose of this thesis, polyphonic music is composed of one singing voice accompanied by one or more instruments. *Monaural* audio refers to single channel audio signals, and *stereo* to two channel signals. Commercial music is usually a polyphonic stereo audio signal.

1.2 Overview of Singing Voice Related Research Topics

The need for classifying large amounts of music appeared with the digitization of music and the expansion of storage space. Suddenly it was possible to have very large music collections and organizing them became a problem. For music stores, the classification of music according to genre and artist name is not enough anymore, because people are interested in finding new music and asking for recommendations is not always solving the issue. Music needs to be automatically organized, sorted, analyzed, retrieved, transcribed, and all this can be made possible based on automatic analysis of the audio content.

In the development of music related research, the singing voice has its own place. There are many applications related to singing and lots of specific research. As mentioned earlier, humans are very skillful in separating the information from music, following only the singing voice in a song, recognizing the singer and understanding the lyrics. The same tasks are much more difficult for an automatic system.

Singing information processing was introduced as a standalone novel area of research [33] to comprise the research which is directly related to processing the singing from the musical signals. A few proposed subcategories include listening to singing voice – for extracting information about the lyrics, singer identity, singer skills and visualization; music information retrieval based on singing – using singing voice timbre, singing query transcriptions; singing synthesis – for speech to singing synthesis or singing to singing voice transformations.

Main research areas and topics

The research and the applications are oriented towards processing of the identity, music and language content of the singing. Classification tasks include classification based on music genre [96], singer identification [30, 102], voice quality classification (professional/amateur) [49], language identification [63, 83, 91]. Other topics related to singing include phonetic segmentation [56], alignment between midi and music [17] or lyrics and music [27, 42]. Singing voice modeling tasks include pitch tracking and analysis, timbre analysis and voice quality classification and transformations [88], melody transcription for query by humming [1] and methods for query transcription and processing [71].

In many cases, the developed methods are initially applied on monophonic singing and then modified for analyzing the singing voice in polyphonic audio. Some classification tasks applied on polyphonic music do not make a distinction between the pure instrumental parts of the song and the parts containing voice, and do modeling of the entire audio [P1]. Others try to retain information strictly related to the singing voice, by preprocessing the audio using methods for singing detection to segment music into vocal and nonvocal parts [6, 75], or sound source separation [9],[P1] to eliminate the influence of the accompaniment.

Methods developed for processing and recognition/classification of monophonic singing will generally have lower performance when used on polyphonic music. At the same time, human performance does not seem to be in any way bothered by the presence of the instruments, as people can still recognize the voice and understand the words. For an automatic system that analyzes the polyphonic music, the complexity of the signal comes as additional challenge, prompting the development of methods for separating the singing from the polyphonic mixture. Some voice separation methods rely on the fact that the voice is usually mixed in the center of the stereo signal [35], while in case of monaural audio, methods based on singing transcription plus sinusoidal modeling [P2],[27] or non-negative matrix factorization [P3] obtain satisfactory results.

Singer identification

The first main topic of this thesis is *singer identification*. Singer identification is part of the more general audio classification and categorization research. The aim of identification follows the human capability of recognizing a familiar voice, in speech or singing. An automatic system capable of performing the same task is desirable for organizing, browsing and retrieving music collections.

Closely related tasks to singer identification are the automatic musical instruments classification and speaker identification. Systems for instrument classification were first trained and tested on isolated notes, and then extended to comprise individual instrument musical phrases and polyphony [36]. Speaker identification is generally applied in monophonic signal conditions, with the voice and various types of noise being present. The singer identification work started by using monophonic singing, and then developed into identification using com-

mercial polyphonic music, where multiple instruments overlap with the singing.

The instrumental accompaniment present in the commercial music is an important challenge to the singer identification systems, as the richness of the mixture makes it hard to identify the spectral components belonging to the singing. Methods for singer identification are using various preprocessing methods for reducing the influence of the instrumental accompaniment on the singer models. Methods range from statistical estimation [69] to feature transformation [90] and vocal line segregation and reliable frame selection [30]. The reported performance of the singer identification systems varies widely because in most cases the evaluation databases are different, comprising usually a small number of singers. Nevertheless, for studies comprising 8 – 20 singers, the identification results are over 80%.

Lyrics transcription

The second main topic of this thesis is *lyrics transcription* from singing. Lyrics transcription is considered to be the most difficult class of speech recognition from a technical point of view. As the lyrics are a very important aspect of the music, it is not surprising that research continues to pursue the problem and resort to various assumptions and simplifications to tailor the task to a specific application.

Intuitively, the task should be approached in a similar way as speech recognition. However, the development of a phonetic recognizer is hindered by the lack of a publicly available database with annotated singing that would be large enough to be used in training the phonetic models. Efforts in building such a database are not of great interest [56]. To a certain degree, the problem can be overcome by using speech data instead of singing for training the phoneme models. The difference between speech and singing is broad enough to cripple the performance of such models in recognizing singing material. Using a small amount of singing data, the suitability of the models can be improved by model adaptation [P4]. Additional improvement can be obtained by imposing constraints depending on the application, for example a small closed set vocabulary or a fixed grammar using a finite-automaton language model [38].

The results in transcription of lyrics from monophonic singing are by no means comparable with speech recognition levels. Depending on the target application, modest word recognition performance can be

useful, as shown in [P5] for song retrieval based on keywords or sung phrases.

When considering the transcription of lyrics in polyphonic music, the problem is much more complex. It seems impossible to overcome the difference from models trained on speech to the properties of singing in a complex mixture. There are however simplified tasks that can be performed by a speech recognition system on polyphonic music, for example alignment between singing and lyrics, imposing restrictions on the search space in the phonetic recognition [27],[P2].

Applications

Singing voice analysis facilitates applications in various areas. Music information retrieval applications include retrieval tasks such as query-by-humming, which retrieves a song based on analysis of the melody, query-by-singing, which retrieves a song based on recognizing the text, or query-by-example, retrieving music based on similarity.

Melody and text transcription of singing provide the possibility to develop human-computer interaction applications for education, such as singing tutors, and various tools for music processing – pitch correction, voice modifications, various applications converting the singing into audio score or textual transcription, management tools for annotating and organizing music collections. Not of least importance is the entertainment sector: karaoke (automatic alignment tools offer the synchronization of music and lyrics similar to karaoke systems) and other singing oriented computer games.

This thesis presents three applications based on lyrics transcription. One is a home-use karaoke system based on the automatic alignment of lyrics with singing. The second application – song/video browsing – is also based on the lyrics alignment, and allows skipping playback to a point of interest based on lyrics or song structure. A third application – query-by-singing – is based on transcription of lyrics from a monophonic singing query, for song retrieval based on the transcribed text.

1.3 Objectives of the Thesis

The main objectives of this thesis are obtaining the singer identity and the words content from a singing voice audio signal. The methods use as input an acoustic signal containing singing voice, and will output

the desired information, that is, the name of the singer, selected from a predefined group, or the textual transcription of the sung lyrics, respectively.

The first objective is to identify the singer in a given piece of audio. The approach follows the general framework of speaker identification, by training individual models for each voice, based on features extracted from the audio. For monophonic singing the task is straightforward. When the tested material is polyphonic music, an additional task is to select information related to singing by eliminating as much as possible the instrumental accompaniment influence, in order to obtain a reliable classification.

The second main objective of this thesis is to develop a system for recognition of the words from a singing voice, based on state-of-the-art speech recognition techniques. The lack of a large database containing annotated singing prohibits training of singing specific models for phonemes. A method for overcoming this problem is to use a speech database for training the models. The task comes to adapting the phonetic models representing speech to the characteristics of singing voice by employing speaker adaptation techniques. A secondary objective to the lyrics transcription is to evaluate the phoneme models in a simplified scenario, for alignment of lyrics and singing from polyphonic music.

1.4 Main Results of the Thesis

The main results presented in this thesis include a method for singer identification in polyphonic music and the construction of a phonetic recognition system for singing voice. Based on the phonetic recognition, the thesis also presents applications related to music information retrieval: an automatic alignment system for singing to lyrics alignment that can be used for browsing within songs and for karaoke production, and a query by singing application that can be used as such or for speeding up a melody-based retrieval system. The thesis presents the following results:

- A study of singer identification in mono and polyphonic music. The classification methods are tested at different singing to accompaniment ratios, investigating the improvement brought by using a voice separation method prior to feature extraction and classification. [P1]

- A speech recognition system for aligning singing to textual lyrics. The speech recognition system is trained using speech data and the lyrics are processed into their phonetic representation for usage with the phoneme models. [P2, P3]
- An investigation of model adaptation for singing phonemes modeling starting from speech data. The adaptation of phonetic models trained on speech was tested using various strategies: global, gender dependent, voice dependent. [P4]
- A complete study of the automatic recognition of lyrics from singing in mono and polyphonic music. A complete lyrics transcription system was constructed, containing phonetic models trained on speech and adapted to singing, plus a language model constructed from lyrics text. [P5, P6]
- A query by singing application using recognized words from a sung query to retrieve songs. The retrieval is based on searching the recognized words in a text database for providing the queried song. [P6]

The results of each included publication are summarized in the following.

[P1] Singer identification in polyphonic music using vocal separation and pattern recognition methods

The publication evaluates methods for singer identification in polyphonic music, based on pattern classification methods and vocal separation. The study investigates the effect of the mixing level difference between singing voice and instrumental accompaniment on the classification performance. In the evaluations it was found that vocal line separation enables robust singer identification down to 0dB and -5dB singer-to-accompaniment ratios.

[P2] Automatic alignment of music audio and lyrics

The publication proposes a method for aligning singing in polyphonic music audio with textual lyrics. Separation of the vocal line is achieved via melody transcription and sinusoidal modeling. The phonetic representation of the lyrics is aligned to the features extracted from the separated vocal line using a phonetic recognizer. The performance of

the method is measured as average absolute error in aligning a test set of commercial music. The errors are calculated at the beginning and end of each line in the lyrics, on over 1000 lines of lyrics text, and the average absolute error is 1.4 seconds.

[P3] Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music

The publication proposes a novel algorithm for separating vocals from polyphonic music accompaniment. The separation is based on pitch estimation and non negative matrix factorization to allow separation of vocals and noise even when they overlap in time and frequency. In the alignment of vocals and textual lyrics, the performance obtained using this method for separation is higher than that of the reference method (presented in [P2]).

[P4] Adaptation of a speech recognizer for singing voice

The publication presents a detailed study of the speaker adaptation techniques that can be applied for adapting a speech recognizer to singing voice. The adapted system is tested for recognition of clean singing and for alignment of singing and lyrics in polyphonic music. Compared to a non-adapted system, various degrees of improvement are obtained with specific transforms, improving the phoneme recognition by 10 percentage units and obtaining an average alignment error under one second.

[P5] Recognition of phonemes and words in singing

The publication presents a study of n-gram language models in the recognition of sung phonemes and words. The song lyrics have very specific language and quite limited vocabulary. Phoneme and word recognition are tested using clean singing and singing separated from polyphonic music. The word recognition performance on clean singing, even being as low as 24%, facilitates retrieval based on the recognized words. In a query-by-singing application based on text search of the recognized words, the first retrieved song was correct in 57% of the test cases.

[P6] Automatic recognition of lyrics in singing

The publication considers the task of recognizing phonemes and words from a singing input by using a phonetic hidden Markov model recognizer. The system is targeted to both monophonic singing and singing in polyphonic music. The recognizer is trained using a speech database and linearly adapted to singing. The paper presents a detailed investigation of adaptation scenarios: global adaptation, gender-specific, singer specific and corresponding recognition performance details. Combined with language models constructed from lyrics text, the recognizer is used for aligning singing to textual lyrics and in a query by singing retrieval application.

1.5 Organization of the Thesis

This thesis is organized as follows. Chapter 2 gives an overview of the singing voice properties and the information it carries. The differences between speech and singing voice are reviewed to highlight that applying the methods from speech research requires some changes for adapting to the singing characteristics. Chapter 3 makes a complete presentation of the singer identification task, features and pattern recognition algorithms used, and their performance in mono and polyphonic music. Chapter 4 presents speech recognition methods for singing voice, starting with the construction of a phonetic recognizer using speech, adaptation of the models to singing voice and language models. The chapter also presents methods for evaluating the performance of the phonetic recognizer in lyrics transcription: a simplified task of aligning singing with given lyrics, and the full task of recognizing phonemes and words. Based on these results, some applications are presented in Chapter 5: a karaoke production tool based on the lyrics alignment, a song browsing application based on word timing information, and a query-by-singing retrieval service based on words recognized in a sung query. Finally, Chapter 6 summarizes the findings and provides some directions for further development of these topics.

Chapter 2

Overview of Singing Voice Properties

The voice is a very important method for communication between humans, and the exchange of information through speech plays a significant role in the cultural and social development of the society. In a simplified view, speech consists of different sounds arranged in adequate sequences to constitute a communication code. Singing, on the other hand, consists of musical sounds, commonly referred to as notes, and, at the same time, consisting of speech sounds that communicate a message. The speech is produced with the purpose of communication, while the purpose for producing singing can be pleasure, comfort, ritual, artistic education or entertainment.

The sounds of speech and singing are produced when an airstream passes through the vocal folds and through the vocal tract. The different structures that are involved in the production of these sounds are the breathing system, the vocal folds and the vocal and nasal tract. Humans can produce a large variety of sounds: speech, singing, whispering, laughing, whistling, the most important of them being speech.

The speech wave carries a large amount of information: the semantic information needed for the understanding of the message, the speaker's vocal characteristics that make us recognize the person speaking, and the emotional state of the speaker. These all blend into the voice and give it the specific characteristics that we are all inherently capable of recognizing.

The exact characteristics of the sounds produced when speaking or singing depend on many different factors. The sound of a voice depends on the individual shape of the vocal tract and of the vocal folds, and on the pronunciation, and it can be changed by education and train-

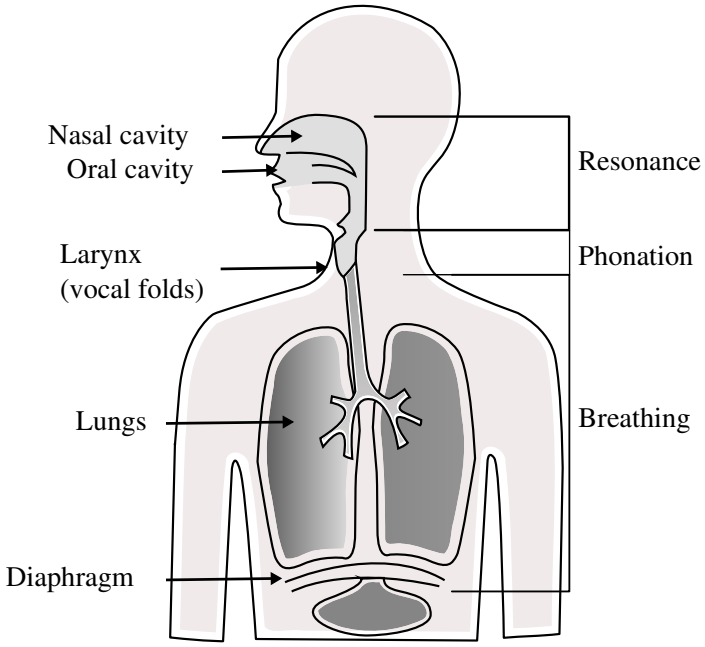

Figure 2.1: Simplified view of the vocal organ: breathing system, vocal folds and vocal tract. The air flow from the lungs is transformed into a tone through phonation, then shaped acoustically through the resonating cavities in the vocal tract

ing. Pronunciation in speech tends to vary according to geographic and sociologic origin, and is perhaps the easiest to change. In singing, pronunciation is usually trained intensively, especially for professional singers. The other important factor is the voice timbre, which is largely determined by the personal characteristics of the voice organ, practically by the shape and size of the vocal cavities.

The vocal organ has three parts: the *breathing system*, the *vocal folds* and the *vocal tract*, each having a specific function. A simplified view of the components can be seen in Figure 2.1. The vocal folds are located in the larynx, at the base of the vocal tract. The opening between the vocal folds is called glottis. The breathing system compresses the air in the lungs, so that a stream of air can be pushed through the glottis and vocal tract. The vibration of the vocal folds generates a sound, which is shaped acoustically when passing through the vocal tract.

The vocal folds open and close the glottis at periodic time intervals, generating a tone, in a process called *phonation*. Tones are voiced sounds characterized by a pitch. They are shaped by the vocal tract

configuration, which is controlled by *articulation*. The vocal tract is a tube-like structure, comprising the throat and the vocal and nasal cavities, and is characterized by resonant frequencies. By keeping the vocal tract in a relatively stable configuration, the vowel sounds are produced. Adjustment of the vocal tract for producing different sounds is called articulation, and the organs which can actively move, such as the tongue, lips, are named *articulators*. The vocal tract resonances and the position of the articulators determine the vowels. Each vowel is therefore associated with a specific shape of the vocal tract and specific pattern of the articulators.

There is a second category of sounds, the consonants, which are articulated with complete or partial closure of the vocal tract. Most of them are unvoiced sounds, that lack pitch, produced by creating a turbulent airflow of pulse-like sounds; for example fricatives are produced when the airflow passes through a constriction created using the tongue or lips, while plosives are created by sudden release of high pressure air. Voiced consonants are called nasals, and have air flowing through the nasal cavity, but not through the mouth. Speech consists of a combination of vowels and consonants, with language dependent ratio, being approximately 60% voiced sounds for the English language [46]. Singing consists of a much higher proportion of voiced sounds, around 90% of its duration, because their harmonic structure can convey the melody. In singing, the vowels get stretched to match the duration of musical notes in the composition.

2.1 Identity and Semantic Content of the Voice

A sung phrase conveys, in the same way as speech, complete information about the identity of the singer and the semantic information present in the lyrics. In addition, musical content is present in the form of melody and rhythm, while emotional content is conveyed by the artistic interpretation.

Both in speech and in singing, the voice identity is largely determined by the characteristics of the vocal folds, the vocal tract and the vocal traits. The additional control required in singing creates a wider variability in pitch and articulation. The vocal tract contributes by its size to identity and through the formants to the semantic content. Studies reveal that the two lower order formants (F1-F2) are most im-

portant for the speech intelligibility, while the higher order formants (F3-F5) contribute to speaker identity information [85]. Additionally, the singing voice, especially in classical singing, can contain *vibrato*, a periodic modulation of pitch that adds artistic expression to the performance. Studies show that vibrato is also related to the identity [64] and singer identification can be achieved based only on features characterizing vibrato [66, 77].

The semantic message in singing, known as lyrics, is the equivalent of its speech content, regardless of the musical information such as pitch or tempo. The information contained in the lyrics is a significant part of the music listening experience.

The organization of the semantic content in singing is the same as in speech, but with a rather specific structure and timing. The words are mapped on a melody, with pitch and duration of the syllables determined by the notes, and the vocabulary is quite specific and much smaller than in speech [P5]. There can be also interjections and filler vowels which do not have any role in the message content (”ah”, ”oh”, ”na na na”).

A more detailed analysis of the differences between singing and speech is presented in the following section, from the voice sound production point of view to the temporal and spectral signal characteristics.

2.2 Speech and Singing: What's Different?

The singing voice is produced in the same way as speech, by the vocal organ, but with added control for producing the musical aspect of the voice. From this perspective, singing can be seen as a special form of speech, having some similar characteristics. At the same time, there are notable differences in the physics of sound production for different phonation types, such as whispering, shouting, normal speech or singing. These differences determine special characteristics of the sounds.

There are many studies about the characteristics of the singing voice, with lots of studies dedicated to soprano voice [3, 40, 64] and other groups of classically trained voice [37], and not so much about commercial pop music. The understanding of the singing voice production mechanism and control is essential in explaining why algorithms developed for speech or music sounds can perform so badly on the singing voice signal. The general properties of the singing voice are pre-

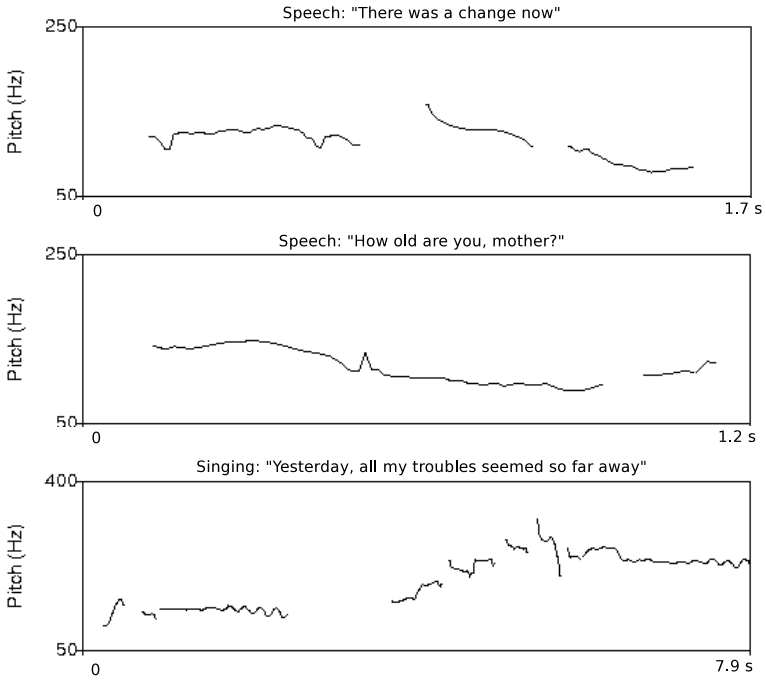

Figure 2.2: Melody in speech and singing. In speech, prosody and accent determine the pitch fluctuation. Differences can be seen between a declarative sentence ”There was a change now” and interrogative sentence: ”How old are you, mother?”. In singing, the pitch changes are determined by musical notes: ”Yesterday, all my troubles seemed so far away”

sented in great detail in [85], and the main differences to speech are reviewed in the following.

Speech has its natural melody called *prosody*, which is different between languages [16]. The spectrum of speech varies freely with prosody, and the pitch contour, loudness variations, rhythm and tempo are used to express emotions. In singing, the singer is required to control the pitch, loudness, and timbre according to the composition. The vowels are sustained much longer in singing than in speech, and independent control of pitch and loudness over a large range is required.

The timbral properties of a sung note depend on the positions of the strong and weak partials, with the strong partials appearing around formant frequencies. The formant frequencies represent resonances of the vocal tract and cavities, and can be controlled to some extent by changing the length and shape of the vocal tract, and the shape and

position of the articulators. The role of the articulators in determining the positions of the formants is well defined, with first formant varying with the jaw opening, second formant determined by the tongue shape, and third formant sensitive to variations in the position of the tip of the tongue. The fourth and fifth formants are less mobile, being mostly determined by the vocal tract length.

The vowels in speech are characterized by specific positions of the formants, especially the first two formants. The values of the formants do not differ very much between individuals, with the formantic regions of speech being more or less the same for a vowel.

In singing, the positions of the lower three formants frequencies can be changed drastically by moving the articulators. Skilled singers can control very accurately the pitch and the formants' frequencies, and different individuals will tune their formant frequencies differently for each vowel. Male singers with low pitch may keep formants approximately in their normal speech positions, but in the high-pitched notes sung by females, articulation must vary with pitch. This is because for example in soprano voices the pitch can be much higher (1000Hz) than the normal speech value of the first formant (500Hz). The first formant will be moved higher in frequency, so that it is approximately equal to the pitch, and the second formant will also move higher. Other resonances are not tuned close to the harmonics of the pitch, except for the first formant [40]. This can cause the vowels to lose their identity, since they are defined by the positions of the first two formants. The fact is, in classical singing the intonation and musical qualities of the voice are the most important aspect, while the intelligibility comes second.

Vowel articulation control in singing is achieved using two different singing techniques: males lower the larynx to control vocal tract length, while females manipulate the first formant by changing the jaw opening depending on the sung vowel and pitch. Singers learn to control the extent of formant changes in such a way that the vowels do not change too much. This is a well known effect in opera singing when sopranos are singing on very high pitches. Even though a vowel will sound very different than in speech, the formant tuning will not change the vowel identity. It will not sound like something else, but it will sound very different than in speech and it will make the words hard to understand.

In singing, the dynamic range is greater than in speech, and also the fundamental frequency variations of singing are much larger, being of about 2 octaves for an average trained singer. More important than range in voice classification is *tessitura*, the region where the voice

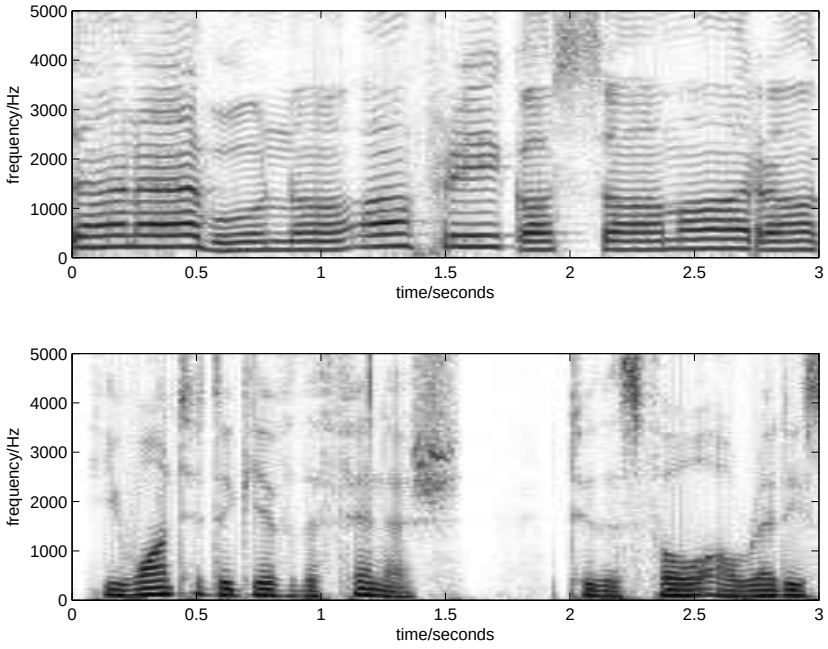

Figure 2.3: Example spectrograms of male singing (upper panel) and male speech (lower panel). In singing the voice contains more continuous blocks, according to notes, whereas in speech the pitch and formants vary more rapidly in time. In speech the amount of unvoiced segments is higher.

is most comfortable singing, and this determines the classification of voice type. The pitch range in a singing phrase is usually higher than in a spoken sentence. In speech the pitch varies all the time, whereas in singing it stays approximately constant during a note, with vibrato being used for artistic singing. This effect is illustrated in Figure 2.3. In singing the variance of the spectrum of a phoneme with a note is smaller compared to speech, while difference between phonemes sung at different pitches can be much larger.

Singing voice timbre shows important differences between male and female singers. One of the important factors contributing to the differences is the voice source, with major difference in the amplitude of the fundamental. In general, the voice spectrum of a male voice has a weaker fundamental than the voice spectrum of a female voice [85]. Different registers of the male voice show the same kind of difference, with modal and chest registers having weaker fundamental than falsetto and middle registers. This brings a large variability between individuals in singing, much more than in speech.

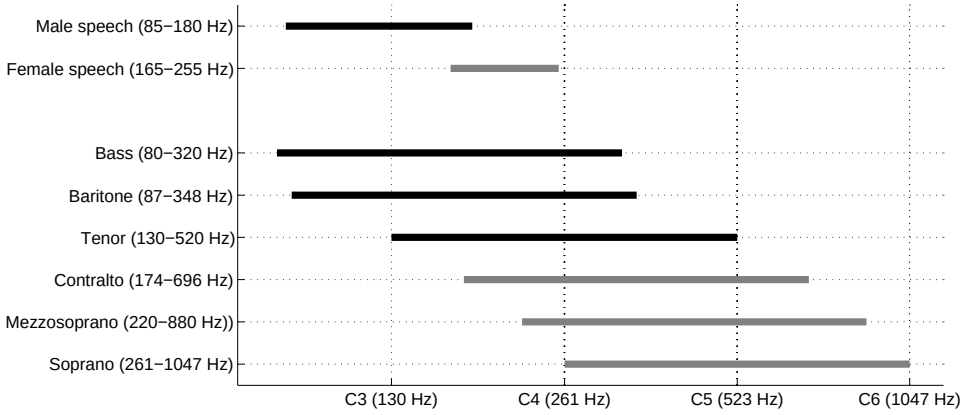

Figure 2.4: Typical range for speech and singing in male (black lines) and female voices (gray lines)

The loudness of phonation is primarily controlled by the subglottic pressure, while the phonation frequency is primarily controlled by laryngeal muscles. In normal speech, the subglottic pressure is much lower than in singing. When singing, the subglottic pressure changes rapidly, and pressure must adapt between adjacent notes for maintaining control over the loudness and pitch. When the loudness is increased, the higher spectrum overtones gain more in amplitude than the lower overtones, therefore a singer must use a voice with a well controlled (ideally constant) content of overtones, to maintain a similar voice quality throughout the phrase while accommodating the changes in pitch.

The subglottic pressure is established by exhaling, therefore for louder phonation, more air is needed. The volume of the lungs is varied within a small range in quiet breathing, but for speaking the volume range is higher, and the speaker needs to take a breath on average every 5 seconds to re-initiate the phonation production. In singing, the phrases are generally longer than in speech, phrases over 10 seconds long are not uncommon, and the pauses where the singer can inhale can be very short. The breathing pattern in singing can be trained, aiming to less air consumption: the less air needed for singing phonation, the more skilled the singer.

In reality, we never think of all these details when we are listening to a singing voice. We simply perceive the properties of the voice, its pitch, the timbre, the loudness, the emotion, the semantic message.

We recognize the characteristics without any special effort, and without considering any objective measures. This makes it difficult to design an automatic system that could react in the same way, based on some numbers representing acoustic features. The struggle for finding human-perception motivated features, algorithms and methods has produced some systems with acceptable performance, but which cannot yet surpass the performance of humans when dealing with the listening experience. The following chapters will give some insight into the signal processing research of singing voice, highlighting methods and challenges for automatic identification of singers and for automatic recognition of the words content.

Chapter 3

Singer Identification

Singer identification is a typical example of supervised classification task: given a set of example signals belonging to different classes, a model is created for each class; when presented with an unknown signal, the classification system uses these models to assign a label to the unknown audio sample. This is the general setup of any supervised pattern classification algorithm: the categories are defined in advance and the model for each category is trained on representative data. In training, a specific model is constructed based on features extracted from the training data, while the testing phase involves assigning one of the predefined category labels to the test data.

In voice identification applications, which include speaker and singer identification, the models must be trained on features that reflect the identity-specific information. The feature sets for singer identification originate from the area of speech recognition. The speech of a person reveals identity-related information through the voice quality, voice pitch, loudness, tempo, intonation, accent, the use of vocabulary. In singing, some of these are imposed by the song, such as pitch, tempo, vocabulary. The most important identity information is related to voice quality, and features representing the static and temporal characteristics of the spectral envelope are mostly used in the classification.

3.1 Features for Singer Identification

The first step in any classification system is to create a representation of the input data. This usually consists in some signal processing methods for extracting a set of features for a discriminative representation of the different categories to be classified. Good features for classifi-

cation are the ones that offer good discrimination of the different categories while having small variability for examples that belong to the same category. Based on the training data, a representative set of features can be chosen, but it is not possible to have a measure of the variability that will be encountered during classification. In the same time, it is desirable to have a compact representation.

In audio processing, the features are typically extracted in short frames, with some degree of overlapping between successive analysis windows. The frame-based features are sometimes averaged over longer windows. Temporal behavior is modeled by derivatives of the static features, typically first and second order only. Various temporal and spectral features have been used in audio classification tasks with variable degree of success. The next sections give a general view of some of them.

Mel frequency cepstral coefficients

The dominant features in speech/speaker recognition are the mel frequency cepstral coefficients [15], abbreviated as MFCC. MFCCs have been frequently used in other music and audio sounds classification tasks for audio content analysis [48]. They encode the rough shape of the signal spectrum in a decorrelated, compact representation. Typically, MFCCs are extracted in frames of 20-30 ms, by applying a windowing function at fixed intervals. Figure 3.1 shows the process of calculating the MFCC features.

Frame by frame, the power spectrum of the signal is calculated using a discrete Fourier transform (DFT). Following the model of human perception of loudness, the logarithm of the power spectrum is used. The resulting spectrum is then transformed from linear to mel-scale frequency. The mel scale is based on a mapping of the frequency into a warped frequency scale that approximates the human auditory sensation of tone height/pitch. The relationship between the frequency in Hz and the mel scale frequency is usually approximated by the equation:

$$F_{mel} = 2595 \log_{10}(1 + \frac{F_{Hz}}{700}) \tag{3.1}$$

The mapping is approximately linear below 1 kHz and approximately logarithmic in the frequency range above 1 kHz. The frequency scale conversion is implemented by triangular band pass filters uniformly distributed on the mel scale, and the transformation of the spectrum into the mel-scale spectrum consists in calculating the power

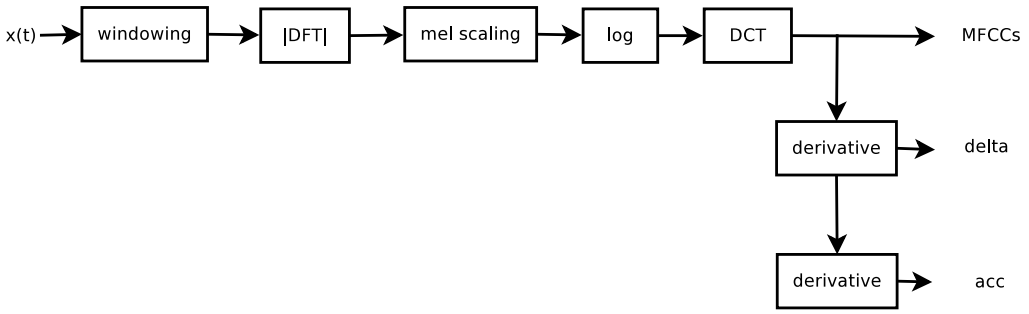

Figure 3.1: Block diagram of MFCC extraction: frame by frame the power spectrum is calculated and redistributed on the mel scale. The mel scaling is obtained by calculating energies in triangular bandpass filters. After logarithmic compression, the cepstral coefficients are obtained using a discrete cosine transform.

within each band. Typically a number of 40 filters are used, resulting in 40 coefficients.

A final step to obtain the cepstrum is the Discrete Cosine Transform (DCT). The DCT decorrelates the feature vectors. The zeroth order coefficient corresponds to the signal energy and is discarded in most applications. The lower order coefficients, usually from 1 to 12-20, are used as features for each frame. Most of the time, the first and second time derivatives are also included in the feature vector.

An investigation on whether the MFCCs are suitable for music classification as they are for speaker identification was presented in [58] in the context of a speech/music classification problem. The study compared linear vs mel scale frequency warping and the DCT as an approximation of the Karhunen-Loeve transform as a decorrelation method. The conclusions of the study are that the mel scale is at least not harmful for being used in music processing and the decorrelation by DCT is appropriate for both speech and music.

Other features

A variety of other spectral and temporal features describing the content of audio are used in audio classification. There are a few variations for cepstral representation, and many features represented by a single scalar. Usually multiple scalar features are collected in a feature vector. A selection of features used in audio classification is shortly presented in the following.

Linear prediction is a powerful tool in speech analysis, and cepstral coefficients derived from linear prediction (LPCC) are an alternative to MFCCs for spectrum representation. Several authors vote in favor of one or the other, claiming better performance of one over the other in classification [7, 30]. However, they both model in some way the power spectrum. Reynolds [78] explains that in fact, if the spectral model does not overly smooth spectral details or provide too much detail to spurious spectral events, then there should be little difference in performance of the features.

Other features used in music and singer classification include, but are not limited to: spectral centroid, spectral rolloff, spectral skewness, spectral kurtosis, zero crossing rate, beat, pitch contour, rhythmic content [95, 96, 102]. Details about these features can be found in literature.

3.2 Classification Methods

In its basic form, a pattern classification procedure consists in the development and usage of a system for classification of some unknown observations into categories, based on a number of given examples of those categories. The development of the system is called *training* and it uses the feature representation of the given examples from the *training set* and their denoted categories to create models of the available data to be used in the actual classification procedure. The *testing set* is represented by a number of unknown observations that will be associated with the existing categories in the *testing* phase.

Supervised classification is a type of classification problem where all the categories are specified in advance and any of the observations in the test set will be classified to one of the pre-defined classes. Ideally, in a supervised classification problem, different classification algorithms would offer the same result, the completely correct classification (100% performance). In contrast, in the *unsupervised* methods, the system forms clusters of the input patterns, with different clustering algorithms leading to different clusters [19].

Supervised classification is suitable for any problem where the categories can be specified in advance. In most situations, for closed set classification in speaker recognition and music classification purposes, the supervised methods are the most used. Methods for classification range from simple and fast to complex ones with big computational

load, depending very much on the problem type, inclusive of the size of the models and the number of classes.

The general approach in training a model based on a set of examples is to specify a model having one or more parameters and estimate the values of the parameters from the training data. For example by setting the form of the underlying probability density, the training samples can be used for estimating the values of its parameters. A wide variety of methods are available for supervised classification, and the most common ones are presented in the following.

Nearest neighbor type of classifiers are simple nonparametric classifiers based on examples. In the training phase, the system is presented with a set of examples that are representatives of each class, and their associated class label, which are all stored. A test observation to be classified is compared with all the available examples and assigned to the class that is represented by the closest training examples. The closeness is evaluated by calculating a distance measure from the test observation to all the available training examples. In the k-nearest neighbor (kNN) classifier, the test observation will be assigned to the class represented by the majority of a number k of the closest points (kNN), with k freely chosen. Distances used for classification can be simple Euclidean distance between feature vectors in the feature space.

The main drawback of the kNN is the storage of all the training examples and the time needed to calculate the distances from the test observation to each of the training examples. When the training set is large, this can be a major computational load. A solution to this is clustering of the training data and storing only the centers of the clusters, or selecting a number of representative examples for each class instead of storing all the examples. The kNN approach was used in singer identification as the main classifier [57], or as a baseline for comparison. In Publication [P1], kNN was one of the classifiers tested in the study.

Linear discriminant functions are relatively easy to compute and are attractive candidates for initial classifiers. The problem of finding a linear discriminant function is formulated as the problem of minimizing the training error – the error when classifying the set of training examples. The discriminant functions produce decision boundaries for the classes. Two classes are said to be linearly separable if there exists a linear discriminant function (decision boundary) that classifies all the samples correctly.

Based on the training data presented as feature vectors $\mathbf{x}$, a set of discriminant functions $L_i = \mathbf{x}^T a_i + c_i$ is constructed, where a_i is a

vector of discriminant coefficients for class i and c is a constant. Given a test observation, the discriminant functions are evaluated and the observation is assigned the to the class having the highest value of the discriminant function. By allowing cross terms, quadratic discriminant functions of the form $L_i = \mathbf{x}^T \mathbf{A}_i \mathbf{x} + c_i$ are obtained ($\mathbf{A}_i$ being a matrix).

Practically the decision boundaries are hyperplanes in the multi-dimensional feature space, and the training stops when the best partitioning of the feature space is obtained, with the least points from the training set being on the wrong side of the decision hyperplane. However, a small training error does not guarantee a small error when classifying unknown test data. In this respect, the classifier based on discriminant functions does not have very good generalization properties, but will give a good picture of the distribution of the data: if the classes are linearly separable, it is not necessary to use a more complex classifier. In Publication [P1], linear discriminant functions were also tested in the study.

Support vector machines are similar to the discriminant functions in the sense of partitioning the feature space with the use of decision boundaries. They aim to find the parameters of the discriminant functions such that the margin between the the decision surface and the training examples is maximized. The training examples that lay on the margin are the only ones that determine the decision surface, and they are called *support vectors*. The basic form of an SVM classifier is a binary classification, but extensions to multiclass problems have been implemented for example by dividing the problem into multiple binary classifications of type one-against-all-others. SVMs are often used in classification because of their ability to deal with cases that are not linearly separable. They have been used in singer identification in conjunction with MFCCs, LPCs or for singing voice detection [46].

Gaussian Mixture Models (GMM) are a popular choice in modeling data, due to their ability to form smooth approximations of arbitrarily shaped densities [79]. A linear combination of Gaussian basis functions has the capability of representing a large class of sample distributions. In addition, the GMMs have good generalization properties when trained on sufficient data.

A Gaussian mixture model for the probability density function (pdf) of x is defined as a weighted sum of multivariate normal distributions:

$$p(x) = \sum_{n=1}^{N} w_n \mathcal{N}(x; \mu_n, \Sigma_n), \tag{3.2}$$

where w_n is the weight of the n-th component, N is the number of components and $\mathcal{N}(x; \mu_n, \Sigma_n)$ is the pdf of the multivariate normal distribution with mean vector μ_n and covariance matrix Σ_n. The weights w_n are nonnegative and sum up to unity. The complete Gaussian mixture model is parametrized by the mean vectors, covariance matrices and mixture weights from all component densities.

In training, the weights, means and variances are estimated using the training data points x represented by the calculated features. The standard procedure to train a GMM is the expectation-maximization (EM) algorithm. The algorithm estimates the parameters of the GMM by relying on unobserved latent variables. In an iterative manner, *expectation* (E) and *maximization* (M) steps are performed, calculating the expectation of the log-likelihood using the current estimate for parameters, and then estimating the parameters by maximizing the expected log-likelihood from the previous step.

In testing, the features of the unknown audio sample will be calculated and the likelihoods of them belonging to the model GMMs are calculated; the class is selected according to the model with highest likelihood. The classification principle in the maximum likelihood classification is to find the class i which maximizes the likelihood L of the set of observations $X = x_1, x_2, ..., x_M$:

$$L(X; \lambda_i) = \prod_{m=1}^{M} p_i(x_m) \tag{3.3}$$

where λ_i denotes the i-th GMM and $p_i(x_m)$ the value of its pdf for observation x_m. The general assumption when using this classification criterion is that the observation probabilities in successive time frames are statistically independent.

GMMs are often used in classification problems due to their capability of representing a large class of sample distributions. Given enough Gaussians, a GMM can estimate any distribution. In speaker or singer identification, it is reasonable to assume that components of the mixture model will be representing broad classes of phonetic events of one speaker [79] or singer. Another advantage of the use of GMMs for classification is that using the EM algorithm it is easy to find the model parameters. Most of the time, models with diagonal covariance matrices are used, because their training is more computationally efficient. For using diagonal covariance matrices, it is preferable that the features used to train the models are not correlated, but this is not entirely necessary, as the density modeling of a full covariance GMM can equally

well be achieved by using a larger order diagonal covariance matrix [80].

As with any parametric classifier, GMMs will have the best generalization performance at the right balance between the accuracy on the training data and the learning capability of the model. Overfitting may occur when the model is unnecessarily complex, because it is trained by maximizing its performance on the training data, while aiming for the best 'working' performance on unseen data. The common practice to avoid overfitting is using a development set for validating the model performance.

Classifiers based on GMMs are the state of the art in many classification problems, including singer identification, singing detection, and speech recognition. Models are usually trained using MFCCs [92], [P1], LPCCs [46, 102], or other features or combinations of features.

3.3 Dealing with Polyphonic Music

When the singing voice signal is available in isolated conditions, the modeling is straightforward, as the features that are calculated from the signal represent entirely the singing voice properties. Monophonic singing databases are usually collected specifically for each study and generally not publicly available.

The large amount of commercial music available to the scientific community and general public at the same time contains singing accompanied by various instruments. There is growing interest in developing methods applicable to polyphonic music, as this has a good potential in offering tools for automatically indexing the continuously growing music collections.

In polyphonic music, the available signal is a mixture of singing voice and various instrument sounds, containing regions where the voice could be missing, and the instruments appear in different combinations. When studying the properties of the singing voice in such a signal, various methods have to be used in dealing with the interference introduced by the instrumental sounds. They can be regarded as preprocessing of the mixture music signal for the desired application of modeling specifically the singing.

A simplistic approach is to ignore the interference and just calculate all the features from the mixture signal. The features would represent the general characteristic of the entire sound, and the models trained using such features might be able to discriminate different

voices. Treating the polyphonic mixture directly and extracting the features for classification relies on the assumption that the singing voice is sufficiently dominating in the feature values.

For applications that consider only information related to singing, it is important to find the segments where singing is present and ignore the instrumental regions in the subsequent analysis. Detection of singing segments within a song is an important subtopic of music research and it was approached in a variety of ways. Furthermore, separation of the singing voice from the mixture musical signal is an important subtopic in sound source separation. The general directions of development in the two areas are outlined in the following subsections.

Detection of singing in polyphonic music

Methods for singing voice detection rely on supervised classification of the signal into vocal/nonvocal segments or on thresholds applied to certain features whose evolution in time correlates to the presence of the singing voice [102].

Classification approaches are based on a wide variety of features that were found to be useful in discrimination of singing voice from polyphonic instrumental background. Unfortunately most of the authors use their own evaluation database and it is difficult to directly compare the results, as in general they do vary with the instrumentation and music arrangement (light vs heavy accompaniment, western vs Asian music, classical music vs pop).

Straightforward classification approaches using cepstral coefficients are used with different classifiers. One example was presented in [20], using MFCCs and a multivariate autoregression mechanism for feature selection for classifying segments of one second length into vocal/nonvocal. Using various classifiers and a database of 147 songs cut into segments of one second, the average results are around 18% error rate. Vocal/non-vocal classification was used also for improving the results of an artist classification system [7], by using 13 PLPs with delta and acceleration coefficients and a multilayer perceptron for detecting the vocal segments.

Berenzweig and Ellis [6] used a speech recognition system to collect a phoneme activation output feature vector to be used in classification. They assume that the speech-trained acoustic models will respond in a detectably different manner to singing, than to instrumental regions. Some combinations of analysis steps outperform slightly a cepstral baseline classifier. The usage of the speech recognition is

an excellent precursor for automatic alignment between known lyrics and the music signal [27]. The authors note that the negative effect of the instrumental accompaniment on the recognition should be levied by using features that can go some way toward separating the singing signal from other sounds.

In contrast with training the classifiers on material consisting of different songs, Tzanetakis presented a semi-automated method for segmentation by training song-specific models [95]. Since the voice of the singer may vary from one song to another due to the song melodic content or singing style, song-specific training data is collected by presenting the user with random snippets from the song for labeling. When an equal number of vocal and nonvocal labels are obtained, a classifier is constructed based on the data and the rest of the song is classified. A number of different classifiers were tested, and the set of features is selected from a list of features used previously by the same authors for music genre classification [96]. The same approach of bootstrapping models using multiple songs instead of song-based performed significantly worse. It is important to note that the annotation of the snippets takes significantly less time than annotating the entire song, and the classification accuracy reached 70% when annotating 24 snippets (48 seconds). One observation in the analysis of the results is that the nonvocal segments are more often missclassified as vocal than vocal being labeled as nonvocal. Generally in a binary classification there is always a tradeoff between precision and recall and the classifier can be tuned to the preferred one.

Another method for dealing with voice variability is multimodel approach: the vocal and nonvocal classes are represented by multiple models, corresponding to different sections and arrangement styles of a song [67]. This means having separate training examples according to section type, tempo and loudness for both vocal and nonvocal, for example having a vocal model for loud, high tempo chorus. The authors of the study used a statistical hypothesis for checking the reliability of the classifier decision in order to select high confidence frames and refine the models. Correct classification results reported on 14 test songs using different types of cepstral coefficients stand between 77.3% (LPCC, no model refining) and 86.7% (harmonic attenuation power coefficients, bootstrapped models).

A second group of methods for singing voice detection in polyphonic music rely on thresholds applied to specific features that are varying with the presence of the singing voice. For example, a twice iterated Fourier transform was used to measure the harmonic structure of each

frame in a song. The sum of the first 50 bins was thresholded for vocal frame detection [60]. The chosen number of bins and the threshold were empirically determined. After correcting some of the segmentation boundaries to the beat structure, the reported performance was 80% frame level correct segmentation.

A threshold-based decision was used also by Regnier and Peeters for detecting singing voice using vibrato, as the expected average vibrato for singers is 0.6 to 2 semitones, while for instruments it is 0.2 to 0.35 semitones [77]. The singing segments were detected by tracking the partials and thresholding the vibrato and tremolo that appears on the partials. The reported results are 76% F-score for detection of the singing class, similar with the performance of a baseline classifier using MFCCs and GMMs.

As a short conclusion of the segmentation problem, it seems that the state of the art performance of systems for vocal/nonvocal segmentation of songs is on average around 85% at frame level, and the choice of features and classifiers does not seem to make much difference. Most of the systems report their performance relative to a baseline classifier constructed using MFCCs and GMMs, and the improvements are generally not impressive. Even with carefully designed statistics, and attempts of exploiting musical characteristics of the singing voice, the automatic segmentation does not match the human annotations.

The inaccuracy of human subjects for annotating the data is neglected. However this does have an influence on the measured performance of the automatic methods, because humans tend to segment music in larger chunks, ignoring short pauses in singing. Such a pause can be a breathing pause at the end of the phrases, or just a longer pause between words. There is no agreed consensus on what is the length (in seconds or number of beats) of a nonvocal segment between two vocal segments that needs to be annotated as such, and not ignored, and therefore there is always some ambiguity in the ground truth annotations.

Separation of the vocal line from polyphonic music

A refinement in modeling the characteristics of the singing voice is to find a way to extract the vocal line from the mixture signal. The separation of sound sources is an important part of the general auditory scene analysis [8], and in this specific case the signal of interest is the singing voice from a musical mixture.

There are a few different methods for voice separation from music, with two main groups depending on the signal, monophonic or stereophonic. Monophonic approaches are based on methods such as blind source separation [97], source-adapted models [68, 69, 93], or segregating the melodic line [30, 53, 54]. Methods for stereo signals are using inter-channel differences in intensity, time or phase, and various refinements [35, 84].

When using blind separation for monophonic mixture signals, a method for identifying and grouping components that belong to the vocal source is needed [97]. Unfortunately in this kind of situation is hard to objectively evaluate the quality of the separated audio.

Segregation of melodic lines does not necessarily refer to singing voice. In [45] a *voice* could represent singing, or an instrument, and chords were also allowed to belong to a voice, after segregation of multiple melodic lines. A method that avoids transcription is presented in [9], by using sinusoidal modeling of a single-channel mixture signal. After onset detection for the sinusoids, the ones that start together were grouped to the same source.

A statistical approach to vocals separation is presented by Ozerov et al., by inferring the model for vocals from models of instrumental only and mixture signal [69]. The general models are learned from recordings different than the ones to be separated, and then the apriori instrumental model is adapted according to the song that is to be processed. The adaptation does the adjusting of the source models with respect to the actual properties observed in the mix, in what is denoted as adaptation with *missing acoustic data* – because the model parameters are estimated from the mixture, whereas the actual acoustic data (the sources) are unknown [68].

Another method to accomplish vocals separation is extracting the harmonic components of the predominant melody from the sound mixture and then resynthesizing the melody by using a sinusoidal model. This method assumes that the predominant pitch is the singing voice. The time frequency components of the detected predominant pitch can be selected using a binary mask and grouped together to form the voice signal [53, 54, 75]. The sinusoidal model can represent the voiced parts of the singing signal, but will have difficulties in extracting the the unvoiced regions. By an explicit step of classification into accompaniment, voiced singing voice and unvoiced singing voice, the unvoiced regions can be modeled and extracted too [39].

The main challenge present in melody extraction is that the accompaniment is usually correlated with the singing, and harmonics of voice

and instruments that are accompanying the singing parts could be overlapping. Other harmonic instruments performing solo parts can be mistakenly extracted as voice. A solution was presented in [76], by tracking two predominant melodies, assuming that there could be maximum one instrument that leads at one time, if not voice. Features based on the temporal instability of singing voice are used to identify the voice pitch out of the two tracked melodies.

Another class of signal separation methods is based on non-negative matrix factorization (NMF) of the spectrogram of the signal. For example in Publication [P3] a binary mask determined based on the pitch of the transcribed melody was used to select the spectral regions where singing is present, and NMF was used for obtaining the separated singing and background accompaniment.

There are a number of approaches based on the stereophonic signal. In most commercial song recordings, the placement of the lead vocal element is at the center of the stereo field and its amplitude usually dominates the other music sources. This is the main element in the approaches for separating singing from stereo music [35, 84]. Generally the methods based on panning fail in time-frequency zones where several sources overlap, because time-frequency points containing energy from both left side and right side sources are erroneously associated with center. Different solutions for better separation combine the panning information with pitch tracking to exploit the fundamental frequency information [10, 13].

The quality of the separation is often subjectively judged, or comparative signal to noise ratios of the monophonic singing and separated voice signals are provided as a measure. It is however hard to interpret and estimate how much the separation quality influences subsequent processing of the separated signal.

3.4 Systems for Singer Identification

The term *identification* explicitly describes the process of naming one out of a group, therefore the methods for singer identification are all part of the supervised classification where models are available for each singer and the decision has to be taken which of the available models is better representing the test data. A clustering application based on singer voice characteristics was formulated by Tsai et. al. for blind clustering of popular music recordings [93]. After a segmentation into vocal and nonvocal segments, a vocal GMM is inferred using feature

transformations from observations of the mixture signal represented by the vocal segments and the background components represented by the nonvocal segments. The clustering method has no knowledge of the number of singers. Ideally if there would be ten singers and the number of clusters is set to 10, the outcome of the clustering should consist of the ten singer classes. The supervised methods are all based on constructing models for voices, using various features.

Monophonic singing is straightforward to classify according to singer. Classification systems using MFCCs and GMMs are of very good performance when using small test databases. Tsai et.al. [89] expand the direct modeling to recognizing simultaneous singers in duet recordings, by training models for each combination of two voices. GMMs with high enough number of mixtures (64) are capable of modeling the classes, obtaining a performance of 100% for recognizing ten solo voices, and 70-80 % for the duets, depending on the partitioning of the test set.

Bartsch and Wakefield proposed a singer identification system based on estimation of the spectral envelope through a composite transfer function. This transfer function is derived from the instantaneous amplitude and frequency of the signal harmonic partials, and is claimed to be better at characterizing vocal variations like vibrato. The test singers are 12 classically trained female singers performing vocal exercises, and the test data consisted of separate vowels. The method was further developed to handle polyphonic music with light accompaniment [5].

Another envelope based method is proposed for detecting segments that the authors name *mp3 phonemes*, in mp3 encoded audio [55]. Segmentation is done using onset detection. Modified discrete cosine transform coefficients computed from the segments are used as features for singer identification, assuming that these are characteristic to each singer. On a set of ten male and ten female singers, the best result is 80%, obtained using 100 nearest neighbors. There is no mentioning if the music is mono or polyphonic.

Polyphonic music is harder to deal with. When the usage of polyphonic music is explicitly mentioned, there are a variety of methods for selecting the voice-related content. Preprocessing for selecting the voice components is based on segmentation, statistical estimation of the voice models from the available data, feature transformation or vocal line separation.

One approach for endpoint detection is presented in [102], by following the values of energy (start of the singing is a sudden rise in energy), spectral flux (high peaks when singing starts), zero-crossing rate

(high amplitude ZCR peaks produced by consonants) and harmonic coefficients (high values with singing) to decide where the singing starts. At the point where singing is detected, a 25 seconds fragment was selected and represented using frame based cepstral coefficients derived from LPC of order 12, and used to train GMMs for singers. It is quite a bold assumption that the 25 seconds following detection will contain voice, but nevertheless the performance of the method is 82% in classifying 45 songs belonging to eight singers. The authors observe that the lighter the instruments sounds, the better the results are.

Another example of a method for selecting the voice content from the mixture is band filtering of the signal [46]. The mixture signal was filtered using a bandpass filter to retain the content between 1200 Hz and 2 kHz, considering that the majority of the singing voice energy falls in this region. Voiced frame detection is accomplished by using a measure of harmonicity of the filtered signal. The classifier was constructed using warped and linear LPCs and GMMs and alternatively SVMs, for classifying 17 singers. Highest performance was obtained using SVMs, being 41.5% when using only the vocal frames, 54.3% when using the entire song data. The author notes that given this difference in performance, practically it is unclear whether the classifier is actually training on vocal features or is using some other aspect of the recordings, which is given by the specific instrumental accompaniment.

Singer identification was used as a way to improve the results of an artist classification system [7]. The artist classification as such was regarded as identifying the band or the singer performing the song. The authors selected the most vocal-like segments in a song based on a vocal/nonvocal classifier, and the artist classification system was constructed using only the vocal parts, in contrast of using the entire length of audio available. The vocal segmentation was done using multi-layer perceptron and cepstral coefficients derived from LPC, and the singer identification was done using multi-layer perceptron and MFCCs. The same study brings into attention the *album effect*: when using different albums of an artist in training and testing, classification using selected voice segments did not outperform the classification using the entire song. The album effect has been noticed and studied by other authors as well [47], concluding that the effect stems from the remastering, as the postprocessing of the recordings usually does dynamic range compression or expansion to make the songs sound the same.

There are a number of studies that perform singer identification by using both the singing segments and the instrumental segments of the songs, as a way of including as much information as possible in the classification. Practically this would be defined as the more general artist identification task, but always using music that contains vocals. One such system for singer identification based on vocal and instrumental models [61] does the segmentation of the music piece into instrumental and vocal using beat length segments, on the assumption that the acoustic properties vary more likely in inter-beat intervals. The voice identification performance of 81% is boosted to 87% when adding the content of the instrumental segments. Another system using multiple vocal and instrumental models and vibrato related features is presented in [66]. Vibrato is extracted using triangular subband filters on each note with a +/- 1.5 semitone bandwidth. This work highlights the album effect too by testing how the performance deteriorates when using for training recordings from 2004 and a test set of music recorded in '85-'95.

In [92], a method for statistically inferring the vocal model from the mixture was presented, in a study using mandarin pop music. The authors assume that the stochastic characteristics of the background music could be approximated by those of the instrumental-only regions in a music recording. Having the background music information, the characteristics of the singer's voice can be estimated from the regions containing singing and accompaniment. Using 20 MFCCs calculated in 32 ms Hamming-windowed frames with a 10 ms frame shift, the obtained performance is 66-100% individual identification for 20 singers. An observation made by the authors is that mandarin and other Asian pop music often sounds like the vocals are mixed louder than in western music, and the results should be compared by applying the method on Asian and western pop music.

The same authors published methods for estimating the characteristics of singing voice using feature transformation applied to a signal related to singing voice. In [90], singer characteristics were modeled by estimating a transformation of the cepstrum from accompanied to solo voice, using a large set of paired data, solo and accompanied singing generated by manually mixing the solo with the instrumental accompaniment. Using the learned transforms, the cepstrum of a voice plus accompaniment signal was transformed into a solo-like voice cepstrum. The method presents a small improvement from their previous system that infers the vocal model from the mixture and background. The disadvantage of the cepstrum transformation method is that it needs

parallel data in order to estimate the relationship between solo-like singing and the available signal, making the implementation of a large system practically impossible.

A large number of studies related to singing voice originate from Fujihara et. al. [22, 25, 30], and a method for accompaniment reduction and reliable frame selection is their choice for dealing with the influence of the musical accompaniment. The accompaniment sound reduction consists in estimating the fundamental frequency of the melody, extracting the corresponding harmonic structure, and resynthesizing the signal corresponding to the melody using a sinusoidal synthesis. After obtaining the segregated vocals, a small number of frames are selected as representative for the singing voice. The reliable frame selection avoids segmentation into vocal/nonvocal segments, and it selects frames that have the highest likelihoods to be vocal, according to a threshold that depends on the likelihood ratio of a vocal/nonvocal classifier. The frame selection step discards most of the content for each song, by using only 15% of the total frames for constructing singer models.

This procedure was used for singer identification [30] and information retrieval based on singing voice timbre [22, 25]. The authors choose mel-cepstral coefficients of LPC spectrum (LPMCC) for representing the spectral shape of the singing signal, and claim these are much better for representing information related to identity than the classic MFCCs. However, they compare the performance of the system using accompaniment reduction and LPMCCs to a baseline system in which the calculation of MFCCs does not use the accompaniment reduction step. Using a test set of 40 songs by ten different singers, the performance of the system using accompaniment reduction and frame selection is 83% compared to 75% of the baseline using MFCCs. The gap in the results is present also in the song retrieval [25], as in the retrieval system based on MFCCs the song similarity is better, while in the one with LPMCC the voice similarity is better.

The proposed system

The system presented in [P1] considers a number of aspects of the singer identification problem. The study includes different classification methods, single and multi-model singer class modeling. Novelty of the study includes an investigation of the influence on performance

of different signal to noise ratio of the singing versus accompaniment, and performance improvements through singing voice separation.

The features chosen for representing the spectral shape of the signal are the MFCCs, as according to [78] there shouldn't be any significant difference between the cepstral representations, as long as they are not too coarse or too detailed. The MFCCs do not make any assumptions about the nature of the signal, and the ease of the calculation is attractive.

Different methods tested for classification include discriminant classifiers, Gaussian mixture models using maximum likelihood classification and nearest neighbor based on Kullback-Leibler divergence between GMMs. The discriminant classifiers give a general baseline performance, describing the difficulty of the problem. In the nearest neighbor classifier, a singer class was represented by one GMM or by a collection of GMMs with each one representing one song of the singer. Modeling each song by a separate GMM allows representation of more details, useful for cases in which a singer uses different singing style for different songs. The symmetric Kullback-Leibler divergence was used as a measure of the distance from the test data GMM to the singer models, in order to classify the test data using kNN.

The previous studies have observed that the performance of singer identification systems was higher when the accompaniment was lighter [102], or that Asian music is mixed so that the vocals are quite loud [92]. The obvious way to study the degradation of performance that is caused by the relative level of the accompaniment was to create mixture signals at different signal to noise ratio, from a 30 dB level, where the accompaniment is barely noticeable, to -5 dB which is close to the way commercially available western music sounds. To avoid modeling of characteristics causing the album effect, the vocals were mixed with midi synthesized accompaniment; this way all the songs and singers in the database had similar accompaniment.

The separation of vocals from the mixture signal was achieved using a melody transcription system [81] followed by sinusoidal modeling resynthesis. Within each frame, the melody transcriber estimates whether significant melody line is present, and estimates the MIDI note number of the melody line. For resynthesis of the melodic line, harmonic overtones were generated at integer multiples of the estimated fundamental frequency. Amplitudes and phases were estimated at every 20 ms from the polyphonic signal by calculating the crosscorrelation between the signal and a complex exponential having the over-

tone frequency. The synthesized signal was obtained by interpolation of the parameters between successive frames.

Summary

The performance of all the classifiers degrades visibly with the decrease of the SNR, but preprocessing of the audio by vocal line separation helps the systems to gain back some of the performance lost due to polyphony. The multimodel approach shows some flexibility in modeling, allowing more variation in the modeling of each singer.

The vocal separation seems to be a successful method in dealing with the instrumental accompaniment. When the voice is mixed at much higher level than the accompaniment, the separation is not that crucial, however for mainstream music it is really important that we are able to separate the singing signal so that we extract the features only for the voice characteristics.

Chapter 4

Speech Recognition Methods for Singing Voice

The lyrics of a song are as important as the melody in providing the listener with the full music listening experience. For automatic analysis, lyrics have been lagging behind melodic information for the simple reason of being difficult to deal with. However, it is evident that information retrieval based on lyrics has a significant potential. The extensive online databases with lyrics offer the possibility of finding the lyrics of a particular piece of music when knowing the name of the artist or name of the song, or finding the song and artist based on a fragment of the lyrics.

Lyrics recognition from a song would allow searching in audio databases, by automatically transcribing the lyrics of a song being played, or by providing automatic indexing of music. For now, the most popular form of audio based retrieval is query-by-humming, using the melody information [94]. Only very few works include the lyrics information into the retrieval, by using a constrained grammar for the lyrics, provided as a finite state automaton with predetermined evolution [82, 87]. Other works related to retrieval using lyrics do not include the aspect of recognizing them from singing, but only processing of text information [26, 65].

Transcription of lyrics from singing voice has not received much attention in the speech recognition community, as it is difficult to accomplish. The main reasons for this are the lack of databases specifically collected for training singing phonemes, the large variability of the singing phonemes themselves, and not least, the polyphonic nature of music in general. Recognition of phonetic information in polyphonic music was studied as phoneme recognition in individual frames [34],

but there is no significant work done using large vocabulary recognition of lyrics in English. A simpler task, alignment of audio with textual lyrics, has been implemented by different authors using a speech recognition system and known text to be synchronized with the singing [23, 27, 59, 100], [P2].

The objective of developing a lyrics transcription system is certainly over-optimistic, but some key issues in approaching the problem are within possibility. The basis for the techniques can be found in automatic speech recognition. Despite the differences between singing voice and spoken voice, highlighted in Chapter 3, it is possible to use the speech recognition techniques on singing, with some adjustments, as will be presented in the following.

4.1 Phonetic Speech Recognition

Automatic speech recognition is a well established area of research. Multiple decades of research have produced a state-of-the-art approach for speech recognition based on hidden Markov models (HMM). The speech recognition system structure contains an acoustic component and a language-dependent component. The acoustic component contains the basic pattern recognition problem, which deals with recognizing the phonemes based on models constructed from audio training data. The language component uses a statistical representation of the possible combinations of words in the recognition task. The language model parameters and the acoustic model parameters are considered to be independent. When performing recognition, the likelihoods from the acoustic models and from the language model are combined to provide the final result.

HMM phoneme models

State of the art speech recognition relies on hidden Markov models for representing the speech unit models. The speech unit models are chosen based on the application and they represent the categories for the pattern recognition system. In most cases the speech unit models are words or subword units such as syllables, phonemes or triphones, but they can as well be sentences or other linguistically meaningful units. HMMs are a widely used statistical method for characterizing the spectral properties of the frames in a time-varying signal. The underlying assumption of parametric models is that the signal can be

characterized by a parametric random process and that the parameters of this process can be estimated in a well defined manner [74, p. 322].

A hidden Markov model is a finite state machine consisting of a number of states with associated observation probability distributions and a transition matrix defining transition probabilities between the states. The emission probability density function of each state is modeled by a Gaussian mixture model.

An HMM is specified as:

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi). \tag{4.1}$$

$\mathbf{A} = \{a_{ij}\}$ is the state transition probability distribution that defines the transitions between states, $\mathbf{B} = \{b_j\}$ is the observation probability distribution that defines the symbol distribution in each state, and π is the initial state distribution.

Given the state q_t of the machine at time t, the state transition probability matrix $\mathbf{A}$ consists of elements a_{ij} defined as:

$$a_{ij} = p(q_{t+1} = j | q_t = i). \tag{4.2}$$

The actual state of the machine at a given time t cannot be observed directly, but can be estimated based on an observation vector o_t and its state-conditioned likelihood distribution.

The three problems associated with HMMs are [74]:

1. Given the observation sequence $O = (o_1, o_2, ..o_T)$ and a model $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, how to efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model?

2. Given the observation sequence $O = (o_1, o_2, ..o_T)$ and the model $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$, how to choose a corresponding state sequence $q = (q_1 q_2 .. q_T)$ that is optimal in the sense of best explaining the observations?

3. How to adjust the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$ to maximize $P(O|\lambda)$?

The training of the models is accomplished by using the solution to the third problem in the list. The observation sequence used to optimally estimate the models is the training sequence, and the problem is solved by using Baum-Welch reestimation [73]. The solution to the first problem allows computing the probability that the observed sequence was produced by the model, and comparison of different models for choosing the model that best matches the observations.

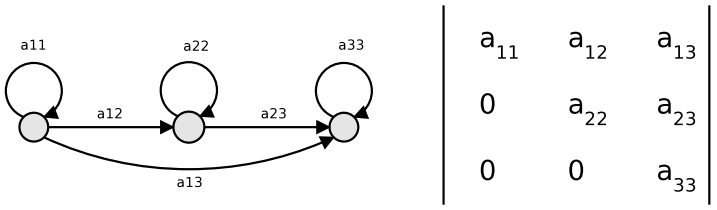

Figure 4.1: Left-to-right HMM and corresponding transition matrix: $a_{ij} = 0$ for $i > j$.

The solution to the second problem is the core of the speech recognition process, solved by the Viterbi algorithm [21], by finding a single best path $\hat{q}$ that maximizes the total observation likelihood

$$\hat{q} = \arg\max_{q}\{p(q|O, \lambda)\}. \tag{4.3}$$

In HMM-based speech recognition it is assumed that the observed sequence of speech feature vectors is generated by a hidden Markov model. The preferred model for the application is a left-to-right model, as this type of HMMs directly explains the sequentiality of the speech signal. In the training process, the transition matrix and the means and variances of the Gaussian components in each state are estimated to maximize the likelihood of the observation vectors in the training data.

A basic speech recognition system for English language can be constructed using the set of phonemes present in the language. One such set is the CMU phoneme set[1], based on the ARPAbet, consisting of 39 phonemes, presented in Table 4.1. By use of a pronunciation dictionary, all the words can be represented as sequences of these phonemes. Each phoneme is represented by a left-to-right HMM, and a silence model represented by a fully connected HMM is added to complete the system. A complete toolkit[2] for speech recognition using HMMs is provided by Cambridge University Engineering Department (CUED).

The typical features used in speech recognition for training phoneme models are the mel frequency cepstral coefficients. Other features such as linear prediction coefficients, filterbank coefficients have been used by different authors, but they did not achieve higher performance [78]. The usual MFCC feature vector contains the lower order static coefficients 1 to 12-13, along with their delta and accelera-

[1]Carnegie Mellon University Pronouncing Dictionary

[2]The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/

Phoneme	Word Example	Phonetic transcription
AA	father	F AA DH ER
AE	at	AE T
AH	sun	S AH N
AO	off	AO F
AW	how	H AW
AY	my	M AY
B	buy	B AY
CH	chair	CH EH R
D	day	D EY
DH	that	DH AE T
EH	men	M EH N
ER	her	HH ER
EY	say	S EY
F	for	F AO R
G	go	G OW
HH	he	HH IY
IH	big	B IH G
IY	bee	B IY
JH	just	JH AH S T
K	key	K IY
L	late	L EY T
M	me	M IY
N	no	N OW
NG	sing	S IH NG
OW	show	SH OW
OY	toy	T OY
P	pay	P EY
R	run	R AH N
S	say	S EY
SH	show	SH OW
T	take	T EY K
TH	through	TH R UW
UH	could	K UH D
UW	you	Y UW
V	very	V EH R IY
W	way	W EY
Y	yes	Y EH S
Z	zero	Z IH R OW
ZH	measure	M EH ZH ER

Table 4.1: CMU phoneme set and word examples

tion coefficients. The zeroth order coefficient is usually discarded, as it represents the energy of the signal. The features are calculated in short frames of 20-25 ms, with some degree of overlap between adjacent frames.

Databases used for training the phoneme models vary in size and content, depending on the purpose of the application. Speech databases can consist of readings of isolated phonemes or words, reading text fragments, or spontaneous speech, with a number of speakers ranging from few to many (over 50).

N-gram language models

The linguistic information in speech recognition is represented using *language models*. A language model consists of a *vocabulary* – a set of words that can be recognized by the system, and a set of rules describing how these words can be combined into sequences. The vocabulary can be defined at different abstraction levels, by using units such as phonemes, syllables, letters, or words. The language model works by restricting the possible unit sequences from which the speech recognition system has to choose the most probable one. The mechanism is that the model provides probabilities for different sequences, and these probabilities are used together with the likelihoods of the acoustic models to find the most likely phonetic sequence for an input signal.

Language models for speech recognition are constructed using n-grams to model probabilities of word sequences. N-grams are probabilistic models for predicting the next item in a sequence. In probabilistic terms, an n-gram language model provides conditional probabilities $P(w_i|w_{i-1}, w_{i-2}, ...w_{i-n})$, meaning that it uses the previous $n-1$ words $w_{i-1}, w_{i-2}, ..., w_{i-n}$ to obtain the probability of the next word w_i [41]. The probability of a whole sequence can be obtained as the product of above conditional probabilities over all units in the sequence.

An n-gram of size one is referred to as *unigram*, size two is a *bigram*, size three is a *trigram*, while those of higher order are referred to as n-grams. Bigrams and trigrams are the ones commonly used in automatic speech recognition.

The language models are estimated from text databases comprising as many examples of n-grams as possible. If the number of occurrences of the sequence of three words $w_{i-2}\ w_{i-1}\ w_i$ and the sequence of two words $w_{i-2}\ w_{i-1}$ are $C(w_{i-2}w_{i-1}w_i)$ and $C(w_{i-2}w_{i-1})$, then the conditional

probability is estimated by the relative frequency:

$$P(w_i|w_{i-1}w_{i-2}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} \tag{4.4}$$

One problem of language models is that not all the possible combinations of words will be encountered in the training database enough times for a reliable estimation of the conditional probabilities. Methods for smoothing the conditional probabilities exist, such as introducing a small probability for any unseen combination. A back-off mechanism is another method for smoothing, such that in case the n-gram conditional probability is not reliably estimated (the n gram was seen less than k times), the system will use the conditional probability of the $(n-1)$-gram.

Another problem of language modeling is that it is not possible to include all possible words in a language model. The percentage of out of vocabulary (OOV) words affects the performance of the speech recognition system, since the system cannot output them. Instead, the system will output one or more words from the vocabulary that are acoustically close to the word being recognized, resulting in recognition errors. While the vocabulary of the speech recognizer should be as large as possible to ensure low OOV rates, increasing the vocabulary size increases the acoustic confusions and does not always improve the recognition results.

The quality of a language model is assessed by using a measure called *perplexity*, which measures the uncertainty in each word based on the language model. Perplexity can be seen as the average size of the word set from which a word recognized by the system is chosen [74, p. 450]. Lower perplexity means better accuracy of the language model in representing the text. An ideal language model should have small perplexity and small out-of-vocabulary percentage on an unseen text.

In a large vocabulary speech recognition system, the goal is to decode the word string W based on a given observation sequence O, such that the decoded string has the maximum aposteriori probability [74, p. 424]:

$$P(\hat{W}|O) = \max_{W} P(W|O) \tag{4.5}$$

According to Bayes's rule, the right side of the equation can be written as:

$$P(W|O) = \frac{P(O|W)P(W)}{P(O)} \tag{4.6}$$

Because $P(O)$ is independent of the decoded string, the decoding rule becomes:

$$\hat{W} = \arg\max_{W} P(O|W)P(W) \tag{4.7}$$

The first term is given by the acoustic model – the probability of the acoustic observation sequence given a word string. The second term is given by the language model – the probability associated with the word sequence. The balance of the two components in the Viterbi decoding can be controlled using a grammar scale factor and an insertion penalty parameter. The grammar factor multiplies the log likelihood of the language model to balance the acoustic and linguistic information influence on the result. The number of words output by the recognizer can be controlled by the word insertion penalty which penalizes the log likelihood by adding a cost for each word [74, p. 454]. The values of these parameters have to be tuned experimentally to optimize the recognizer performance. The preferred values are at the equal error rate (EER) balance point, where the number of insertion errors and deletion errors in recognition is nearly equal.

A language model for singing

For a good coverage of the vocabulary and words combinations in a specific application, the language model should be trained on text with the same topic as the speech recognition application that the system is expected to perform well on. In this respect, singing is quite specific in its choice of words, and the vocabulary of the language model can be chosen as the most frequent words from lyrics text. As presented in [P5], a moderate size vocabulary of 5000 words is representative enough for popular music. The lyrics text of 4470 songs[3], containing over 1.2 million word instances, contains approximately 26000 unique words, out of which only 5167 words appeared at least 5 times. A language model constructed on this data should reflect well the lyrics of other popular songs.

For the more general purpose of phoneme recognition from continuous singing, a language model representing the possible sequences of phonemes can be built from any text. In every language there are some restrictions on the permissible combinations of phonemes. The permissible syllable structure, consonant clusters and vowel sequences are

[3]retrieved from http://www.azlyrics.com/

language specific, making the n-grams suitable for language identification. We can therefore assume that a phoneme-level language model is characteristic to English language, no matter if estimated from general text or from lyrics. This is confirmed by the study in [P5], by constructing a phoneme level bigram and trigram using a training database of phonetically balanced sentences with over 48000 phoneme instances. Perplexities of this language model on the test data consisting in lyrics text is very similar with its perplexity on the training data. For a phoneme language model there is no concern over OOV, since all the phonemes are included as units in the vocabulary.

4.2 Adaptation to Singing Voice

An important part of building a speech recognition system is the training database, as the training material should be large and diverse enough for a reliable estimation of the parameters of the acoustic models. For constructing acoustic models for singing, the lack of an annotated database for this purpose makes the problem more difficult.

Publication [P4] presents the methodology and few combinations for adapting phonetic models trained on speech material to cope with the variability of singing voice. This is a solution used in speech recognition research, and its use in speech recognition refers to adapting speaker independent models to the speech characteristics of one specific speaker in order to have higher performance in recognizing the speech of that speaker. In the same way, models trained using speech acoustic data can be adapted to the characteristics of singing voice.

The acoustic material used for the adaptation is called *adaptation data*. In speech recognition the adaptation data is typically a small amount of speech from the target speaker. The adaptation is done by finding a set of transforms for the model parameters in order to maximize the likelihood that the adapted models have produced the adaptation data. When the phoneme sequence is known, the adaptation is done in a supervised manner, by providing the system with the correct transcription.

The most commonly used technique for model adaptation is maximum linear likelihood regression (MLLR) [51]. Given the mean vector $\boldsymbol{\mu}$ of a mixture component of a GMM, the MLLR estimates a new mean vector as

$$\hat{\boldsymbol{\mu}} = \mathrm{A}\boldsymbol{\mu} + \mathrm{b}, \tag{4.8}$$

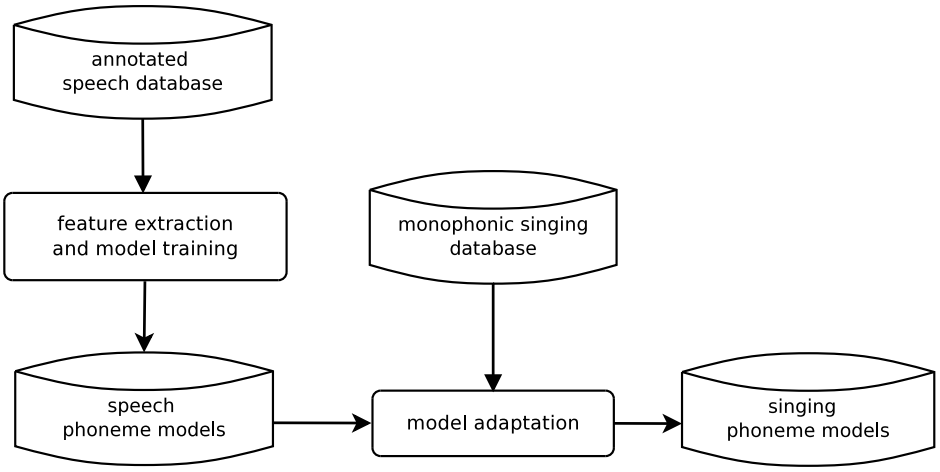

Figure 4.2: Obtaining the models for use with singing data. Phoneme models are trained using speech data and adapted using a small amount of singing data

where $\mathbf{A}$ is a linear transform matrix and $\mathbf{b}$ is a bias vector. In constrained MLLR (CMLLR) [31], the covariance matrix $\mathbf{\Sigma}$ of the mixture component is also transformed as

$$\hat{\mathbf{\Sigma}} = \mathbf{A}\mathbf{\Sigma}\mathbf{A}. \tag{4.9}$$

The transform matrix and bias vector are estimated using the EM algorithm. When the same same transform and bias vector are shared between all the Gaussians of all the states, the method is called *global adaptation*. If enough adaptation data is available, multiple transforms can be estimated separately for sets of states or Gaussians. The states or Gaussians can be grouped either by their phonetic similarity or their acoustic similarity into groups called *base classes*.

Global adaptation is the most suitable method when only a small amount of adaptation data is available, as it will result in a robust transform [52]. When a sufficient amount of adaptation data is available, the adaptation can be done in two passes, using a global transform, followed by a second transform with more classes constructed in a data-driven manner, by means of a regression tree [72]. It is possible to use predefined base classes. For example, base classes can be defined according to the broad phonetic classes: monophthongs, diphthongs, approximants, nasals, fricatives, plosives, affricates [52]. Few possibilities of grouping sounds by phonetic similarity into base classes for adaptation are presented in Table 4.2. Base classes defined by acoustic similarity are determined by clustering the Gaussians of the states.

Number of classes	Classes
3	vowels consonants silence/noise
8	monophthongs diphthongs approximants nasals fricatives plosives affricates silence/noise
22	one class per vowel approximants nasals fricatives plosives affricates silence/noise

Table 4.2: Divisions of phonemes into classes by phonetic similarity.

Gender-Dependent Adaptation can be used as a refinement of the phoneme models, as the gender differences in singing voices are very prominent. Gender dependent phoneme models are common in speech recognition too [101], even though in speech the differences between male and female voices is not that prominent as in singing. It is beneficial especially to the female voices to have the gender dependent adaptation [P4], as in general, the recognition systems perform visibly worse when tested on female voices.

Singer-Specific Adaptation can be used for adapting the models to a specific voice. This is closer to the speaker adaptation procedure, when a speaker independent set of phoneme models are adapted to a specific speaker. The amount of available data and the degree of model mismatch is important, as in the results presented in publication [P4], singing adapted models that were further adapted to a specific singer did not perform any better than the initial singing adapted models.

4.3 Automatic Alignment Between Lyrics and Singing

The automatic alignment between lyrics presented as text and singing from a polyphonic audio piece can be seen as a simplified phoneme recognition task for singing, in the sense of forced alignment of text to speech. However, alternative methods for solving the alignment for music have been proposed in literature, by exploiting characteristics of music, such as song structure, melody and the language.

Dynamic time warping (DTW) has been used for aligning signals or symbolic representations. Authors of [50] use DTW to align a hand labeled structure of the text lyrics with labels resulting from an automatic song segmentation algorithm. In [62], the lyrics are synthesized using a text-to speech system, using a duration of one quarter note per syllable and C4 for pitch, then the two signals, the original song and the synthesized singing, are aligned using DTW.

Wong et al. [100] present alignment of a music signal with singing in Cantonese to a corresponding lyric file. A preprocessing method is used to enhance the singing voice and to suppress background music. The features for alignment are based on the tonal characteristics of Cantonese language, where for each syllable a certain pitch is used. Such a method cannot be generalized to non-tonal languages such as English.

A system exploiting song structure is presented in [42]. The system aligns first the higher level structure of a song by locating the chorus and vocal sections. Within the boundaries of the detected sections, a line-level alignment is performed, based on an uniform estimated phoneme duration. The system does not do phoneme recognition, it just searches for a target number of vocal segments, corresponding to the number of lines in the corresponding lyrics section. The method is based on strong assumptions about the structure and meter of the song and is limited to certain types of songs.

The solution borrowed from speech recognition is to create a phonetic transcription of the word sequence comprising the text in the lyrics and to align the resulting phoneme sequence with the audio using a speech recognition system. This simplifies the phoneme recognition task because the system is given the complete transcription. This way, the possible paths in the Viterbi search algorithm during recognition are restricted to just one string of phonemes, representing the input text. The solution to the decoding is the temporal relationship

between the two inputs, providing the timestamps of the phonemes in the audio signal. The methods relying on phoneme models have been applied both on monophonic singing and polyphonic music. One of the first works highlights low-delay alignment [59], focusing on the Viterbi decoding in almost-real time, after a few frames.

A complete system for synchronizing lyrics with music CD recordings is presented in [27], by using the method for segregating vocals and a vocal activity detection [30]. The phoneme models of ISRC [43] software are used as initial models and adapted to singing voice, then used for forced alignment. The (Japanese) lyrics are transformed into a sequence of phonemes, and only the vowels are retained for the alignment process. The alignment performance is evaluated at phrase level, a phrase being defined as a section that was delimited by a space or line feed in the original lyrics. Alignment accuracy is given as the proportion of the length of the sections which are correctly labeled to the total length of a song. An average of 81% is obtained on 10 test songs.

A further refinement follows this system by introducing fricative and filler model recognition [23, 28]. The fricatives (actually only 'sh') are forbidden to be aligned to certain regions where it is confirmed that they cannot exist, however this does not bring improvement for the line level evaluation of the alignment performance. A novel method for vocal/instrument classification brings some improvement in performance, due to more accurate detection of vocal regions. When using songs with parts of lyrics in English, the English phonemes were approximated with the closest Japanese phonemes. There is some limitation in using this approach for English language, given the differences in the phonetics of the two languages: in Japanese all vowels are pure, there are no diphthongs, only monophthongs, and phonetic transcriptions of English using these is very approximate.

Publication [P2] presents a lyrics alignment system based on phoneme models, applied to polyphonic music. The method uses the complete set of phonemes for English language (the 39 phonemes from CMU), while the previously mentioned studies were based on a subset of Japanese phonemes (five vowels). The models are trained using a speech database and adapted to singing using MLLR. As a preprocessing of the audio signal, vocal line separation is performed using melody transcription and sinusoidal synthesis of the main melody. The acoustic features used in alignment are calculated from the separated vocals.

The lyrics input is transformed into a sequence of words that will be used by the recognizer. An optional short pause 'sp' is inserted between

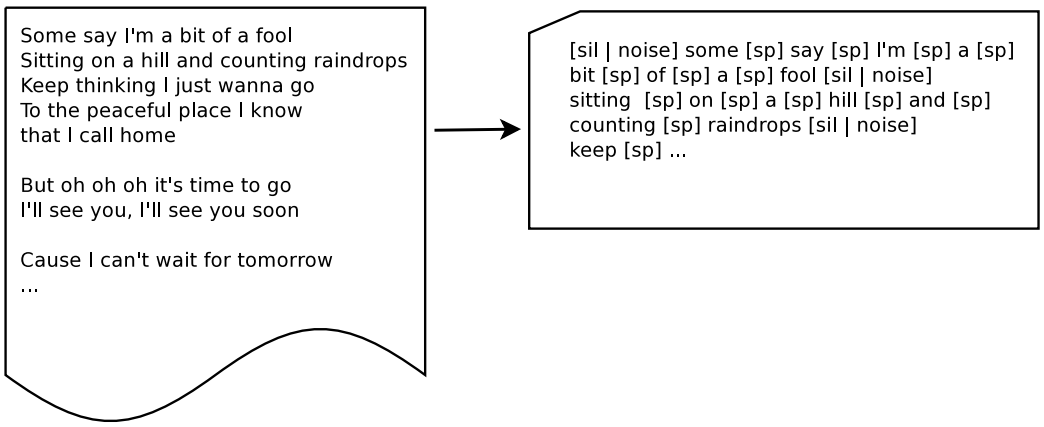

Figure 4.3: Transforming the lyrics text into the grammar for forced alignment: optional short pause is inserted between each words, while at the end of each line a longer pause is represented by optional silence or, alternatively, noise

each two words in the lyrics. At the end of each line an optional *silence or noise* ('sil | noise') event is inserted, to account for the voice rest and possible background accompaniment. The resulting sequence is the *recognition grammar* and the process is presented in Figure 4.3.

Symbol [ ] encloses options and | denotes alternatives. With these alternatives, the alignment algorithm can choose to include pauses and noise where needed while skipping them when they are not necessary. The grammar is formed for the entire audio signal, without any specific vocal detection, assuming that the 'silence | noise' insert will be used by the system to represent the nonvocal regions. The phonetic transcription of the recognition grammar is obtained using the CMU pronouncing dictionary. The features extracted from the separated vocals are aligned with the obtained string of phonemes, using the Viterbi forced alignment.

It is difficult to evaluate the performance of the alignment for each word, as annotating such details is very time consuming. Using line level annotation, with a line being defined as a section delimited by a line feed in lyrics text, the errors at the start and end of each line can be calculated. The system performance was evaluated as having an average absolute error of 1.4 seconds on a test database consisting of fragments of popular music, totaling over 1000 lines of text. The performance is improved to less than one second average absolute er-

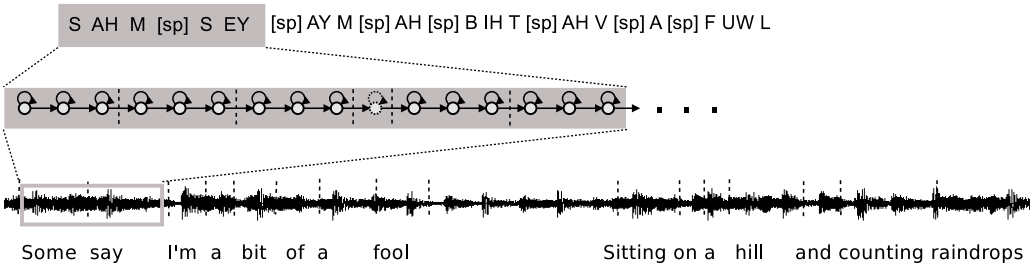

Figure 4.4: The lyrics alignment process: Phonetically transcribed lyrics are represented using the phoneme HMMs, and the models are aligned with the features extracted from the audio.

ror in the followup [P3] by using NMF to separate the vocal line for alignment.

It is important to note that measuring the average absolute error in seconds is probably not the best way of expressing the result, as the perception of the same duration error can be different depending on the tempo of the song. However, the performance obtained in [P2] and [P3] is comparable to the results presented in [42], where the performance is evaluated in musical bars. In [42], the reported average error is less than one bar, which amounts to at least 1.3 seconds, considering the restricted structure and meter of the songs used as test data.

4.4 Recognition of Phonemes and Words in Singing

The work about phoneme and word recognition in singing is fragmented, with the studies belonging to one of two approaches: basic sound classification approach, or speech recognition approach. The classification approach uses isolated phonemes and solves the problem in the classical way, by using a small number of sound classes and classifying them based on certain features. One such system was presented for phoneme recognition in popular music [34]. The preprocessing of the audio signal consists of pitch detection, harmonics extraction and resynthesis of the vocal line to eliminate the instrumental accompaniment. The training and testing material consisted in short segments of isolated vowels selected from polyphonic music, and it was classified into one of 15 classes of voiced phonemes with a 57% performance.

An alternative framework for recognizing phonemes of singing voice in polyphonic music uses multiple vocal templates for each vowel [29] for recognizing a number of five (Japanese) vowels: a, i, u, e, o. This approach performs recognition directly from the polyphonic mixture, without using a voice separation step. The models are constructed using probabilistic spectral templates and modeling the generation of the spectrogram of a singing voice plus accompaniment. The training material for generating the vocal templates consists of monophonic recordings with manually annotated pitch. The system was developed further to use polyphonic music for training the models, and to concurrently estimate the pitch from the mixture [24].

The speech recognition approach to recognizing lyrics was simplified in some cases by using a restricted grammar or a small vocabulary. One such example is the method for lyrics recognition based on a finite state automaton, used for music information retrieval in [38] and [86]. The recognition grammar was constructed as a finite state automaton with only two transitions permitted from each word: either to the next word in the lyrics or to the end symbol. The sung query was also restricted to exactly five words, and a language model was trained from the lyrics in the database.

Another example of restricting the decoding using a small vocabulary is presented in [82], by preparing a language model and a dictionary for each song. The method was used for testing recognition on a number of 12 songs, but the focus was on on signal parameter estimation, therefore the speech recognition aspect was not explained in detail.

Publications [P5] and [P6] take a large vocabulary continuous speech recognition (LVCSR) approach to lyrics transcription. A set of phoneme models is trained using speech data and adapted to singing voice using a small database of monophonic singing. A language model specific to lyrics text was built using as training data the lyrics text of over 4400 songs. By selecting the most common words, a vocabulary of just over 5000 words was included in the language model. For phoneme recognition, language models were built from general phonetically balanced text.

The most common words encountered in the lyrics database used for creating the language model can be seen in Figure 4.4. Excluding pronouns, functional words such as articles (*a, the*), adpositions (*on, in, by, for, with*, etc), and other grammatical particles (*so, but*, etc), the most common words in the used lyrics database are *don't, like, know* and *love*. From the pronouns, most common is *you*, followed by *I* and

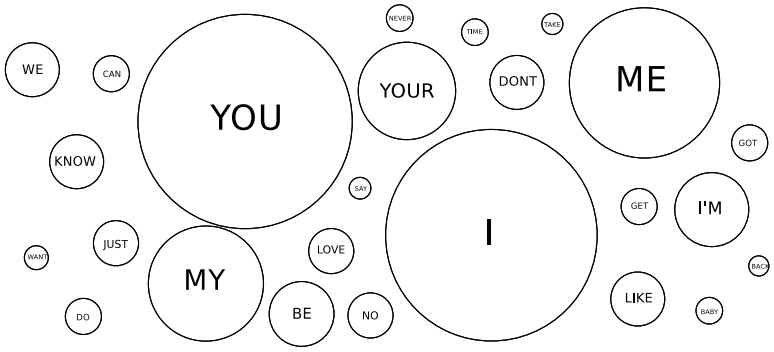

Figure 4.5: Most common words in the lyrics language model. Size of the representation is proportional to the word frequency.

me. This obviously reflects in some way the type of music, as it is hard to believe these would be the most common words in black metal or hip hop music.

The recognition results for monophonic singing are at about 24% correct word recognition rate, which is quite low. For comparison, a small vocabulary language model was constructed including all the words of the test database, similar to the approach in [82], resulting in a vocabulary of 185 words. By allowing free decoding (any number of words), the results were 55% correct word recognition, significantly higher than when using the large vocabulary language model.

Summary

As a conclusion, we can say that the transcription of lyrics from singing voice remains a difficult problem. The topic has attracted interest in the research community, and has prompted the development of various approaches. The methods rooted in speech recognition are elegant but do not offer a satisfactory performance, due to multiple factors. As most of the time the singing voice is encountered in a complex polyphonic mixture containing many musical instruments, signal separation methods are one way of obtaining the singing voice signal from the mixture. Alternatives that do not do the separation have also been proposed, but studied at small scale for now.

Chapter 5

Applications Based on Lyrics Alignment and Transcription

The methods presented so far for lyrics transcription and lyrics alignment to audio facilitate applications in the music service area. This chapter briefly introduces three applications. Two of them are based on the singing voice alignment to lyrics text [P2]: automatic annotation for karaoke and song browsing. The third application is based on lyrics transcription from monophonic singing [P5] for query-by-singing, using the recognized words from the query for the database search.

5.1 Automatic Karaoke Annotation

Karaoke is a form of interactive entertainment in which amateur singers sing along with a music video using a microphone. The karaoke performer has the lyrics of the song displayed on a screen, and the color of the lyrics changes (or there is some symbol, like a ball bouncing on top of the displayed words) for guiding the singer for timing. The music used consists usually of popular songs without the lead vocal, and the karaoke performer completes the song with his vocal interpretation. Karaoke is a popular form of entertainment in some cultures, often performed in a social gathering among friends. There are both public and home use systems, and the included songs are either individually purchased or in collection. The systems are commercially available, with costs due to copyright issues regarding commercial use and lyrics synchronization work which is done manually.

The method for alignment of the singing voice with the lyrics text [P2] provides the complete synchronization information for each word

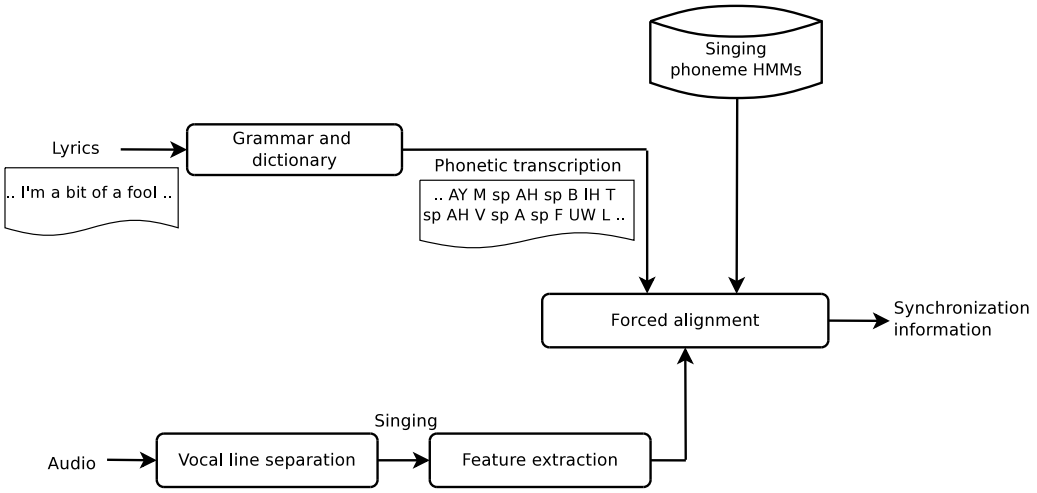

Figure 5.1: Steps in creating the sync information for a karaoke system, staring from the lyrics text and audio.

in the lyrics to corresponding timestamp in the audio. The system can therefore provide a home-use karaoke system to create annotations on-the-fly to any song, as long as the audio and the corresponding lyrics are available.

There are a few commercial software applications available for creating karaoke type of annotations, such as the Creative Karaoke Player[1], the Karaoke CD+G Creator from Power Karaoke[2], or the freeware Del mp3 Karaoke[3]. They rely on the user to create the annotation. The usual action required from the user is to tap at each word or syllable start, to export the timestamps to a special file format, and then to edit this file and adjust the timestamps where needed. After a few iterations, it is ready for singing along.

The method presented in [P2] performs the alignment automatically, and would need the intervention of the user only for correcting the timestamps, if the user is not satisfied with the synchronization. Figure 5.1 presents the block diagram for obtaining the alignment. The word level alignment might not be very exact, but typically people will want to sing karaoke on songs that they are familiar with, therefore scrolling the lyrics lines with good sync might be already acceptable quality of alignment for home use entertainment. One method for fin-

[1]Creative Karaoke Player
http://www.creative.com/soundblaster/products/software/subsoft.asp?id=3

[2]Karaoke CD+G Creator http://www.powerkaraoke.com/src/prod_karaokecdgcreator.php

[3]Del mp3 Karaoke http://www.delmp3karaoke.com/index_en.shtml

ishing the karaoke application is for the user to create a video with the audio file as a soundtrack, for which the text file with synchronization information can be used as subtitles.

5.2 Song Browsing

When looking for new music, people are interested in hearing a few samples before making a purchase. Usually they want to skip the intro, looking for a more representative fragment, most of the times the chorus. They do the search by pressing the fast forward or randomly skipping forward until they reach some interesting spot in the song. In order to provide some automatic tool for this situation, various methods were developed for automatic detection of chorus [32], audio thumbnailing [2, 4], and music summarization [14, 70].

Lyrics are often included when publishing music, therefore it is easy to take advantage of the additional information and use automatic alignment for locating the chorus. In fact, the synchronization information output by the alignment system in [P2] can be used to locate any word, within the audio.

Figure 5.2 shows a possibility of using this information for navigating within a song. The lyrics are displayed, and each word is linked to its corresponding time in the video. The video can be streamed from online services like for example YouTube. By reading the lyrics, the user can select a point in the song where he wants to hear the music, and by clicking on the point of interest in the lyrics, the video playback will jump to the corresponding point. While the song is playing, the words displayed are changing color, with the current word being highlighted. This allows the user to know exactly at each time where the playback is.

The structure of the song can also be displayed for guidance, so that the chorus is highlighted. The structure represented in this figure by different colors was obtained through automatic segmentation of the lyrics text, using the longest common sub-sequence algorithm and comparing the beginning of each paragraph.

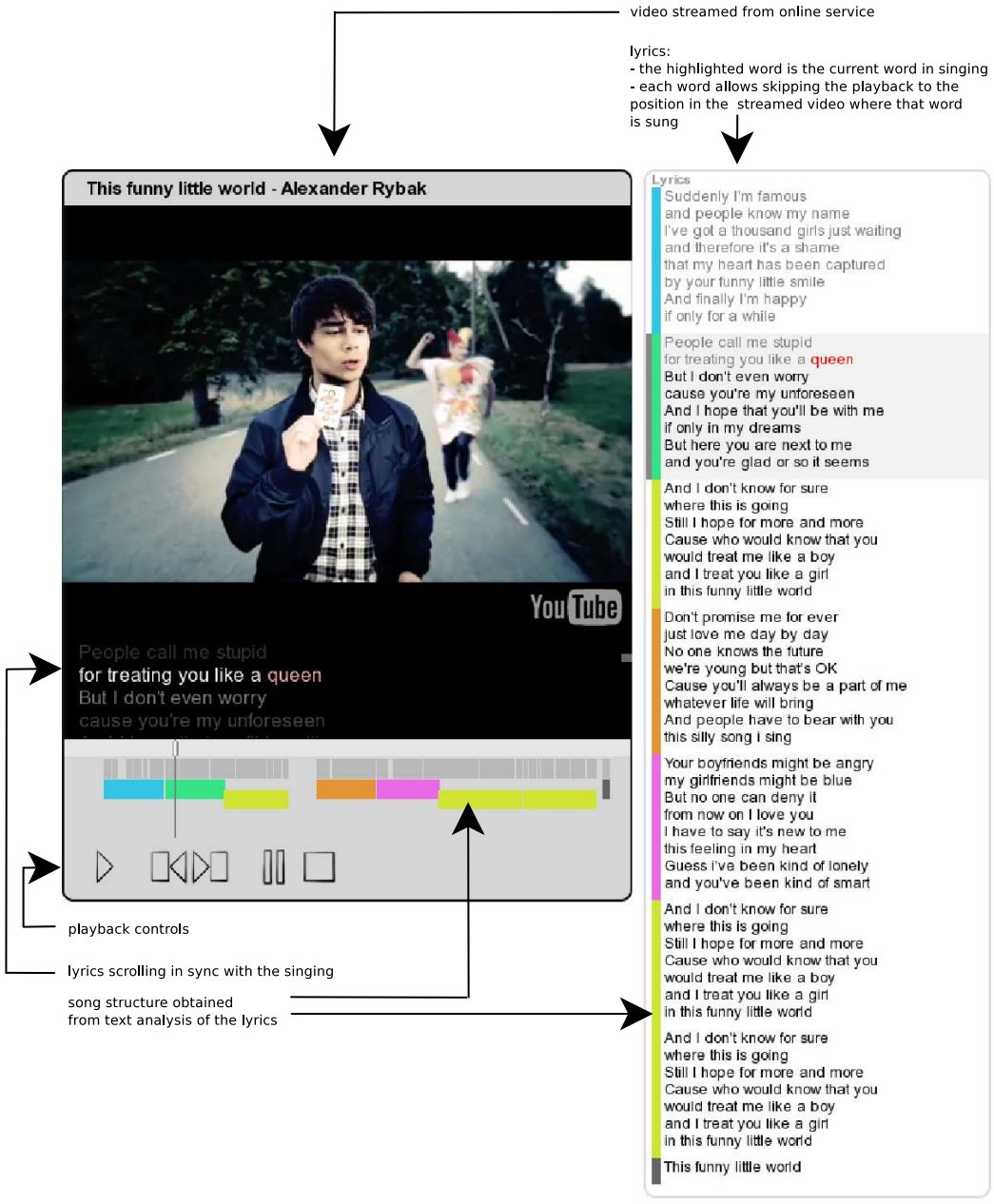

Figure 5.2: Song browsing: The alignment information between lyrics and audio is used to skip to points of interest in the video playback. Song structure and individual words in the lyrics are all linked to corresponding positions in the video

5.3 Query-by-Singing

Query by singing or humming (QBH) is a music information retrieval type of application in which the aim is to find a song in a database based on a query sung by the user. QBH is a research topic that has been active for decades, and has generated lots of development especially in the area of melody transcription and symbolic representation [94]. The query by singing/humming systems typically use only the melody information, while the use of lyrics for finding songs is limited to performing text search in lyrics databases. In the QBH system, the melody sung by the user is transcribed and a match to this melody is searched for in the database. For large databases, the search time can be significantly long.

If the lyrics of the query are available, they can be used for searching a match in lyrics text files. Using a text search as a first step can offer the advantage of speed in narrowing down the melody search space to a few songs that best match the lyrics. The lyrics transcription from the query can be thus used to exploit the additional information brought by the recognized words, even if not all of them are in fact correct. The melody-based search also faces the difficulty of matching melodies in case of less skilled singers, who cannot follow the original melody or tempo. In such a situation, the lyrics could be more reliable in offering retrieval hypotheses. One example of system is presented in [38] for retrieval from lyrics and melody, with a query restricted to exactly five words and a recognition system using a closed vocabulary .

The method presented in [P5] is based on free decoding, therefore offering the possibility of singing any number of words, or even combining humming and singing where the words are not known. A few correct words can give important clues for identifying the song. Another method of exploiting the phonetic content of the query was presented in [98]: after a free decoding of the query, the number of phonemes in the transcription was counted to decide if the query was singing or just humming. When the query was singing, the retrieval used a similarity measure to compare the query to the first 30 syllables of each song in the database, assuming that the user sings the beginning of the song.

The prototype query by singing system [P5] is based on text search from a transcribed query. It uses a small database of lyrics text files and a number of 49 queries representing 12 different songs. The recognition performance of the lyrics transcription system on the 49 queries is approximately 24% correct word rate. The output of the recognition system offers sometimes words that are acoustically very similar with

Correct transcription	Recognized
yesterday	yes today
seemed so far away	seem to find away
my my	mama
finding the answer	fighting the answer
the distance in your eyes	from this is in your eyes
all the way	all away
cause it's a bittersweet symphony this life trying to make ends meet you're a slave to the money then you die	cause I said bittersweet symphony this our life trying to maintain sweetest ain't gettin' money then you down
I heard you crying loud all the way across town you've been searching for that someone and it's me out on the prowl as you sit around feeling sorry for yourself	I heard you crying alone all away across the sign you been searching for someone I miss me I don't apologize you see the rhyme feelin' so free yourself

Table 5.1: Examples of errors in recognition.

the correct ones, sometimes cases with different spelling but same phonetic transcription. Some examples representing the transcriber "misheard" lyrics can be found in Table 5.1. For recognition performance evaluation, all that is incorrect counts as error, but for music information retrieval purpose, the words which are correct are the valuable information, especially when obvious key words of the song were correctly recognized (e.g. "bittersweet symphony").

A basic bag-of-words approach was used for searching the text: each word in the transcription of the query was searched for in each of the lyrics files, and the songs were ranked by the number of matched words. When the queried fragment appears among the first N ranked lyrics, the song is considered to be correctly identified. In our small scale experiment, the retrieval results are promising, with the first ranked song being the correct retrieval result in 57% of the cases.

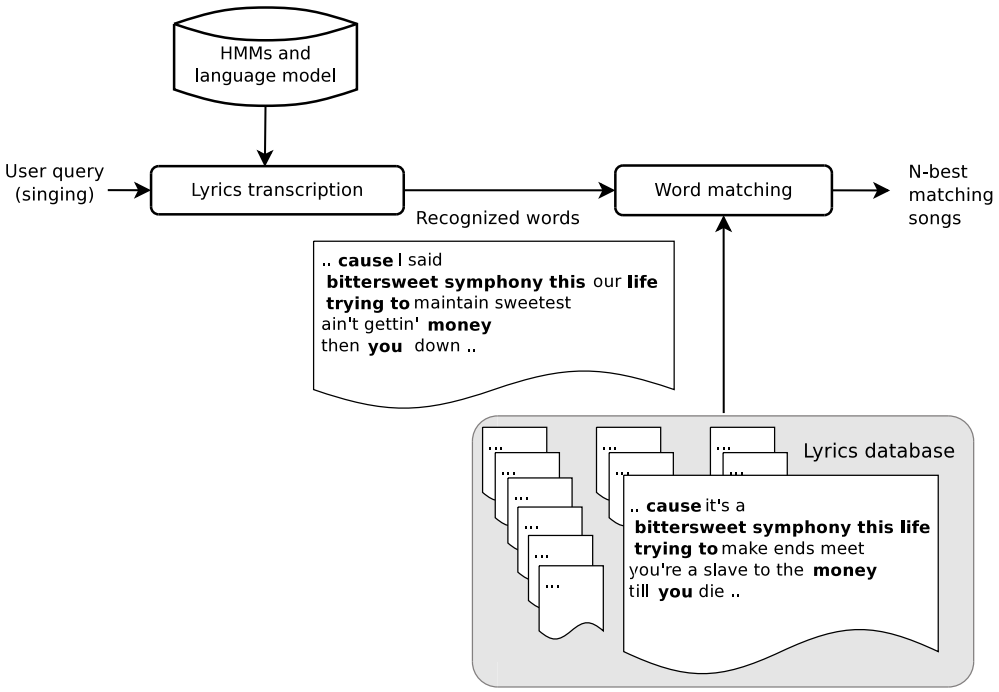

Figure 5.3: Query-by-singing: The words from the output of the recognition system are searched one by one in the database containing lyrics text and the results are ranked by the number of words that are found in each file.

Summary

The applications presented in this chapter prove that despite the transcription results being far from high performance, there is potential for using them in particular tasks. The automatic alignment of lyrics and singing can be satisfactory for browsing music and for line level synchronization, while the query-by-singing indicates that it is possible to recognize correctly a song even by using one quarter of the words from the query.

Chapter 6

Conclusions and Future Work

This thesis has presented methods for analysis and recognition of the singing voice signals, aimed at extracting information about the singer identity and the lyrics. The presented approaches encourage development of automatic music retrieval applications targeting general public. In this sense, the developed methods are shifting interest from studying monophonic music towards the polyphonic music.

The singing voice is a signal with specific characteristics, some related to speech, some related to musical signals. The singing, just like speech, carries the individual characteristics of the person, which allows people to recognize who is singing. At the same time, singing carries a semantic message, just like speech, which allows people to understand its meaning. Yet another aspect of singing is related to the musical sounds, in carrying melodic information in the form of musical notes and artistic expression in frequency modulations.

Related to singer identification, the thesis described methods for classification of singing voice in monophonic and polyphonic music. Publication [P1] studies the degradation in performance of various classifiers when decreasing the signal to noise ratio, which practically represents the relative level of the singing voice in the polyphonic mixture compared to the accompaniment. Different classification setups show similar degradation in performance from loudly mixed voice down to a level where accompaniment is louder than voice, when classifying the polyphonic mixture. A signal separation approach used for preprocessing the audio in order to extract the singing voice from the mixture is introduced. The results suggest that separating the singing from the mixture is a very good approach to classifying the singers, as it significantly improves performance in the low signal to ratio mixtures.

Recognition of lyrics from singing is practically speech recognition from a singing input, and the presented solution for this is tuning speech phoneme models to the singing voice characteristics. In the recognition process, Viterbi algorithm is used for decoding a word or phoneme sequence that best represents the observation sequence. In a simplified setup, the same decoding process can be used for alignment of a given phoneme sequence with the observation sequence. Publication [P2] presents a system for alignment between known lyrics and singing voice extracted from the polyphonic mixture. The audio is preprocessed to extract the vocal line using melody transcription and sinusoidal modeling and the lyrics text is preprocessed to obtain a phoneme sequence. The performance of the alignment is evaluated by the average absolute alignment error at the beginning and end of each line. In publication [P3], the same system is evaluated using vocal line separation based on NMF. The results encourage development of applications based on alignment, as the performance is already good enough for selected applications. Chapter 5 presented a home-use karaoke production system that uses automatic annotation of the lyrics based on the automatic alignment between singing and lyrics, and a song browsing application that highlights the song structure and allows skipping the playback to any specific word from the lyrics.

For lyrics transcription approached as a LVCSR, the thesis presented methods for adaptation of speech phoneme models to singing and creation of a singing specific language model. Publications [P4] and [P6] present few different setups for model adaptation, from global to gender specific and voice specific. The adaptation step seems especially beneficial to the female singing voice, and this is somehow easy to understand when considering the variability of singing voices and male/female voice differences. Publications [P5] and [P6] present also word recognition evaluation of a complete speech recognition system complete with language models and singing adapted models, for evaluating lyrics transcription from monophonic singing and from singing separated from polyphonic music. Chapter 5 presented an application developed based on the lyrics transcription system: query-by-singing. The singing query input is transcribed by the system, and all the recognized words are matched against lyrics text from a database. The songs are ranked based on the number of words from that were found in the lyrics, and the top N are provided as a retrieval result. Despite quite low word recognition performance of the lyrics transcription system, the query-by-singing system shows good performance in retrieving the correct song.

The methods presented for singing voice classification, lyrics transcription and alignment combine methods from pattern recognition, speech recognition and music signal analysis. There are however many things left unsolved and plenty of space for improvement. Further research in singer identification may be oriented towards recognition of singers in audio containing multiple voices and instrumental accompaniment, as most of the research until now was directed towards single voice material.

The semantic aspect of singing is quite complex, with large variability in articulation for each phoneme. If anyone would start a task of building phonetic models from singing material, this would require a considerably large database, covering different pitches and phonation types for producing singing. The task should probably be approached by building separate models for male and female voices, and possibly different models for the different phonation types. It is hard to say if the cases for lyrics transcription from monophonic singing and polyphonic music should be treated separately, as the possibility of collecting databases is much better if considering polyphonic music. Voice separation methods are already a successful tool in obtaining the vocal line from a song, and this can be useful in creating a database of separated singing, with possibility of automatic annotation by forced alignment such as the method presented in this thesis.

Bibliography

[1] N. Adams, M. Bartsch, and G. Wakefield. Coding of sung queries for music information retrieval. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 139–142, oct. 2003.

[2] J.-J. Aucouturier and M. Sandler. Finding repeating patterns in acoustic musical signals : Applications for audio thumbnailing. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, 6 2002.

[3] J. J. Barnes, P. Davis, J. Oates, and J. Chapman. The relationship between professional operatic soprano voice and high range spectral energy. *Journal of Acoustical Society of America*, 116:530–538, 2004.

[4] M. Bartsch and G. Wakefield. To catch a chorus: using chroma-based representations for audio thumbnailing. In *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 15–18, 2001.

[5] M. A. Bartsch. *Automatic singer identification in polyphonic music*. PhD thesis, University of Michigan, 2004.

[6] A. Berenzweig and D. Ellis. Locating singing voice segments within music signals. In *2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 119–122, 2001.

[7] A. Berenzweig, D. P. W. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. In *AES-22 International Conference on on Virtual, Synthetic and Entertainment Audio*, 2002.

[8] A. S. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, MA, 1990.

[9] J. J. Burred and T. Sikora. Monaural source separation from musical mixtures based on time-frequency timbre models. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, pages 149–152, 2007.

[10] P. Cabanas-Molero, D. M. Munoz, M. Cobos, and J. J. Lopez. Singing voice separation from stereo recordings using spatial clues and robust f0 estimation. In *Audio Engineering Society Conference: 42nd International Conference: Semantic Audio*, 7 2011.

[11] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01).*, volume 2, pages 865–868, 2001.

[12] L. P. Clarisse, J. P. Martens, M. Lesaffre, B. D. Baets, H. D. Meyer, H. Demeyer, and M. Leman. An auditory model based transcriber of singing sequences. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 116–123, 2002.

[13] M. Cobos and J. J. Lopez. Singing voice separation combining panning information and pitch tracking. In *Audio Engineering Society Convention 124*, 5 2008.

[14] M. Cooper and J. Foote. Summarizing popular music via structural similarity analysis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 127–130, oct. 2003.

[15] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, aug 1980.

[16] D. Deutsch. Speaking in tones. *Scientific American Mind*, (July/August):36–43, 2010.

[17] J. Devaney, M. Mandel, and D. Ellis. Improving midi-audio alignment with acoustic features. In *IEEE Workshop on Applications*

of Signal Processing to Audio and Acoustics, WASPAA '09., pages 45–48, oct. 2009.

[18] J. S. Downie. The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, 29(4):247–255, 2008.

[19] R. Duda, P. Hart, and D. Stork. *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. Wiley, 2001.

[20] L. Feng, A. B. Nielsen, and L. K. Hansen. Vocal segment classification in popular music. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, 2008.

[21] G. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268 – 278, march 1973.

[22] H. Fujihara and M. Goto. A music information retrieval system based on singing voice timbre. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[23] H. Fujihara and M. Goto. Three techniques for improving automatic synchronization between music and lyrics: Fricative detection, filler model, and novel feature vectors for vocal activity detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2008.*, pages 69–72, april 2008.

[24] H. Fujihara and M. Goto. Concurrent estimation of singing voice F0 and phonemes by using spectral envelopes estimated from polyphonic music. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2011*, pages 365–368, may 2011.

[25] H. Fujihara, M. Goto, T. Kitahara, and H. Okuno. A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):638–648, march 2010.

[26] H. Fujihara, M. Goto, and J. Ogata. Hyperlinking lyrics: A method for creating hyperlinks between phrases in song lyrics.

In *Proceedings of the 9th International Conference on Music Information Retrieval ISMIR*, 2008.

[27] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H. G. Okuno. Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. In *ISM '06: Proceedings of the 8th IEEE International Symposium on Multimedia*, 2006.

[28] H. Fujihara, M. Goto, J. Ogata, and H. Okuno. Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics. *IEEE Journal of Selected Topics in Signal Processing*, PP(99):1, 2011.

[29] H. Fujihara, M. Goto, and H. Okuno. A novel framework for recognizing phonemes of singing voice in polyphonic music. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA '09.*, pages 17–20, oct. 2009.

[30] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 329–336, 2005.

[31] M. J. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12, 1998.

[32] M. Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1783–1794, sept. 2006.

[33] M. Goto, T. Saitou, T. Nakano, and H. Fujihara. Singing information processing based on singing voice modeling. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5506–5509, march 2010.

[34] M. Gruhne, K. Schmidt, and C. Dittmar. Phoneme recognition in popular music. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[35] A. Härmä and M. Park. Extraction of voice from the center of the stereo image. In *Audio Engineering Society Convention 130*, 5 2011.

[36] T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR 2009)*, 2009.

[37] N. Henrich, J. Smith, and J. Wolfe. Vocal tract resonances in singing: Strategies used by sopranos, altos, tenors, and baritones. *Journal of Acoustical Society of America*, 129:1024–1035, 2011.

[38] T. Hosoya, M. Suzuki, A. Ito, S. Makino, L. A. Smith, D. Bainbridge, and I. H. Witten. Lyrics recognition from a singing voice based on finite state automaton for music information retrieval. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 532–535, 2005.

[39] C.-L. Hsu and J.-S. Jang. On the improvement of singing voice separation for monaural recordings using the mir-1k dataset. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):310–319, feb. 2010.

[40] E. Joliveau, J. Smith, and J. Wolfe. Vocal tract resonances in singing: The soprano voice. *Journal of Acoustical Society of America*, 116:2434–2439, 2004.

[41] D. Jurafsky and J. H. Martin. *Speech and language processing*. Prentice Hall, Upper Saddle River, New Jersey, 2000.

[42] M.-Y. Kan, Y. Wang, D. Iskandar, T. L. New, and A. Shenoy. Lyrically: Automatic synchronization of textual lyrics to acoustic music signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):338–349, feb. 2008.

[43] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano. Recent progress of open-source lvcsr engine julius and japanese model repository. In *Proceedings of ICSLP*, pages 3069–3072, 2004.

[44] S. Khine, T. L. Nwe, and H. Li. Singing voice detection in pop songs using co-training algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2008.*, pages 1629–1632, 31 2008-april 4 2008.

[45] J. Kilian and H. H. Hoos. Voice separation - a local optimization approach. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, 2002.

[46] Y. E. Kim. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 164–169, 2002.

[47] Y. E. Kim, D. S. Williamson, and S. Pilli. Towards quantifying the album-effect in artist classification. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, 2006.

[48] A. Klapuri and M. Davy, editors. *Signal Processing Methods for Music Transcription*. Springer, New York, 2006.

[49] B. Kostek and P. Zwan. Automatic classification of singing voice quality. In *5th International Conference on Intelligent Systems Design and Applications, ISDA '05.*, pages 444–449, sept. 2005.

[50] K. Lee and M. Cremer. Segmentation-based lyrics-audio alignment using dynamic programming. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, 2008.

[51] C. Leggetter and P. Woodland. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. ARPA Spoken Language Technology Workshop*, 1995.

[52] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9, 1995.

[53] Y. Li and D. Wang. Singing voice separation from monaural recordings. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, 2006.

[54] Y. Li and D. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1475–1487, may 2007.

[55] W.-N. Lie and C.-K. Su. Content-based retrieval of mp3 songs based on query by singing. In *IEEE International Conference*

on Acoustics, Speech, and Signal Processing (ICASSP '04), volume 5, pages 929–932, may 2004.

[56] C.-Y. Lin and J.-S. Jang. Automatic phonetic segmentation by score predictive model for the corpora of mandarin singing voices. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2151–2159, sept. 2007.

[57] C.-C. Liu and C.-S. Huang. A singer identification technique for content-based classification of mp3 music objects. In *Proceedings of the 11th international conference on Information and knowledge management*, pages 438–445, 2002.

[58] B. Logan. Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*, 2000.

[59] A. Loscos, P. Cano, and J. Bonada. Low-delay singing voice alignment to text. In *Proceedings of the International Computer Music Conference*, 1999.

[60] N. Maddage, K. Wan, C. Xu, and Y. Wang. Singing voice detection using twice-iterated composite fourier transform. In *IEEE International Conference on Multimedia and Expo, ICME '04*, volume 2, pages 1347–1350, june 2004.

[61] N. Maddage, C. Xu, and Y. Wang. Singer identification based on vocal and instrumental models. In *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004.*, volume 2, pages 375–378, aug. 2004.

[62] N. C. Maddage, K. C. Sim, and H. Li. Word level automatic alignment of music and lyrics using vocal synthesis. *ACM Trans. Multimedia Comput. Commun. Appl.*, 6(3):19:1–19:16, Aug. 2010.

[63] M. Mehrabani and J. H. L. Hansen. Language identification for singing. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011*, pages 4408–4411, may 2011.

[64] M. Mellody, F. Herseth, and G. H. Wakefield. Modal distribution analysis, synthesis, and perception of a sopranos sung vowels. *Journal of Voice*, 15(4):469–482, 2001.

[65] M. Müller, F. Kurth, D. Damm, C. Fremerey, and M. Clausen. Lyrics-based audio retrieval and multimodal navigation in music collections. In *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries*, 2007.

[66] T. L. Nwe and H. Li. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2):519–530, feb. 2007.

[67] T. L. Nwe, A. Shenoy, and Y. Wang. Singing voice detection in popular music. In *12th annual ACM International Conference on Multimedia*, pages 324–327, New York, NY, USA, 2004. ACM.

[68] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1564–1578, july 2007.

[69] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot. One microphone singing voice separation using source-adapted models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 90–93, oct. 2005.

[70] G. Peeters, A. L. Burthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *In Proc. International Conference on Music Information Retrieval*, pages 94–100, 2002.

[71] E. Pollastri. A pitch tracking system dedicated to process singing voice for music retrieval. In *IEEE International Conference on Multimedia and Expo, ICME '02*, volume 1, pages 341–344, 2002.

[72] D. Pye and P. Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1997.

[73] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257 –286, feb 1989.

[74] L. Rabiner and B. Juang. *Fundamentals of speech recognition*. Prentice-Hall signal processing series. PTR Prentice Hall, 1993.

[75] V. Rao, S. Ramakrishnan, and P. Rao. Singing voice detection in polyphonic music using predominant pitch. In *INTERSPEECH'09*, pages 1131–1134, 2009.

[76] V. Rao and P. Rao. Vocal melody extraction in the presence of pitched accompaniment in polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2145–2154, nov. 2010.

[77] L. Regnier and G. Peeters. Singing voice detection in music tracks using direct voice vibrato detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009*, pages 1685–1688, april 2009.

[78] D. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, oct 1994.

[79] D. Reynolds and R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, jan 1995.

[80] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(13):19–41, 2000.

[81] M. Ryynänen and A. Klapuri. Transcription of the singing melody in polyphonic music. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, 2006.

[82] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka. An autoregressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '05).*, volume 1, pages 237–240, 18-23, 2005.

[83] J. Schwenninger, R. Brueckner, D. Willett, and M. Hennecke. Language identification in vocal music. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pages 377–379, 2006.

[84] S. Sofianos, A. Ariyaeeinia, and R. Polfreman. Towards effective singing voice extraction from stereophonic recordings. In

IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2010, pages 233–236, march 2010.

[85] J. Sundberg. *The science of the singing voice*. Northern Illinois University Press, 1987.

[86] M. Suzuki, T. Hosoya, A. Ito, and S. Makino. Music information retrieval from a singing voice based on verification of recognized hypotheses. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, 2006.

[87] M. Suzuki, T. Hosoya, A. Ito, and S. Makino. Music information retrieval from a singing voice using lyrics and melody information. *EURASIP Journal on Advances in Signal Processing*, 2007.

[88] F. Thibault and P. Depalle. Adaptive processing of singing voice timbre. In *Canadian Conference on Electrical and Computer Engineering*, volume 2, pages 871–874 Vol.2, may 2004.

[89] W.-H. Tsai, S.-J. Liao, and C. Lai. Automatic identification of simultaneous singers in duet recordings. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, 2007.

[90] W.-H. Tsai and H.-P. Lin. Popular singer identification based on cepstrum transformation. In *IEEE International Conference on Multimedia and Expo (ICME), 2010*, pages 584–589, july 2010.

[91] W.-H. Tsai and H.-M. Wang. Towards automatic identification of singing language in popular music recordings. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, pages 568–576, 2004.

[92] W.-H. Tsai and H.-M. Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):330–341, jan. 2006.

[93] W.-H. Tsai, H.-M. Wang, D. Rodgers, S.-S. Cheng, and H.-M. Yu. Blind clustering of popular music recordings based on singer voice characteristics. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 2003.

[94] R. Typke, F. Wiering, and R. C. Veltkamp. MIREX symbolic melodic similarity and query by singing/humming. In *International Music Information Retrieval Systems Evaluation Laboratory(IMIRSEL)*, URL http://www.music-ir.org/mirex2006/.

[95] G. Tzanetakis. Song-specific bootstrapping of singing voice structure. In *2004 IEEE International Conference on Multimedia and Expo, ICME '04.*, volume 3, pages 2027–2030, june 2004.

[96] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, jul 2002.

[97] S. Vembu and S. Baumann. Separation of vocals from polyphonic audio recordings. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 337–344, 2005.

[98] C.-C. Wang, J.-S. R. Jang, and W. Wang. An improved query by singing/humming system using melody and lyrics information. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, 2010.

[99] C.-K. Wang, R.-Y. Lyu, and Y.-C. Chiang. An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003*. ISCA, 2003.

[100] C. H. Wong, W. M. Szeto, and K. H. Wong. Automatic lyrics alignment for Cantonese popular music. *Multimedia Systems*, 12(4-5), 2007.

[101] P. Woodland, J. Odell, V. Valtchev, and S. Young. Large vocabulary continuous speech recognition using HTK. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 1994.

[102] T. Zhang. Automatic singer identification. In *International Conference on Multimedia and Expo, ICME '03.*, volume 1, pages I–33–6 vol.1, july 2003.

Publication P1

A. Mesaros, T. Virtanen and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods," in *Proceedings of the 8th International Conference on Music Information Retrieval*, (Vienna, Austria),pp. 375–378, 2007.

Publication P2

A. Mesaros and T. Virtanen, "Automatic alignment of music audio and lyrics", in *Proceedings of the 11th International Conference on Digital Audio Effects*, (Espoo, Finland), pp. 321–324, 2008.

Publication P3

T. Virtanen, A. Mesaros and M. Ryynänen, "Combining Pitch-Based Inference and Non-Negative Spectrogram Factorization in Separating Vocals from Polyphonic Music", in *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, (Brisbane, Australia), pp. 17–22, 2008.

Publication P4

A. Mesaros and T. Virtanen, "Adaptation of a speech recognizer for singing voice", in *Proceedings of the 17th European Signal Processing Conference*, (Glasgow, Scotland), pp. 1779–1783, 2009.

Publication P5

A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing", in *Proceedings of the 35th International Conference on Acoustics, Speech, and Signal Processing*, (Dallas, Texas, USA), pp. 2146 – 2149, 2010.

Publication P6

A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing", *EURASIP Journal on Audio, Speech and Music Processing*, Volume 2010, 11 pages, 2010.

Tampereen teknillinen yliopisto
PL 527
33101 Tampere

Tampere University of Technology
P.O.B. 527
FI-33101 Tampere, Finland

ISBN 978-952-15-2893-4
ISSN 1459-2045