UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

Time-frequency reassignment for acoustic signal processing

From speech to singing voice applications

Georgina Tryfou

Advisor

Maurizio Omologo

Fondazione Bruno Kessler (FBK)

April 2017

To all the mighty girls,
and their mothers.

Acknowledgements

I would like to thank my advisor, Maurizio Omologo, for his support and continuous guidance. I really appreciate all the time and resources that he has spent for this work. I also want to thank my colleagues in the SHINE group for offering their help, and scientific or technical expertise whenever it was asked.

In our phd office, I had the chance to meet some very interesting people, and enjoy a friendly and multicultural environment. I want to thank all the students I met, for all the help they offered, in a professional and personal level. Among them, Marco Pellin excelled in both, being one of the most resourceful persons I have met. I also want to thank my friends outside the office, who were either in Trento, or anywhere in the world, for all the time, drinks and discussions shared.

The most special thank you goes to Cristina Guerrero, who has been a valuable colleague and a true friend. It is hard to express my deepest gratitude to her in these few lines, as she taught me so many things, inspired me in new ways and supported me as family does.

I would also like to thank my parents for teaching me to work hard to achieve my goals, and showing me that the sky's the limit. My sister for always being available despite her heavy work schedule, and for visiting, calling and advising me all the time, lifting me up with her can-do attitude and never ending positivity.

Finally, the two most special persons for me, my husband and daughter. Petro, thank you for taking this journey with me, and never losing your faith in me. This, and many more, would not have been possible without your constant love and support. Youlie, thank you for showing me a whole new perspective in life, and smiling at me every morning.

Abstract

The various time-frequency (TF) representations of acoustic signals share the common objective to describe the temporal evolution of the spectral content of the signal *i.e.,* how the energy, or intensity, of the signal is changing in time. Many TF representations have been proposed in the past, and among them the short-time Fourier transform (STFT) is the one most commonly found in the core of acoustic signal processing techniques. However, certain problems that arise from the use of the STFT have been extensively discussed in the literature. These problems concern the unavoidable trade-off between the time and frequency resolution, and the fact that the selected resolution is fixed over the whole spectrum.

In order to improve upon the spectrogram, several variations have been proposed over the time. One of these variations, stems from a promising method called *reassignment.* According to this method, the traditional spectrogram, as obtained from the STFT, is reassigned to a sharper representation called the Reassigned Spectrogram (RS). In this thesis we elaborate on approaches that utilize the RS as the TF representation of acoustic signals, and we exploit this representation in the context of different applications, as for instance speech recognition and melody extraction.

The first contribution of this work is a method for speech parametrization, which results in a set of acoustic features called time-frequency reassigned cepstral coefficients (TFRCC). Experimental results show the ability of TFRCC features to present higher level characteristics of speech, a fact that leads to advantages in phone-level speech segmentation and speech recognition. The second contribution is the use of the RS as the basis to extract objective quality measures, and in particular the reassigned cepstral distance and the reassigned point-wise distance. Both measures are used for channel selection (CS), following our proposal to perform objective quality measure based CS for improving the accuracy of speech recognition in a multi-microphone reverberant environment. The final contribution of this work, is a method to detect harmonic pitch contours from singing voice signals, using a dominance weighting of the RS. This method has been exploited in the context of melody extraction from polyphonic music signals.

Keywords

Time-frequency reassignment, reassigned spectrogram, speech recognition, objective quality measures, channel selection, melody extraction

Contents

List of Tables

List of Figures

Abbreviations

T_{60} reverberation time.

f_0 fundamental frequency.

ASR automatic speech recognition.

CD cepstral distance.

CMN cepstral mean normalization.

CS Channel selection.

DFT discrete Fourier transform.

DNN deep neural network.

DRR direct to reverberant ratio.

DRS dominance reassigned spectrogram.

DSR distant speech recognition.

DTW dynamic time warping.

EV envelope variance.

GD group delay.

GMM Gaussian mixture model.

HMM hidden Markov model.

IDCT inverse discrete cosine transform.

IF instantaneous frequency.

IR impulse response.

LLR log likelihood ratio.

LPC linear prediction coefficients.

MFCC Mel frequency cepstral coefficients.

MIREX Music Information Retrieval Evaluation eXchange.

MMWM modified moving window method.

MPC melodic pitch contour.

MPD mixed partial derivatives.

PDA pitch detection algorithm.

PER phone error rate.

PLP perceptual linear prediction.

RCD reassigned CD.

RMS root mean square.

RPWD Reassigned point wise distance.

RS reassigned spectrogram.

SDM single distant microphone.

SNR signal to noise ratio.

SNRseg segmental SNR.

STFT short-time Fourier transform.

TFD time-frequency distribution.

TFR time-frequency reassigned.

TFRCC time-frequency reassigned cepstral coefficients.

WER word error rate.

WVD Wigner-Ville distribution.

Chapter 1

Introduction

Sound is an important source of information for humans, and one of the most intuitive forms of communication among them. In an era characterized by a complete automation and a wide range of affordable personal electronic devices, which provide constant access to services such as the internet of things and cloud computing, users demand a satisfying machine understanding of sounds. In an example scenario, a user would be able to use voice commands in order to access and organize a huge, untagged collection of music according to a certain criterion, such as the artist or the genre. A viable solution for this scenario comprises, among other parts, a robust speech recognizer and a content-based music information retrieval module. Over the last years, several steps towards these directions have been made, fuelled by extensive scientific activities in the area of *acoustic signal processing*. However, there are still many open challenges and limitations in addressing the user expectations for quality, mobility and functionality in a wide range of audio, or voiced based, applications.

This thesis is concerned with signal processing topics that aim at extending the current state-of-the-art in audio based solutions, and in particular speech and singing voice applications.

1.1 Time-frequency representations

One of the most important tools in analysing and understanding acoustic signals in an automatic way is the *Fourier* analysis, which allows the decomposition of acoustic signals into the individual frequencies they are composed of, and establishes the relative intensities of these frequency components. Although Fourier analysis is undeniably a powerful tool for audio signal processing, it has an important limitation, since it does not provide meaningful temporal information on the occurrence of the various frequency components. Most real life acoustic signals are non-stationary, which means that their frequency components are changing over time. For example, if we Fourier-analyse a 2-minute song, the obtained energy density spectrum will not give us any information of the various distinct notes that compose the piece, nor of their temporal arrangement.

In the above example, if we split the song into many consecutive short slices, each comprising a single note, the Fourier analysis of each slice and the concatenation of the results would enable an easy identification of each note. This is the main idea behind the short-time Fourier transform (STFT) and its graphic representation called spectrogram [Koenig et al., 1946, Potter et al., 1947]. Since its invention, and with the various subsequent developments on it [Allen and Rabiner, 1977, Dziewonski et al., 1969], the spectrogram lies in the core of the vast majority of acoustic signal processing techniques, being the standard analysis tool for non-stationary signals. It provides a description of the temporal evolution of the frequency components of the signal, *i.e.,* it is a *time-frequency representation* of the input acoustic signal [Cohen, 1989, 1995, Flandrin et al., 2013]. Although it is the most well-known time-frequency representation, the spectrogram is not the only valid approach. Alternative representations derive from other methods to map the energy, or the intensity, of an acoustic signal in the time and frequency domains simultaneously, thus describing the evolution of the spectral content of the signal.

In this thesis we study one of these time-frequency representations called the RS, which is obtained by the method of time-frequency reassignment of the traditional spectrogram. Although this method is relatively old and can be traced back to [Kodera et al., 1976], it has not been extensively studied in the literature. The particularly *sharp* description of the various spectral components and their evolution, as offered by the RS, can lead to improvements in systems designed to understand, and further process the spectral components of acoustic signals.

1.2 Speech and singing voice processing

Acoustic signal processing has numerous, and increasingly important applications, including speech recognition, music signal analysis, seismic data analysis, fetal imaging through sonograms, and radar based tracking. In this work we focus on applications that are concerned with the analysis and understanding of *human voice*, and in particular in two of the most common modes of it, *i.e.,* speech and singing. The acoustic signals created by the human voice production system are very complex, and highly informative signals, and human communication is principally accomplished through them. Speakers and singers send messages, which encode specific ideas, are transmitted through a certain communication channel, are often mixed, on purpose or not, with secondary acoustic signals, and are finally received and decoded by a listener. Each of these stages is critical to the quality and effectiveness of the communication process, as well as the successful conveyance of the encoded message. First, the physical characteristics of the speaker (or signer), the mood they are in and even their cultural background influence the acoustic signal at its source. During the transmission, factors such as other acoustic signal sources (noise), reflections on surfaces (reverberation) and even the medium (air, phone) shape and, most often, degrade the quality of the speech signal. Particularly in singing, the acoustic signal is consistently mixed with additional signals, *i.e.,* background music generated by various musical instruments. Finally, at the side of the listener the perception and decoding are further affected by physiological and psychological responses of humans to sound, often called *psychoacoustics*.

In addition to this complex communication process, the nature of the signal itself introduces further challenges to the automatic machine understanding of human voice. Human voice has a very rich structure, which is determined by the voice organ and, more specifically the collaboration of three main parts: the lungs, the vocal folds (or chords), and the articulators. When the lungs produce an adequate airflow the vocal folds vibrate creating an audible sound source, which can be fine tuned to a certain frequency from the muscles of the larynx. Alternatively, during unvoiced sounds the vocal folds do not vibrate and a noise like audible source signal is produced. In any case, the various articulators filter the source, and in these two steps the voice production system is capable of producing a highly sophisticated array of sounds, entirely unique for each individual.

When observed simultaneously in the time-frequency domain, speech signals can be viewed as a combination of melody, harmony and rhythm [Fulop, 2011]. The goal of the majority of systems that perform speech signal analysis is to make sense of one or more of these three attributes, while simultaneously decoding the message carried by voice signals, and minimizing variabilities attributed to the encoding and transmission processes. Time-frequency analysis, *i.e.,* the study of various time-frequency representations of the speech and singing voice signals, is essential for achieving these goals, as it facilitates the detection and full description of melody, harmony, and rhythm.

1.3 Motivation

Over the last years we observed a rapid development in high tech industry, which keeps producing affordable electronic devices, with ever increasing computing power and storage capabilities. These tremendous technological advances resulted in a huge demand for intuitive, voice-based interaction, whether it is for hands-free control of devices, for enhanced quality of living, or for entertainment.

One of the application groups, which pre-existed the technological growth of the last years but gained further industrial and scientific interest because of it, is related to speech analysis and understanding. Systems such as automatic speech recognition (ASR), speech synthesis, speaker recognition, speaker diarisation fall in this category. The relevance and scientific interest of these directions is further supported by extensive evaluation campaigns and industrial solutions. Directly related to ASR, which is perhaps the most popular of the mentioned systems, some examples of such evaluation campaigns are the CHiME-3 [Barker et al., 2015], REVERB [Kinoshita et al., 2013], ASpIRE [Harper, 2015] and ACE [Eaton et al., 2016] challenges. Concerning industrial solutions we can mention the Google API[1], the Alexa Voice Service from Amazon[2], the speech recognition API offered by Apple[3] and the Custom Speech Services from Microsoft[4].

On the other hand, numerous systems are concerned with the automatic analysis, synthesis and understanding of singing voice. Popular examples here are the various query-by-singing/humming systems, that perform content-based search in huge music collections, as for instance the Shazam[5] and SoundHound[6] services. Singing voice synthesis is also gaining popularity as indicated by the various voice synthesizers, such as the VOCALOID[7] and the Cantor[8]. Regarding the research activities in this area, the increased interest in singing voice is also indicated by the increasing amount of related tasks within the Music Information Retrieval Evaluation eXchange (MIREX)[9] evaluation campaign, which is a community based framework setting the research directions in the area of music analysis.

Given this great demand for methods that improve voice based applications and deliver excellent results in applications as the ones mentioned above, we contribute to this direction by exploring how the RS can improve the representation and automatic analysis of human voice signals, in particular speech and singing voice.

1.4 Scope of the thesis

The scope of this thesis is to explore the RS, and to exploit this representation in the context of acoustic signal processing applications, that are concerned with the analysis and understanding of speech and singing voice. Various aspects complicate this study, as for instance the noisy nature of the RS, the particularly rich structure of human voice signals, and the complex communication process that delivers the signal to a listener. To

[1] http://cloud.google.com/speech/
[2] http://developer.amazon.com/alexa-voice-service/
[3] http://developer.apple.com/reference/speech/
[4] https://cris.ai/
[5] http://www.shazam.com
[6] http://www.soundhound.com/
[7] http://www.vocaloid.com/en/
[8] http://virsyn.com/en/E_Home/e_home.html
[9] http://www.music-ir.org/mirex/wiki/MIREX_HOME

address all these, we study the behaviour of the RS under diverse acoustic conditions, and we investigate possible ways to further improve this time-frequency representation, according to the goals of specific applications. The findings and insights acquired by this study are applied in order to propose concrete solutions that can be reused within the context of different applications. Specifically, these solutions fall in three categories, which are briefly introduced in the following.

(i) *Acoustic feature extraction* is the process of extracting sets of descriptors that represent specific properties of acoustic signals. Opposite to transformations, *e.g.* the Fourier transform, the feature extraction aims, first, at representing higher level characteristics and, second, at significantly reducing the signal dimensionality. Particularly in speech recognition and related applications feature extraction is a method to compress information for a successive effective statistical modelling of the feature vectors.

(ii) *Objective speech quality measures* are designed to predict the overall quality of a given speech signal. Initially, these measures were introduced in the speech coding community in order to evaluate the distortions introduced by speech codecs. Later, objective speech quality measures were used in early speech recognition systems, and for evaluation purposes in systems concerned with speech enhancement, noise reduction and dereverberation. Similar measures have been introduced for music and audio coding.

(iii) *Pitch contour extraction* refers to the detailed description of the harmonic content of speech, or singing voice. Generally speaking, pitch is a very important perceptual quality of sound that makes listeners able to order a given sound in a frequency related scale, as “lower” or “higher”. Pitch contours can be exploited in a wide range of applications, such as pitch detection algorithm (PDA), melody extraction, and music transcription.

The evaluation of the proposed solutions in the above areas, is performed with their inclusion in real applications and, in particular, the following.

(i) *Speech segmentation*, or speech alignment, refers to the segmentation and labelling of speech into its building blocks, *i.e.,* phonemes. This process was traditionally performed manually, but it is a time consuming and error-prone activity. The high demand for accurate and fast speech segmentation, for instance for the initialization of speech recognizers and the evaluation of the performance of speech recognizers, led to the design of automatic methods.

(ii) *ASR* refers to the process of converting spoken utterances into textual form. State-of-the-art ASR systems achieve excellent results under the constraint that the speech signal is recorded in a quiet environment, by microphones used in close proximity to the speaker mouth. Results start deteriorating quickly when we introduce some noise, or some distance between the speaker and the microphone.

(iii) *Channel selection (CS)* is a method that has been exploited in order to improve speech recognition performance in certain challenging scenarios, *i.e.,* when the speaker is located far from the microphone, and the acquired signal is degraded by various environmental factors. This scenario is commonly referred to as distant speech recognition (DSR) and a common group of solutions suggests, among other steps, to use signals recorded by multiple microphones and exploit the overlapping information they provide. In CS this is achieved by a comparison of the different signals, and the selection of the one that it is assumed to lead to the best recognition results.

(iv) *Melody extraction* is the task of estimating and tracking the fundamental frequency (f_0) of the main melodic instrument playing in a piece of music. A complete melody extraction system must detect the regions where the main melody is active, as opposed to the regions that the corresponding instrument stopped playing, and then provide a detailed temporal tracking of the frequency of this melody. This task in usually complicated due to the multiple instruments playing simultaneously in polyphonic music.

Summarizing, the main goals of this thesis are:

- To provide a comprehensive overview of the RS, and discuss its limitations.
- To suggest mechanisms to exploit the RS, for analysing and understanding speech and singing voice.
- To incorporate the proposed methods in final applications.
- To perform detailed evaluations and study the behaviour of the proposed techniques, with the use of various data sets and variable acoustic conditions.

1.5 Contributions

The specific contributions of this Ph.D. thesis comprise:

- A comprehensive literature review in various topics of interest, including, but not limited to, time-frequency representations, acoustic feature extraction and objective signal quality evaluation.

- A comprehensive discussion of the RS and its use for acoustic signal processing. The characteristics of the RS are experimentally demonstrated, and this is done under different acoustic conditions.

- A new reassigned front-end module for speech feature extraction, applied on speech segmentation and speech recognition. The proposed set of acoustic features is studied under different acoustic conditions and their properties are demonstrated.

- A new set of reassigned objective signal quality measures, which are exploited for the characterization of reverberant conditions.

- A new CS method, based on traditional and reassigned cepstral distance (CD)s. This represents part of a common work with Cristina Guerrero [Guerrero, 2016].

- A new pitch contour extraction method. The use of the RS enables a fine tracking of the harmonic components of the input signals, both in time and in frequency domains. The pitch contour extraction method is applied for melody extraction from polyphonic music signals.

- The dissemination of the proposed methods and related results:
  Georgina Tryfou, Marco Pellin and Maurizio Omologo. Time-frequency reassigned cepstral coefficients for phone-level speech segmentation. In Proceedings of the 22nd IEEE European Signal Processing Conference (EUSIPCO), 2014

  Cristina Guerrero, Georgina Tryfou and Maurizio Omologo. Channel selection for distant speech recognition exploiting cepstral distance. In Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2016.

  Cristina Guerrero, Georgina Tryfou and Maurizio Omologo. On the Use of Objective Signal Quality Measures for Channel Selection in Distant Speech Recognition. Submitted for publication in "Computer, Speech and Language" Journal, on November 9^{th} 2016. Submitted revised manuscript on May 2^{nd} 2017.

  Georgina Tryfou and Maurizio Omologo. A reassigned based singing voice pitch contour extraction method. In Proceedings of the 42nd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

  Georgina Tryfou and Maurizio Omologo. A reassigned front-end for speech recognition. Submitted in 25th IEEE European Signal Processing Conference (EUSIPCO), *2017.*

In addition, the following paper has been submitted for publication with a negative result: *Georgina Tryfou and Maurizio Omologo. A reassignment-based melody line extraction system for polyphonic music. Submitted in Transactions on Audio, Speech and Language Processing, May 2015.*

1.6 Thesis outline

The remainder of this thesis is organized as follows. In Chapter 2 we review various fundamental signal processing topics. The first part of the chapter is concerned with different acoustic signal representations, such as the waveform, the spectrum and various time-frequency representations. In the second part, we present details for the production and the characteristics of human voice, as well as differences between speech and singing voice. Following this, we overview the area of feature extraction from speech signals. Finally, we discuss the effects of certain acoustic conditions, in particular reverberation and music, in speech and singing voice signals.

In Chapter 3 we recall the theory behind the RS, and discuss the strengths and the limitations of this representation. We overview the most important implementation details and review the main applications within which RS has been exploited so far. Following this, we propose two different representations, built upon the RS, namely the reassigned cepstrum and the dominance reassigned spectrogram (DRS).

Chapter 4 is concerned with the use of the time-frequency reassigned cepstral coefficients (TFRCC) in speech segmentation and speech recognition. First, we summarize existing strategies in the fields of speech segmentation, ASR and DSR. After this, we describe the details of the proposed TFRCC feature set and, finally, we present experimental results from the use of these features in various contexts.

In Chapter 5, we investigate the RS as the time-frequency representation upon which objective speech signal quality measures are computed. After the review of state-of-the-art measures, which are extensively used for speech quality estimation, we propose reassigned alternatives. The remaining of the chapter is focused on the study of the various characteristics of objective speech quality measures, reassigned and not, particularly in relation to distortions introduced due to reverberation.

In Chapter 6, we exploit a subset of the signal quality measures discussed in the previous chapter, and in particular the CD and the reassigned CD (RCD)) in the context of CS for improving the recognition performance in a multi-microphone DSR system. First, we overview the literature in what concerns multi-microphone DSR and CS approaches. Then we discuss the details of the proposed CS method. Finally, we present experimental activities that exploit different corpora and DSR configurations.

The last contribution of this work, namely the extraction of harmonic pitch contours that describe the frequency components of singing voice, is presented in Chapter 7. We provide related work in the areas of pitch and melody extraction and, then, we describe the extraction of melodic contours from the DRS representation. Finally, various experimental activities study the ability of the proposed method to describe the melodic line of polyphonic music signals.

In Chapter 8 we summarize the contents of this work, we point out the main contributions and we draw our conclusions. Future directions of this research are envisioned before concluding this thesis.

Chapter 2

Acoustic signal processing

In this chapter we review the basic concepts of acoustic signal processing. First, in Section 2.1 we discuss the various signal representations widely used as a basis to perform further processing. In Section 2.2 the main characteristics of human voice, in speaking and singing modes are presented. In Section 2.3 some background in speech feature extraction is presented, along with details on the processing steps for the extraction of the most commonly used sets. The various acoustic environments that can complicate a simple scenario of speech or singing voice processing are overviewed in Section 2.4.

2.1 Signal representations

2.1.1 The waveform

Formally, a sound wave is a succession of different pressure levels which travels through a propagation medium, such as the air. It is created by a vibrating object, for instance the string of a musical instrument, and manifests by making other objects, for example the human eardrum, vibrate.

The simpler signal that acoustic signal processing may be concerned with is a sinusoid wave with a single frequency component, which is commonly called a *pure tone*. The waveform that represents a *sinusoidal* function of time t is given by

$$x(t) = \alpha \sin(\omega t + \phi) \quad , \tag{2.1}$$

where α is the maximum amplitude, ω is the angular frequency in radians per time unit, and ϕ is the phase angle with respect to the time origin, in radians. The amplitude α of the wave corresponds to the maximum air pressure deviation from the ambient atmospheric pressure caused by the propagation of this wave. The phase ϕ is the relative displacement of the sine wave from the origin of the axis. A positive phase shifts the sine waveform to the left introducing an advance, and a negative phase shifts the waveform to the right, introducing a delay or lag, with a time shift equal to ϕ/ω seconds.

Commonly, the angular frequency is replaced by the cyclical as follows

$$x(t) = \alpha \sin(2\pi f t + \phi) \quad , \tag{2.2}$$

where f is the frequency, defined as the number of times that a full cycle is repeated in a second. It is measured in cycles per second, or *Hertz.* The period T is the inverse of the frequency *i.e.,* $T = 1/f = 2\pi/\omega$, and it is measured in seconds, indicating the duration of a full oscillation of the sinusoid signal. In the above simple case, the sine wave comprises a single frequency. More generally a *complex* waveform is represented by a sum of N components, as follows

$$x(t) = \sum_{n=1}^{N} \alpha_n \sin(2\pi f_n t + \phi_n) \quad , \tag{2.3}$$

where α_n, f_n and ϕ_n are the amplitude, frequency and phase of the $n-th$ component, respectively. In such a complex wave, the greatest common divisor of the frequencies f_n is called *fundamental frequency*, or f_0. A complex waveform can be expressed as an infinite sum of pure tones, called *partials* or *harmonics*, with frequencies at integer multiples of the fundamental frequency

$$x(t) = \sum_{n=1}^{N} \alpha_n \sin(2\pi n f_0 t + \phi_n) \quad . \tag{2.4}$$

A notion that is closely related to the fundamental frequency, and particularly the perceptual aspects of it, is the *pitch*, often described as how *high* or *low* a sound is. The pitch of a complex tone is related to the pitch of a sinusoidal tone of the same f_0, but it is influenced by other factors, such as the timbre. More details on the notion of *pitch*, and other related topics are given in Chapter 7.

2.1.2 The power spectrum

The power spectrum is a representation of the distribution of energy of a waveform among the various frequency components. According to Fourier analysis, any "well-behaved" periodic signal can be decomposed into an infinite set of sines and cosines of discrete frequencies, called Fourier series. When we pass from periodic to nonperiodic functions, the Fourier series become the well known Fourier transform. For a continuous waveform the Fourier transform is given by

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt \quad . \tag{2.5}$$

In practice we normally deal with sampled, real valued discrete signals, which we denote $x(n)$, where n is the discrete time. The corresponding discrete Fourier transform (DFT) is given by

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}dn \quad . \tag{2.6}$$

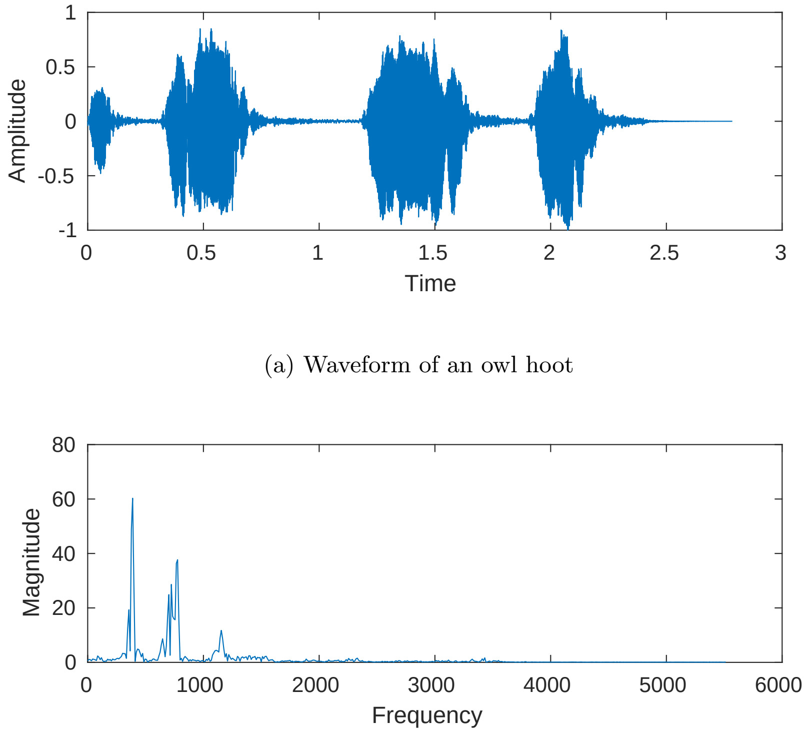
Both the continuous and the discrete Fourier transforms map the signal from the time to the frequency domain. The obtained representation is complex valued, and decomposes the signal into the frequencies that make it up. From the Fourier transform, the *power spectrum* is defined as $P(\omega) = |X(\omega)|^2$

The Fourier transform has a long list of interesting properties. First, it is a linear operation, *i.e.,* the complex spectrum of the sum of two signals is equal to the sum of their spectra. Linear operations performed in the time (or frequency) domain have corresponding operations in the frequency (or time) domain, which can be easier to perform. For instance, according to the *convolution theorem*, the Fourier transform maps the convolution operation between two signals into the point-wise product of their spectra. This means that a linear time-invariant system, such as a filter applied to a signal can be expressed as a multiplication operation in the frequency domain.

In the discrete case, the *Shannon theorem* states that the Fourier transform is meaningful for frequencies lower than $f_N = f_s/2$, where f_s is the sampling frequency. The maximum allowed frequency f_N is called Nyquist frequency. Another aspect in the discrete case concerns the complexity of its implementation. The direct calculation of the sum in (2.6) requires N^2 complex multiplications and $N(N-1)$ complex additions. This is drastically reduced by the *Fast Fourier transform* (FFT) algorithm, which performs only $(N/2)log_2(N)$ complex multiplications for the calculation of the same complex spectrum.

Problems of the waveform and the power spectrum

Let us consider a simple acoustic signal, for example a short owl hoot. Figure 2.1a shows how the amplitude of this signal changes over time, and Figure 2.1b shows the various frequency components of the signal in the magnitude spectrum. In the power spectrum, we can see high energy concentration around certain frequencies, which can be related to three frequency components. Although this is already a lot more from what we can deduce from the waveform, both of these representations are insufficient for a complete analysis of the hoot signal. An important reason for this is the fact that, in the creation of



(a) Waveform of an owl hoot

(b) Frequency representation of the same owl hoot

Figure 2.1: A short owl hoot represented in the time (a) and the frequency (b) domains.

the power spectrum, the *phase* has been discarded. As mentioned, phase is important in specifying the time-delay of a tone or in other words, in describing when each frequency component takes place.

The analysis and understanding of nonstationary signals demands more information than what is present in the waveform and the power spectrum. In the owl example, we would be interested to know how are the frequencies distributed in time; do the three frequency components appear sequentially in the three short phrases that can be seen in the waveform, or each phrase is a sum of more than one frequencies? In the latter case, it would also be important to know the extent to which each frequency component affects the final sum, *i.e.*, the relative intensity of each component.

2.1.3 The spectrogram

The main idea of the STFT is to analyse the nonstationary signals in short, consecutive segments, within which it can be assumed that their content does not change significantly. Instead of performing Fourier analysis one time over the whole excerpt, we can select to

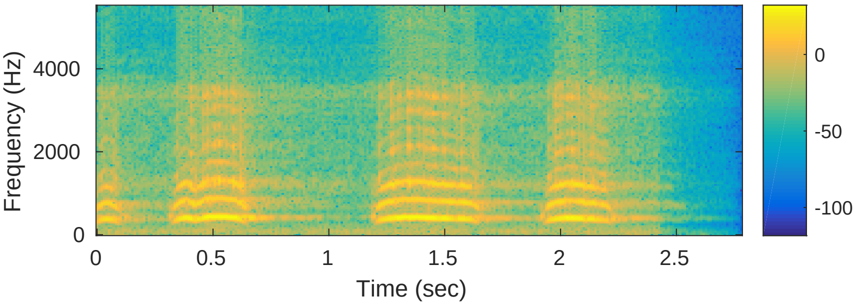


Figure 2.2: The power spectrogram of the owl hoot introduced earlier. The color is a representation of the magnitude of the energy in each time-frequency point, as shown in the color bar on the right.

cut it into many pieces, and concatenate the output of the Fourier analysis of each of these segments into a new representation. This visual representation of this process, called *spectrogram*, is three dimensional and assigns a certain energy, or intensity value for each point in time and in frequency. As shown in Figure 2.2, for the owl example, the spectrogram offers more information than the waveform and the power spectrum, as for instance how each frequency component evolves over time.

Here, it is useful to give more details in the notion of *framing*. The *frames* are portions of the signal, and are widely used in signal processing. They are created at specific time instants, with the process of *windowing*. Windowing aims at limiting the analysis scope of a method to a short duration, at which the signal can be considered *stationary*, *i.e.,* its properties do not change rapidly in relation to time. A frame of the signal $x(t)$ centred at the time instant t_0 is computed as follows

$$x_{t_0}^h(t) = x(t)h(t_0 - t) \quad , \tag{2.7}$$

where $h(t)$ is a window function, which can be, among other, Gaussian, Hamming, Hanning, or rectangular (although not commonly used) and typical frame durations for acoustic signals vary from $20ms$ to $100ms$.

From the framed signal, the STFT is built as the Fourier transform of the successive frames

$$STFT_x^h(t,f) = \int_{-\infty}^{\infty} x(\tau)h(t-\tau)e^{-j2\pi f\tau}d\tau \tag{2.8}$$

$$= M(t,\omega)e^{j\phi(t,\omega)} \quad , \tag{2.9}$$

where $M(t,\omega)$ is the magnitude of the STFT and $\phi(t,\omega)$ its phase. The spectrogram is then defined as the squared magnitude of the STFT

$$S_x^w(t,f) = |STFT_x^w(t,f)|^2 \quad . \tag{2.10}$$

In general, the spectrogram is converted to decibels (dB) using the formula

$$S_{x,dB}^w(t,f) = 10\log_{10} S_x^w(t,f) \quad , \tag{2.11}$$

in order to alter the dynamics and make them appropriate for visualization purposes. Due to the simplicity of its construction and interpretation, the spectrogram is the most widely used method to analyse non-stationary signals and, in many cases, the obtained representation clearly shows the structure and evolution of the signal.

However, the spectrogram can be problematic. First, as indicated by (2.10), it completely discards the phase, as in general, the information in the short-time phase spectrum is difficult to interpret. Nevertheless the short-time phase spectrum is known to contain important temporal information about the signal, and has been used for instance to improve frequency estimates of quasi-harmonic sounds [Dolson, 1986] and in the sinusoidal model which is largely based on the phase spectrum for the reconstruction of analysed signals [McAulay and Quatieri, 1986]. Second, the spectrogram suffers from the well-known trade-off between the temporal and frequency resolutions.

The uncertainty principle As discussed, the STFT can be viewed as the Fourier transform of a signal framed with an analysis window $h(t)$. In this case its temporal resolution, *i.e.,* the ability to distinguish two successive events is limited by the duration Δt_h of the analysis window. However, (2.9) can also be written as a function of the frequency domain of the signal, as follows

$$STFT_x^h(t,f) = \int_{-\infty}^{\infty} X(v)H(v-f)e^{j2\pi(v-f)t}dv. \tag{2.12}$$

In this case, the STFT can be interpreted as the output of a band-pass filter, in which case the frequency resolution , *i.e.,* the ability to distinguish between two closely located frequencies, is limited by the bandwidth Δf_h of the filter $H(f)$. The quantities Δt_h and Δf_h are linked by the Heisenberg-Gabor uncertainty principle, according to which the more concentrated a time function is, the more spread its Fourier transform must be, and vice versa. Although the notion of uncertainty mainly refers to a property of quantum mechanics, it can also be stated in terms of harmonic analysis as in [Folland and Sitaram, 1997]: “A nonzero function and its Fourier transform cannot both be sharply localized.”

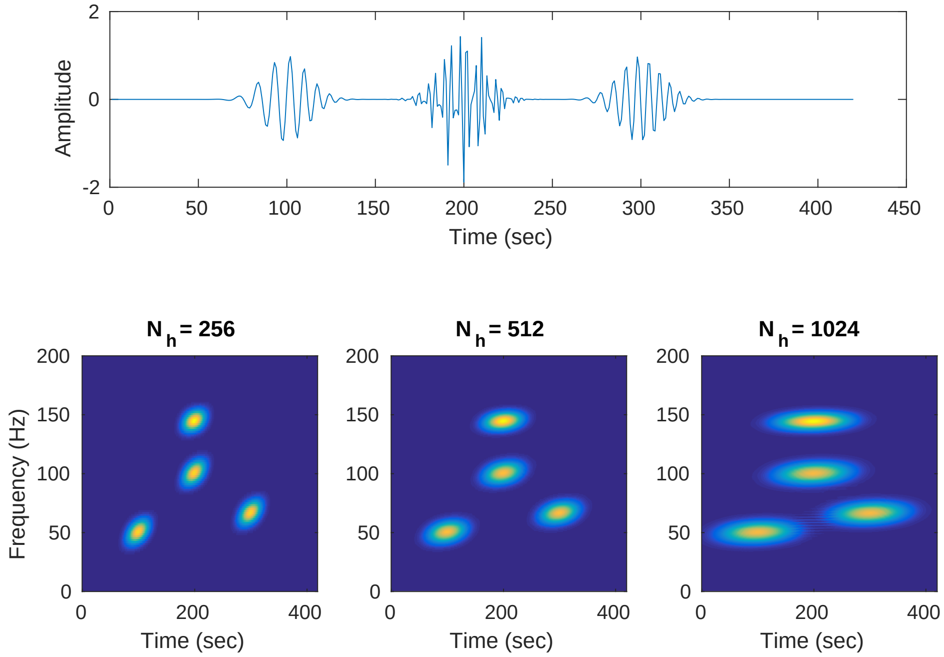


Figure 2.3: The time representation of the real part of an analytic signal (top) and the spectrogram of the same signal (bottom) calculated with three different analysis window lengths N_h.

Following Gabor's definition given in [Gabor, 1946]

$$\Delta t_h \Delta f_h \geq \frac{1}{4\pi} \quad . \tag{2.13}$$

Therefore, it is not possible to obtain simultaneously a very good temporal and a very good frequency resolution. This limitation is demonstrated in Figure 2.3, where we can observe that, according to the choice of the analysis window length, either the time (shorter windows) or the frequency (longer windows) localization of each component is improved.

2.1.4 Time-frequency representations

The motivation to analyse human speech in a systematic way led in the mid 40s to the introduction of the spectrogram [Koenig et al., 1946, Potter et al., 1947]. At the same time, works such as those of [Gabor, 1946], [de Ville et al., 1948] and [Page, 1952] led to alternative methods of analysing the spectra of time varying signals. The main idea behind all these works was always the same, *i.e.,* the derivation of a combined time-frequency representation, which shows the behaviour of a signal in both domains. Such a representation can be achieved by a time-frequency distribution (TFD), that describes the temporal evolution of the various signal components. TFDs are ideal tools to dissect, analyse and interpret complex signals whose spectral content is time-varying [Cohen, 1989]. In addition, methods that relate the time and frequency distributions facilitate the synthesis of signals that follow certain desirable attributes.

Before a deeper discussion on TFDs some related notions should be defined. First, a term that we used several times so far without a proper definition, and of critical importance in the study, and particularly the evaluation of TFDs, is the *component*. A component is a concentration of energy in either the time, the frequency or the time-frequency domains [Baraniuk et al., 2001, Cohen, 1992, Williams et al., 1991]. Pointing back to Figure 2.1, we can see for example 4 components comprising the waveform of the acoustic signal, and 3 components comprising its power spectrum. In the spectrogram of the same signalFigure 2.2, we can see multiple components appearing in different time-frequency regions. Therefore, it is relevant to notice that the representation used affects the visualization of the components that comprise a signal.

In the ideal case, a time-frequency representation should clearly represent the *instantaneous frequency (IF)* spectrum of each signal component. As already discussed, the frequency of a sinusoidal signal is defined as the number of cycles completed within one time unit. Although this is a well defined quantity for stationary pure tones, in practice signals are non stationary and the notion of frequency does not account for the time-varying nature of the signal. The IF of a signal is a concept created to address this limitation, and it defines the location of the spectral peak of a mono-component signal, as this varies with time [Boashash, 1992]. Other definitions for the IF can be found in the literature, as for instance in [Carson and Fry, 1937, Van der Pol, 1946] where IF of a component is defined as the rate of change of the component's phase angle at time t. Here, we assume the definition presented in [de Ville et al., 1948] according to which the IF of a signal $s(t) = a(t)\cos\phi(t)$ is given by

$$f_i(t) = \frac{1}{2\pi}\frac{\partial}{\partial t}\mathrm{arg} z(t) \tag{2.14}$$

where $z(t)$ is an analytic signal built as

$$z(t) = s(t) + jH[s(t)] \tag{2.15}$$

$H[s(t)]$ is the Hilbert transform[1] of $s(t)$. The complex spectrum of the analytic signal, given by

$$Z(f) = A(f)e^{j\theta(f)} \quad , \tag{2.16}$$

[1] The Hilbert transform [Hildebrand, 1949] is a linear operation commonly used to obtain the analytic representation of a signal. It can be thought of as the convolution of the signal with the function $h(t) = 1/(\pi t)$.

is used for the definition of the *group delay (GD)*, as follows

$$\tau_g(f) = -\frac{1}{2\pi}\theta(f) \quad . \tag{2.17}$$

The GD represents the time delay, as a function of the frequency, of an impulse passing through a linear filter with impulse response $h(t) = s(t)$. GD is an interesting attribute, as the function $\tau_g(f)$ describes the localization of the various components of the signal in the time domain. In the ideal case, the two notions defined above, *i.e.,* the IF and the GD pave the way for the definition of a time-frequency representation, since they enable a complete characterization of a signal by its IF at a given time, or the main time at which a frequency appears. Such theoretical representations however, that localize the energy at the points $(t, f_i(t))$, or $(f, \tau_g(f))$, are not of practical use since they do not extend in the case that a signal is a sum of several components localized at the same time or the same frequency.

Of central importance in the field of time-frequency representations is the Wigner-Ville distribution (WVD) [de Ville et al., 1948, Wigner, 1932]. Two main reasons contribute to this, first the fact that the WVD is a highly concentrated distribution in time and frequency, which means that it shows exactly the IF of a frequency modulated sinusoid, *i.e.,* a linear chirp. This is in fact related to the second reason of its importance, which is that the WVD only depends on the signal and it is not affected by the choice of an analysis window. The WVD has, by definition, a better resolution than the spectrogram:

$$W_x(t,\omega) = \int_{-\infty}^{+\infty} x(t + s/2)x^*(t - s/2)e^{-j\omega s}ds \quad . \tag{2.18}$$

This definition can be interpreted as the STFT of a signal using the analysis window $h(t) = x(-t)$, *i.e.,* the time-reversed version of the analysed signal [Flandrin, 1998]. However, the difficulty in physically interpreting the often negative values of the WVD output makes the WVD having very little practical use. Furthermore, the WVD is prone to noise and generates cross-components that can even mask components of interest, in multi-component signals. Such masking components, which are oscillatory in nature could be reduced with low-pass smoothing. Of course, such a smoothing spreads out the perfectly localized components of the WVD. A class of such smoothed WVD is defined in [Cohen, 1995], as bilinear time-frequency representations

$$C_x(t,\omega) = \int\int W_x(\tau,\nu)\Phi(\tau - t, \nu - \omega)d\tau d\nu \quad , \tag{2.19}$$

where $\Phi(t,\omega)$ is a smoothing kernel designed to suppress noise and cross-components.The spectrogram can be viewed as a member of the Cohen's class of distributions, with the

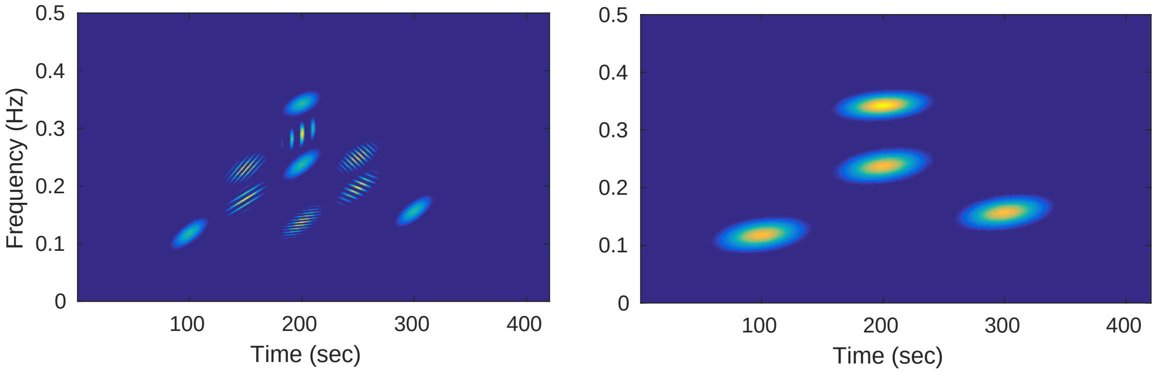


Figure 2.4: Wigner-Ville distribution (left) and spectrogram (right) of a signal that comprises four components, each with Gaussian amplitude and linear frequency modulation.

WVD of the window function $h(t)$ used to smooth the WVD of the signal [Flandrin, 1998]

$$S_x^h(t,\omega) = \int\int_{-\infty}^{+\infty} W_x(\tau,\nu)W_h(\tau - t, \nu - \omega)\frac{d\tau d\nu}{2\pi} \quad . \tag{2.20}$$

The above smoothing means that the spectrogram values do not express the energy at a certain point (t, ω) of the time-frequency plane. Instead, each point results from the summation across a whole distribution of values. The obtained sum is assigned to the *geometric center* of the time-frequency domain.

Evaluation of TFDs In Figure 2.4 the WVD and the spectrogram of the same multicomponent signal are presented. It can be observed that the WVD localizes the components much better than the spectrogram, but it introduces six additional cross-components. The spectrogram on the other hand does not present any interference, but the auto-components have a much worse localization.

From this illustration it becomes evident that the adequacy of a TFD to represent a particular signal relates to (i) the suppression of TFD cross-components, (ii) the concentration and resolution of autocomponents, and (iii) separation of signal components, such as parallel chirps, that overlap in both time and frequency. Measures that have been used for the evaluation of TFDs include mainly moment-based measures, for instance the time-frequency bandwidth [Boashash and Sucic, 2003], measures of information stemming from probability theory [Sang and Williams, 1995, Williams et al., 1991], and parametric decomposition techniques [Orr, 1991]. Classical moment-based measures though have been criticised for not really measuring signal complexity and information content, as in a commonly used example of a signal that comprises two components of compact support, such measures constantly increase when the separation of the components increase, while the signal complexity does not. On the other hand, measures borrowed from probability theory are more promising. The analogy between TFDs and the probability densities makes the classical Shannon entropy [Shannon, 2001] a very good candidate to measure the amount of information encoded in a signal, as this information is represented by the TFDs. Using the same signal, and changing its time-frequency representation, a peaky representation that has only a small number of components will result in lower entropy values, compared to the diffuse and more complex, due to the appearance of cross-components, representations. Nevertheless, the negative values that can appear in most of the time-frequency representations prohibit the use of the Shannon entropy. This problem was sidestepped in [Williams et al., 1991], where the generalized Rènyi entropy [Rényi et al., 1961] was employed for the evaluation of TFDs. The Rènyi entropy of order α is defined as

$$R_a = \frac{1}{1-\alpha} \log_2 \sum_{l=-L}^{L} \sum_{k=-K}^{K} [C_x(l,k)]^{\alpha} \quad , \tag{2.21}$$

where $C_x(l,k)$ is a time-frequency representation. It has been shown that better TFDs are those with smaller uncertainty measure [Baraniuk et al., 2001, Sang and Williams, 1995], often calculated for $\alpha = 3$. As an example, the Rènyi entropies of the spectrogram presented in Figure 2.3 are, from left to right, 10.62, 10.08 and 10.49 bits respectively. For the Figure 2.4, the WVD corresponds to 11.57 bits, and the spectrogram to 10.08 bits.

2.2 Human voice

"Human voice" is the result of a series of sounds, created by a human using the voice production organs, in order to speak, sing, laugh and more. The automatic analysis and understanding of human voice is of interest in a wide range of applications. In this thesis, we are particularly interested in speech recognition, *i.e.,* the automatic understanding of the content of speech, and in singing voice melody extraction, *i.e.,* the automatic identification of the predominant frequency of a singing voice. However there is a very long list of other voice related applications such as speaker identification, forensic analysis, emotion recognition, speech/singing synthesis, to mention only a few.

2.2.1 Voice production

A schematic diagram of speech production is presented in Figure 2.5. From the components shown there, the voice organs comprise the lungs, the larynx, the pharynx, the nose

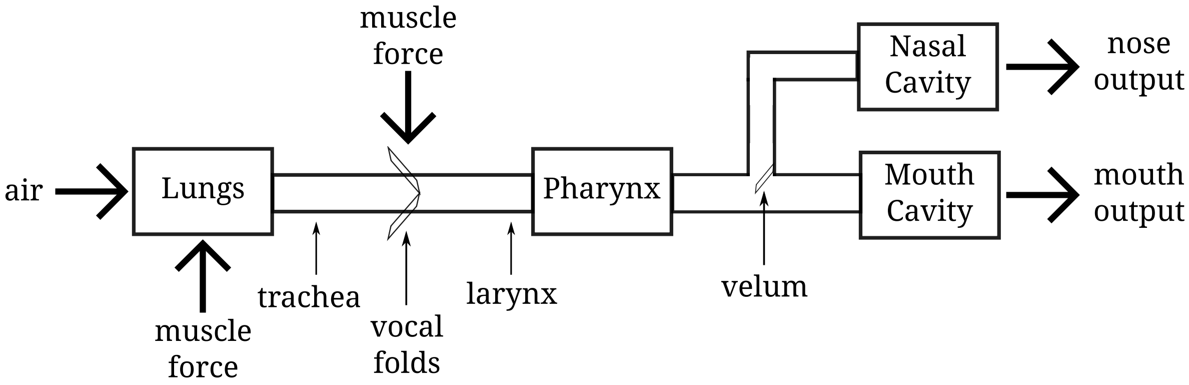


Figure 2.5: Schematic representation of the voice production mechanism, based on [Flanagan et al., 1970]

and the mouth cavities [Sundberg et al., 1977]. The first part of the voice organ, the lungs, are responsible to power the voice production apparatus with air. The vocal folds are thin membranes which can be controlled by a complex structure of muscles [Orlikoff and Kahane, 1996], and are located in the bottom end on the larynx. The tube-like larynx leads to a wider cavity, called the pharynx, which in turn leads to the mouth cavity. At the top end of the pharynx, the velum is the "door" to the last part of the voice organ, the nasal cavity. The larynx, pharynx and mouth cavity form the *vocal tract.* The role of the vocal tract is to "shape" the produced sound, a process done with a series or *articulators:* the larynx, the lips, the jaw, and the tongue.

Once the intention to produce voice has been transmitted by the brain to the voice organs, the lungs expand in order to produce an excess of air pressure. This process is called *breathing.* The air passes through the trachea and encounters the vocal folds. The next step of speech production, called *phonation*, is to convert the air pressure coming form the lungs into raw audible sound, which is called *voice source.* There are several types of phonation [Fulop, 2011] and in the most common one, called *voicing* the produced sound is a quasi-periodic wave. During *voicing*, in preparation to produce sound the vocal folds are adducted. The positive air pressure from the lungs forces them to open momentarily, but the Bernoulli effect brings them back together. This sets the folds into a self-sustaining oscillation, *i.e.,* they open and close periodically. The rate at which the vocal folds open and close determines the fundamental frequency of the voice. As shown in Figure 2.6, the f_0 of speech is affected by the gender and the age of the speaker. In singing voice these ranges vary much more, as discussed in the next section.

Although most of the content in speech as well as in singing demands voiced phonation, aperiodic or transient sounds are also needed. These type of phonations are produced when the air passes through the open vocal folds.

At the next step of speech production the vocal tract, and in case the nasal cavity,

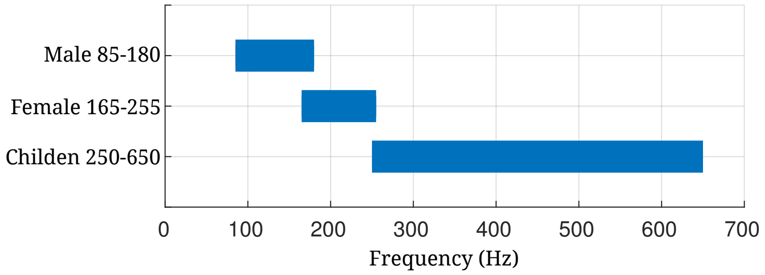


Figure 2.6: f_0 ranges for different genders and age groups. Infants' f_0 has been measured as high as 1000Hz, while a 10 year old child typically has an f_0 around 400Hz.

shapes the produced raw sound wave. The frequencies at which the vocal tract resonates the voice source are called *formants*. In practice, the uniformly slopping spectrum of the voice source is disrupted, and peaks are imposed at the frequencies of the formants, as shown in Figure 2.7. This shaping of the envelope of the voice spectrum is what results in distinguishable sounds. The particular frequencies at which the formants appear are characteristic of each sound. For example, the vowel /ae/, as in the word *bat* is associated with formant frequencies at 660Hz, 1720Hz and 2410Hz. The formants of the /oo/ vowel as in *boot* are in the frequencies 300Hz, 870Hz and 2240Hz. These frequencies are determined by the shape of the vocal tract, which can change in a rather complicated way in order to shift the formant frequencies. The jaw, the body of the tongue and the tip of the tongue are the main articulators that facilitate this change. Each configuration of these corresponds to a set of formant frequencies, which in turn is associated to a specific sound.

The "source-filter" model From the above description of voice production process, and a view similar to the one depicted in the right part of Figure 2.7 the well known"source-filter" model [Fant, 1971, Joos, 1948] has been inspired. According to this, a speech signal $x(n)$ is created when a sound wave is filtered through the vocal apparatus, as follows

$$x(n) = v(n) \star p(n) \quad , \tag{2.22}$$

where $v(n)$ is the impulse response of the filter related to the vocal and nasal tract, and $p(n)$ is the periodic excitation at the vocal folds, *i.e.,* the source signal. The source sound wave can be voiced, unvoiced or a combination of the two, according to the mode of the vocal folds. The source-filter model is a powerful engineering model because it allows the modelling of speech using only three parameters: the voicing state, the fundamental frequency (in the case of voiced speech) and the vocal tract parameters [Loizou, 2013]. Even in its simpler form the source-filter model is successfully used for speech/singing voice synthesis and low-bit-rate speech coding.

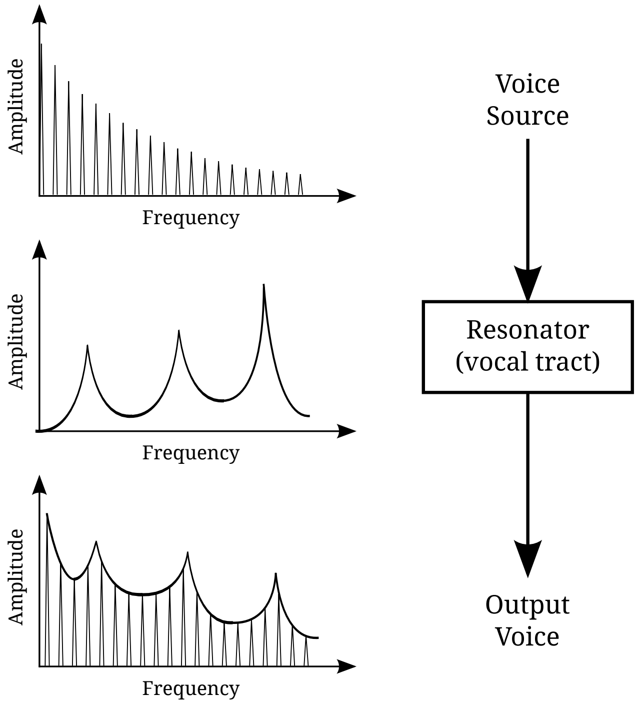


Figure 2.7: Schematic representation of the resonance process.

2.2.2 Singing vs. speech

Speech is the most important means for humans to transmit messages between each other. In spoken language, the main message that is communicated is the underlying thought, which turns into phone units, words and eventually spoken sentences. Speech transmits additional cues about the speaker's personality, emotions, background, health and education. On the other hand, singing is the process during which the human voice is used to produce musical sounds. During singing, apart from the semantic information of the lyrics, a great part of the conveyed message regards the melody and rhythm of the song. Similar to speech, information about the background of the singer, the emotional state and even her musical training is also transmitted through singing. Both singing and speech are produced by the same process, which means that the two share many characteristics. Nevertheless, the differences are several and particularly interesting to discuss.

In terms of pitch, singing voice has a higher average and a wider range. A trained singer should have a range of about 2 octaves, an excellent may have even up to 3 octaves. In Figure 2.8 typical f_0 ranges for different singing styles are presented. In both cases, *i.e.,* during speech and singing, physiology plays an important role to the produced pitch, but in speech it is affected unintentionally by the current mood of the speaker, while in the case of singing voice it is defined by the singer's training and the range of

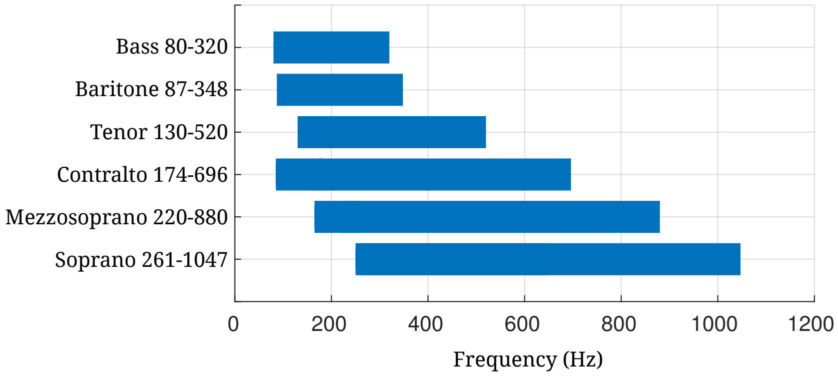


Figure 2.8: Average variations of the f_0, for singing voices.

the composed melody. For what concerns loudness, speech has a lower average and more limited dynamics than singing voice. Furthermore, in singing voice approximately 95% of the total duration is voiced, as opposed to 60% in speech. In speech the same vowels will present very similar formant structure among different speakers; in singing it is often necessary to change the position of the first two formants of the vowels in order to match the target pitch. In extreme cases, for instance in a soprano voice, this formant repositioning can cause the vowels to lose their intelligibility. Finally, the vocal training of professional singers results in a more regular use of the vocal folds, a fact that leads to differences in the source signals and in the breathing mechanisms.

The way that a singer reproduces syllables, having to manage concurrent changes in the reproduced notes, adds another important element of difference between singing voice and speech for what concerns the *prosody* [Deutsch, 2010, Taylor, 2009]. Opposite to this description of prosody in singing voice, in the case of speech the term is used to describe the rhythm and intonation followed by the speaker, which is highly influenced by the emotional state and the contextual information that ones wants to transfer with a sentence. Another term used in both cases and describes different phenomena is the term *articulation*. In the case of singing voice, articulation refers to the different techniques that singers may use in order to perform a phrase or a passage, while in speech the term describes the exact movement of speech organs that produce the different speech sounds.

Apart from the differences between spoken and singing voice signals, in the latter we observe certain characteristics that are not present in speech. The *singer's formant* results from the need of the singers, particular in classical and concert music, to be audible at a great distance from the listeners, without using any amplifying system, while an orchestra is simultaneously playing. Singer's formants are, in general, created with the grouping of the higher order formants together. Furthermore, in order to add expressiveness to an execution, singers often perform effects such as *tremolo* and *vibrato*. Tremolo is the trembling effect that is created when the amplitude of a sung tone changes following a certain frequency, and vibrato refers to small fluctuations from the target tone; again following a certain modulation frequency and amplitude.

2.2.3 Speech sounds

The smallest distinctive units of speech, and singing voice, are called *phonemes*. Each language has a distinct set of phonemes, but they can always be categorized based on the type of the source signal: periodic, noisy, or a combination of the two. Additionally, they can be categorized in terms of the manner of articulation, for example the place of the tongue and the degree of construction in the vocal tract. Focusing on the English language, and in particular the American pronunciation, we can identify a set of about 40 phonemes, which are shown in Table 2.1. The first group, the *vowels* and *glides*, are always voiced and very often bare higher energy than the phonemes of the other groups. The main cue for the perception and identification of a vowel or a glide are the frequencies of the formants and in particular the first three. Other interesting characteristics of vowels, which play an important role in the intelligibility of speech, are the duration and the formant transitions in the beginning and the end portion of the phoneme [Loizou, 2013]. Among all these sounds, the phonemes /w/, /l/, /r/ and /y/ are called glides, or semivowels, because despite their vowel-like characteristics are classified as consonants. For the production of the phonemes in the next group, *i.e.,* the nasals, the velum obstructs the entrance to the mouth cavity, forcing the air to travel through the nasal cavity. This elongates the vocal tract, leading to the appearance of lower resonant frequencies, and low intensities of the formant of higher frequencies. Again, the formant transitions in the beginning and the end of the nasals are very important for identification.

Opposite to vowels, glides and nasals, the next two groups, *i.e.,* stops and fricatives, are characterized by more restricted, or even totally obstructed airflow. Stops, or plosives, are produced in two steps, which are also very important cues for their identification. First, a complete obstruction of the vocal tract results in a complete silence in the acoustic signal, called closure. After that, the released air causes a transient noise called aspiration. Concerning fricatives, the main characteristic of this class is the presence of aperiodic noise, extending in relatively long region. The intensity of this noise, the shape of their spectra and the formant transitions in the beginning and the end parts are all important cues for the intelligibility of the different fricative sounds.

Class	Symbol	Word	Class	Symbol	Word
Vowels/	iy	beat	Nasals	m	mom
Semivowels	ih	bit		n	none
	eh	bet		ng	sing
	ae	bat	Stops	b	bet
	aa	Bob		d	dad
	er	bird		g	get
	ax	about		p	pet
	ah	but		t	Tom
	ao	bought		k	kite
	uw	boot	Fricatives	v	vet
	uh	book		dh	that
	ow	boat		z	zebra
	ay	buy		zh	azure
	oy	boy		f	five
	aw	down		th	thing
	ey	bait		s	Sam
Glides	w	wet		sh	shoe
	l	lid		jh	judge
	r	red		ch	chew
	y	yet		h	hat

Table 2.1: The phonemes of American English, categorized into 5 broad groups.

2.3 Speech feature extraction

The parametrization of the speech signals, or feature extraction, is designed to discard information that is considered irrelevant to the addressed task, as for instance the discrimination of the various speech units in speech recognition. In the following we review of the most interesting aspects of speech feature extraction, and discuss the exact methodology for the extraction of the most commonly used parameters, namely the Mel frequency cepstral coefficients (MFCC)[Davis and Mermelstein, 1980, Mermelstein, 1976] and perceptual linear prediction (PLP) [Hermansky, 1990] coefficients.

2.3.1 Short-time frequency analysis

Short-time frequency analysis has been extensively used in the majority of speech processing front-end techniques, since it was first introduced in 1940s [Koenig et al., 1946]. As discussed in Section 2.1.3, the basis of short-time analysis is the framing process. In front-end methods for speech recognition and related applications, the input signal is multiplied with an analysis window with a duration 15-40ms (frame size). The selection of the analysis window length is critical, due to the discussed trade-off between temporal and frequency resolution, and it is further complicated by the nature of the speech signals. For instance, during voiced speech the frame must by long enough so that it is not affected by the phase of the glottal cycle. Typically, one needs at least two glottal cycles in each signal frame. However, a very long analysis window blurs impulsive effects, such as the sudden burst of stop consonants. For every new frame the window is shifted 5-15ms (frame shift). Apart from the duration and shift of the analysis window, the shape of it is an important attribute, since it defines the spectral characteristics of the analysed signal. In speech literature, many windowing function have been proposed, as for example the Hanning, Blackman, Kaiser, and Bartlett windows [Oppenheim and Schafer, 1989]. However, the Hamming window, defined as

$$h(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N}), & \forall 0 \leq n \leq N \\ 0, & \text{otherwise} \end{cases} \tag{2.23}$$

is the most commonly used one. After the windowing and segmentation of the speech signal the mapping to the frequency domain is traditionally done with the use of the STFT.

2.3.2 The cepstrum

Cepstrum processing was initially introduced for seismic data analysis [Bogert et al., 1963], and soon after was applied for vocal pitch detection [Noll, 1964]. Since then, cepstrum has been introduced to speech recognition [Davis and Mermelstein, 1980], speaker verification [Furui, 1981] and a wide range of other speech analysis applications. In general, cepstrum can be considered as a distinct transform domain, which results from the inversion of the frequency domain. This inversion also inspired the name *cepstrum*, which stems from the reversal of the first four letters of the word spectrum. Formally, the *complex cepstrum* is defined as [Rabiner and Schafer, 1978]

$$\hat{x}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}(e^{j\omega}) e^{j\omega n} d\omega \tag{2.24}$$

where $\hat{X}(e^{j\omega})$ is the complex valued logarithm of the Fourier transform of the analysed signal. Usually, the *cepstrum* is calculated from the magnitude of the complex logarithm instead:

$$c(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |log X(e^{j\omega})| e^{j\omega n} d\omega, \tag{2.25}$$

which can be shown to be equal to the even part of the *complex cepstrum.* This is approximated by computing the inverse DFT of the logarithm of the magnitude of the

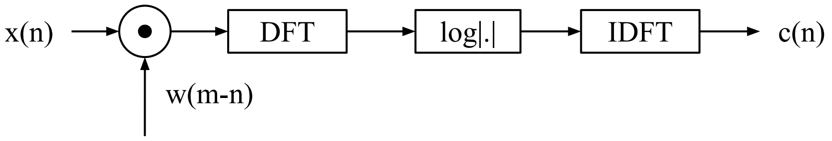


Figure 2.9: The computation steps of the short-time real cepstrum.

DFT of the input signal

$$c(n) = \frac{1}{N} \sum_{k=0}^{N-1} |logX(k)| e^{j\frac{2\pi}{N}kn}, \quad 0 \leq n \leq N-1. \tag{2.26}$$

The computation steps of the short-time cepstrum are shown in Figure 2.9.

The further processing of the spectrum before its inversion, leads to the most popular sets of features for acoustic signals and in particular speech. This can be attributed to some of the characteristics of the cepstrum, as for instance the different meanings of the cepstral coefficients based on their order, in particular the low-order coefficients. For example, the 1_{st} coefficient, often called $0-th$ order, represents the average energy of the input signal. The next value indicates the balance of energy between low and higher frequencies. A negative value shows that most of the energy is concentrated in the higher frequencies, indicating the possible presence of a fricative. Positive values on the other hand, translate in a higher energy concentration in the lower frequencies, as is expected in the cases of vowels, nasal and other resonant sounds [Deng and O'Shaughnessy, 2003]. Higher order coefficients increase the details of the spectral structure represented in the cepstrum, but it is a well known fact that beyond the $12-th$ order coefficient they do not increase the accuracy of systems such as speech recognition [Huang et al., 2001].

Another characteristic of the cepstrum, which makes cepstral features particularly successful in speech modelling, is the possibility to eliminate the effects of the periodic excitation produced by the vocal chords. This can emphasize the spectral envelope of the vocal tract, an attribute which is particular helpful in phoneme discrimination. In more detail, we consider the source-filter model of speech production, as in (2.22). In the log spectrum domain this is expressed as

$$logX(e^{j\omega}) = logV(e^{j\omega}) + logP(e^{j\omega}) \tag{2.27}$$

which then, taking the inverse Fourier tranform is rewritten as

$$\hat{x}(n) = \hat{v}(n) + \hat{p}(n) \quad . \tag{2.28}$$

In [Oppenheim and Schafer, 1989] it is proven that, when $p(n)$ is a periodic excitation with a period T_0, $\hat{p}(n)$ is also periodic with a period $N_0 = \frac{T_0}{T_s}$, where T_s is the sampling period. This makes $\hat{p}(n)$ non-zero only at the points $\hat{p}(kN_0)$, a fact that makes possible the perfect recovery of $\hat{v}(n)$ from the *liftering* operation, which is

$$\hat{v}(n) \approx \hat{x}(n)l(n) \quad , \tag{2.29}$$

where

$$l(n) = \begin{cases} 1 & \forall 0 \leq n \leq N_0 \\ 0 & otherwise \end{cases} \tag{2.30}$$

In the case of speech, $v(n)$ is the impulse response of the speaker's vocal tract and $p(n)$ the periodic vocal chord excitation during voiced speech. (2.29) indicates that the spectral envelope of the vocal tract can be separated from the periodic excitation of the vocal chords by discarding the higher order cepstral coefficients.

When cepstral coefficients are used as the basis for speech feature extraction, the inverse DFT can be replaced by the inverse discrete cosine transform (IDCT) which has been found to decorrelate the obtained sequence. Another very commonly used alternative is the calculation of the cepstral coefficients from the linear prediction coefficients (LPC) a_n, in a recursive manner as follows

$$c(n) = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c(k) a_{n-k} \quad 1 \leq n \leq p \quad , \tag{2.31}$$

where p is the order of the LPC analysis. It is noted here that LPC analysis is a powerful tool in audio and speech signal processing, according to which the spectral envelope of a signal is represented through the parameters of linear prediction. Detailed reviews of LPC can be found in [Makhoul, 1973, Markel and Gray, 2013].

2.3.3 Pre-emphasis

The human auditory system perceives loudness in a manner that varies with frequency [Dirks et al., 1982, Dubno et al., 1984], usually demonstrated as a set of equal-loudness curves [Ott and Longnecker, 2008]. In order to model the human ear sensitivity, and also flatten the input signal spectrum for specific purposes, a finite impulse response filter, called *pre-emphasis* filter is commonly applied. Its transfer function is given by

$$H_{pre-emphasis}(z) = 1 + a_{pre-emphasis} z^{-1} \quad , \tag{2.32}$$

where $a_{pre-emphasis}$ is a the pre-emphasis coefficient usually in the range $-1 \leq a_{pre-emphasis} \leq -0.95$.

2.3.4 Bark and Mel filter banks

Empirical evidence has shown that recognition performance can be improved if the frequencies are modelled in a manner similar to what is done by the human auditory system. In particular, the cochlea, inside the inner ear, resolves the spectrum in a non-linear way, which can be replicated by the use of specially designed filter banks[Huang et al., 2001]. Two commonly used filter banks are implemented in the Bark and Mel frequency scales. The Bark scale is given by [Zwicker, 1961]

$$bark(f) = 13 \arctan(0.00076f) + 3.5 \arctan\left(\left(\frac{f}{7500}\right)^2\right) \quad . \tag{2.33}$$

An alternative non-linear scale, which models the pitch perception characteristics of the auditory system is given by [Stevens and Volkmann, 1937]

$$mel(f) = 1127.01048 \log\left(1 + \frac{f}{700}\right) \quad . \tag{2.34}$$

In the context of feature extraction for speech segmentation and recognition non-linear scales such as the Bark of Mel-scales are usually applied with the use of multiple passband filters. For instance, the Mel filter-bank is defined by a set of triangular filters, each averaging the spectral energy around its center frequency.

2.3.5 Derivatives and other transformations

In both PLP and MFCC analysis, as preprocessing for speech recognition, it is a common practice to extend the feature vectors with their first and second order derivatives, in order to encode their dynamic properties [Furui, 1981]. This is usually performed with the application of a simple regression formula, that considers a certain number of neighbouring values. Finally, the features are very often normalized, for example employing cepstral mean normalization (CMN), variance normalization or band-pass filtering. Additional transformations aim at the reduction of the effects caused by environmental conditions, for example noise and reverberation, and the variabilities that exist among different speakers [Cohen et al., 1995, Welling et al., 1999].

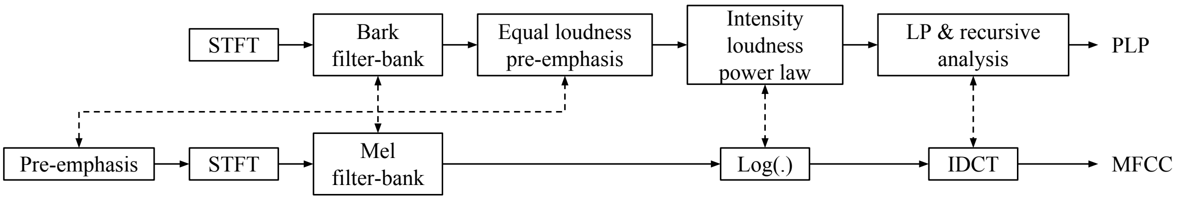


Figure 2.10: The steps of extracting PLP (top) and MFCC (bottom) features. The dashed arrows indicate the analogous processing steps.

2.3.6 MFCC and PLP features

As mentioned above, MFCC and PLP are the most popular choices as a front-end for statistical speech segmentation and recognition systems. PLP features have been reported to be more robust in cases where a mismatch between the training and the testing material exists, while MFCC features have been found to perform better under clean and match conditions [Davis and Mermelstein, 1980, Milner, 2002]. There have been attempts to combine the most interesting characteristics of the two sets of features [Hönig et al., 2005, Milner, 2002], showing that the computation method of both can be further improved.

The block diagrams of the extraction steps for the two sets of acoustic features are presented in Figure 2.10(a). As depicted there, the processing is highly comparable. Both sets derive from the application of the STFT on the acoustic signal and the computation of the magnitude of each frequency bin. This results in the complete loss of the phase information, as well as in a possible loss of accuracy in the power spectrum estimation. In the case of MFCC, a pre-emphasis filtering is applied on the time-domain. In the case of PLP, the pre-emphasis takes place in the spectrum domain, according to an equal-loudness function. The subsequent frequency band analysis comprises the application of a Mel filter-bank in the MFCC computation and a Bark filter-bank in the PLP computation. As discussed, both scales, Mel and Bark, are perceptually inspired and in practice the differences between the resulting filter-banks are negligible. Higher frequencies components are emphasised and more filters are allocated for the lower frequencies. Concerning the intensity law (PLP) and the logarithmic compression (MFCC), both stages model the non-linear relation between the intensity of the sound and its perceived quality. The result of the two approaches has again a very similar effect.

In the last step of analysis the two methods differ significantly. MFCC are computed with the application of the IDCT in the log filter-bank output. Since the filter-banks are overlapping, the output energies are highly correlated with each other. The IDCT decorrelates the energies, a very important step for the subsequent use of the features within a statistical framework.

In PLP analysis the auditory warped filter-bank output is further processed with inverse DFT, a step that yields the autocorrelation function. The values of the autocorrelation function are needed to compute the parameters of an all-pole LP model, which approximates the spectral envelope of the signal. It is noted here that the estimated envelope has sharp peaks, which are decreased with the intensity loudness power law, which raises the power spectrum coefficients to the power of 0.33.

2.4 Acoustic conditions

In everyday situations, sounds originate from a variety of sources simultaneously. For instance, in a very common situation a user is giving voice commands through a microphone, while many people are speaking in the background, a radio is playing some music, a door opens and closes, a phone is ringing, and cars are passing by outside. Each of the sources is reflected and attenuated by various surfaces in the environment, such as walls and furnitures. As a result, the microphone captures a complex mixture of all the sound sources and multiple attenuated reflections of each. For normal listeners, the problem of sorting out and making sense of each source, while focusing on the sound of interest, does not demand a lot of effort. On the contrary, acoustic signal processing systems suffer a lot from the presence of secondary sound sources, *i.e.,* noise, and various reflections, *i.e.,* reverberation.

Several methods have been proposed in order to address noise, reverberation, or both. For instance, certain methods attempt to *separate* the source of interest from the rest of the acoustic contents, while other systems attempt to reduce the effects of the acoustic environment and *enhance* the quality of the source of interest. In-depth reviews of methods such as source separation, dereverberation and speech enhancement, can be found in [Benesty et al., 2005, Naylor and Gaubitch, 2010, Pedersen et al., 2007]. In other applications, knowledge from the characteristics of the distorted signals can be exploited to modify the default processing that takes place. An example here is the use of noise-, or reverberation-robust features in speech recognition performed within noisy and reverberant environments [Benesty et al., 2005].

Overall the acoustic environment, and the characteristics of the various sources, play an unquestionable role in the design and implementation of such methods. In this section we discuss reverberation and music, the two main sources of distortion that are faced when analysing speech and singing voice respectively, and therefore are of great interest in the work presented in the subsequent sections.

2.4.1 Reverberation

Reverberation is created when acoustic signals are reproduced in non-anechoic environments, *i.e.,* enclosures composed of multiple surfaces that reflect sound signals. In such an enclosure, an emitted signal will reach a listener, or a recording device, through multiple paths. This results in a deterioration of the quality that characterizes reverberant signals. Formally, the signal received by the listener will be the sum of many delayed and attenuated version of the source signal $x(t)$, as follows

$$x'(t) = \sum_k a_k x(t - t_k) \quad , \tag{2.35}$$

where a_k is the attenuation, and t_k the time delay that corresponds to the $k - th$ sound reflection, both of which only depend on the characteristic of the acoustic environment. Therefore, we define the room impulse response (IR) as

$$r(t) = \sum_k a_k \delta(t - t_k) \quad , \tag{2.36}$$

where $\delta(t - t_k)$ is a delayed Dirac function. The signal received by a recording device can then be described as

$$x'(t) = r(t) \star x(t) \quad . \tag{2.37}$$

When a sound travels from one point to another, within an indoor non-anechoic environment, the IR fully describes the changes that the sound signal undergoes [Kuttruff, 2007]. The estimation of the IR that characterizes a certain acoustic scenario has been extensively discussed in the literature with various estimation methods, for example the methods *maximum length sequence* [Schröder, 1975], *linear sine sweep*, and *exponential sine sweep* [Farina, 2000]. Alternatively, IRs can be created in a synthetic way with a method known as image method (IM), assuming a shoe-box geometry for a simulated room [Allen and Berkley, 1979, Peterson, 1986].

A room IR, measured in a real acoustic environment as described in [Ravanelli et al., 2012], is depicted in Figure 2.11. As shown there the IR can be split in three parts, each affecting the received signal in a different way [Kuttruff, 2009, Yoshioka et al., 2012]. The first is the *direct sound*, which is a delayed version of the source signal. After the arrival of the direct sound, the *early reflections* are the very first attenuated instances that arrive for the next 50ms. After that, the *late reverberation* comprises numerous and very similar reflections. In general, the early reflections boost the energy of the direct sound and therefore benefit not only the human auditory system [Litovsky et al., 1999], but also automatic sound analysis methods. The time delay at which the direct sound arrives only

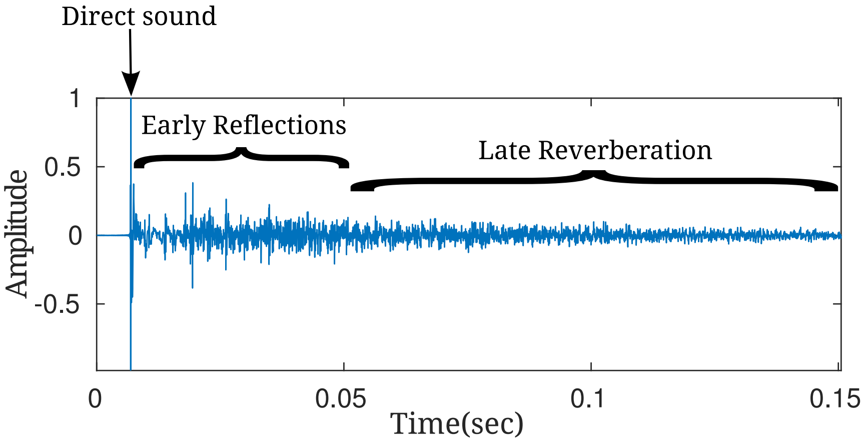


Figure 2.11: An IR measured in a real environment, and the three parts it can be divided: direct sound, early reflections and late reverberation.

depends on the distance between the source and the listener, while the early reflections are additionally affected by the room characteristics. Late reverberation decays exponentially with time, and only depends on the room characteristics - it is not affected by the positions of the sound source and listener. The time required for the late reverberation to decay by 60dB relative to the level of direct sound is called the reverberation time T_{60}, and it is one of the measures commonly used in order to describe a non-anechoic enclosure in terms of reverberation [Kuttruff, 2007]. Another parameter that characterized the reverberation present in a room, and related to the IR is the direct to reverberant ratio (DRR) [Jo and Koyasu, 1975, Kuttruff, 2009], which is defined as the ratio of the sound energy that arrives to the listener via the direct path, *i.e.,* the direct sound, over the energy of the sound that arrives afterwards [Naylor and Gaubitch, 2010].

To offer a better insight for these reverberation parameters, and their relation to different speaker orientations in an enclosure, we create an artificial environment, as the one depicted in Figure 2.12(a). Using the IM we produce several instances of such an enclosure, for different values of reverberation time (T_{60}). The speaker is located at a distance 2m from the microphone and assumes three different orientations. The synthetic IRs are used to estimate the DRR[2]. In Figure 2.12(b) we present the DRR as a function of T_{60} for three different orientations 0^o, 30^o and 135^o. It is observed that DRR directly relates to the orientation of the speaker towards the microphone, with more directive cases, as for example 0^o and 30^o resulting in higher DRRs. Moreover, there is an inverse relation between T_{60} and the corresponding DRR, which is respected in low, average and higher DRR cases, as the different orientations show.

Both parameters, T_{60} and DRR can be directly estimated from the IRS [Schröder,

[2] DRR is estimated from the synthetic IRs with the use of the IR_stats toolbox of MATLAB [Zahorik, 2002]

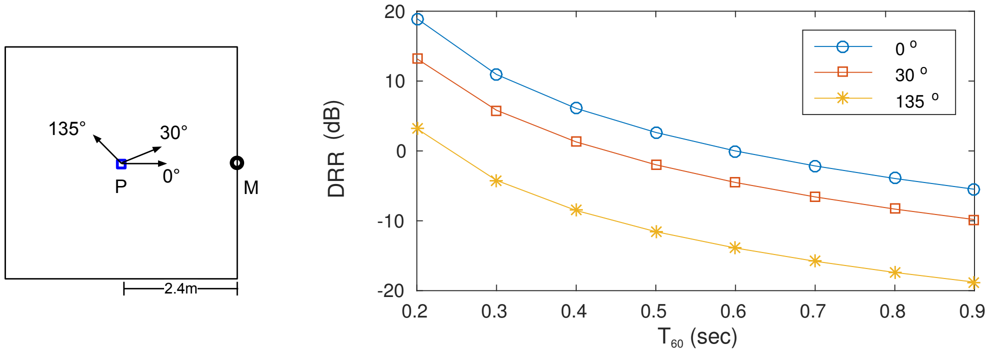


Figure 2.12: (a) The synthetic room used for calculating the relation between DRR and T_{60}. The speaker is positioned in P and assumes 3 different orientations. (b) DRR as a function of T_{60}. T_{60} values range from 0.2sec to 0.9sec which are reasonable for a domestic environment.

1965, Zahorik, 2002] with different approaches, for instance they can be estimated in the full-band or in predefined frequency, since the effect of reverberation is not uniform in the frequency spectrum, *e.g.,* some frequencies are attenuated more than others. Nowadays, different systems are available to derive T_{60} estimates, either in full-band or in sub-bands, starting from a synthetic or from a measured IR. As discussed in [Cabrera et al., 2016], a good agreement between these estimates is generally found, if the IR is characterized by a regular decay curve and low enough noise floor. On the other hand, an accurate estimation of DRR from an IR is a more difficult task, especially in the case of measured IR, due to the uncertainty in deriving the energy of the direct-path [Naylor and Gaubitch, 2010].

Of course, these tasks become much more challenging in real situations in which IRs are not available. Although different approaches have been proposed to estimate the aforementioned parameters blindly [Ratnam et al., 2003], the results are still not satisfactory, particularly for the estimation of DRR. More details on state-of-the-art techniques in this field can be found in [Eaton et al., 2016] that is related to the recent ACE challenge. We believe that improvements in multi-microphone based DSR can also be achieved taking into consideration reverberation parameters as blind DRR.

The effects of reverberation on a speech and a singing voice signal are illustrated in Figure 2.13. Through the spectrogram of the clean and the reverberant version of each signal, the discrepancies between the two become evident. First, we observe a temporal smearing due to reverberation, as for instance the boundaries of phonemes are much more difficult to locate in the second spectrogram. In addition, we can see the harmonic components in the reverberant signal being “extended” and affecting the spectrogram longer than in the clean version.

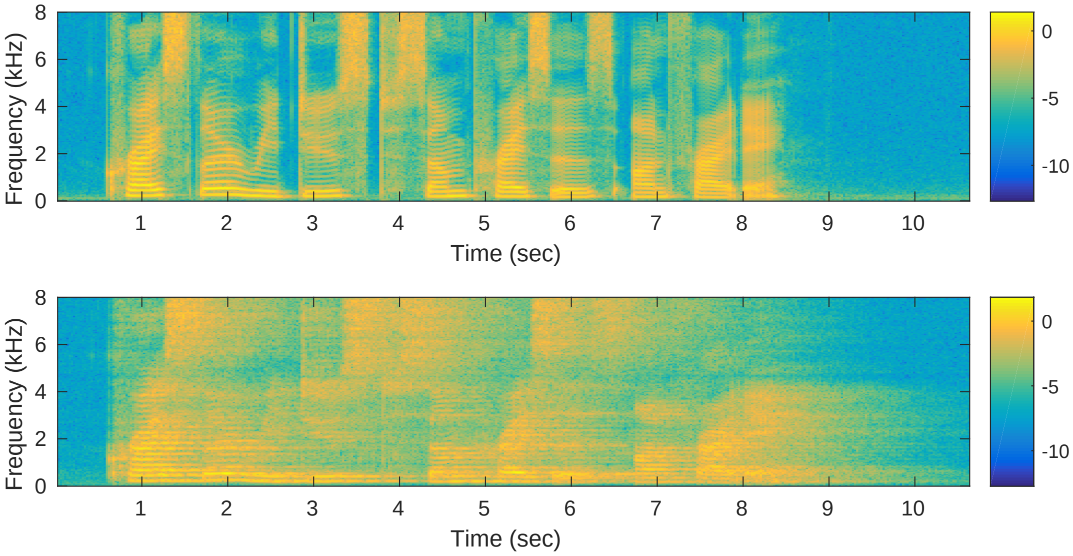


Figure 2.13: The spectrogram of a clean speech utterance (top) and the spectrogram of the same utterance, impinged with reverberation (bottom). The reverberation corresponds to an environment with $T_{60} = 0.9$sec. The uttered sentence is “Chrysler reduced some prices on Friday”.

2.4.2 Music

Music is built from sets of sounds, which are generated concurrently by different sources, and in particular, various musical instruments. The musical instruments, and the produced musical sounds are categorized into *harmonic* and *percussive*. The most fundamental characteristics of a musical sound highly depend on the category it belongs to.

Harmonic sounds are characterized in a great extent by the presence of *pitch*, which is a very important perceptual quality, with several different definitions in the literature. We follow the one found in [Hartmann, 1996, Klapuri and Davy, 2006]: “pitch is defined as the frequency of a sine wave that is matched to the target sound by human listeners”. Pitch is related to the *fundamental frequency* of the target sound but they do not coincide, since pitch is affected by the tones that appear at frequencies approximately equal to the multiple integers of the fundamental frequency, also called harmonics. Another perceived quality of *harmonic* music signals is the *loudness*, according to which sounds can be ordered on a scale from quiet to loud. The physical property of the acoustic signals, that the loudness relates to, is the dynamic range, but psychoacoustics play an important role in its perception [Plack and Carlyon, 1995], as the ear has a non-linear response to sounds of different intensity. Finally, *timbre* is the perceptual attribute that can characterize a sound, and make it distinct from another of the same pitch and loudness [Handel, 1995]. Often referred to as the “colour” of a sound, *timbre* is affected by the energy distribution in the frequency domain, as well as the temporal evolution of this distribution. In practice, *timbre* can be viewed as a multidimensional concept, which, in a music analysis task, should be represented through a vector of values, opposite to pitch and loudness which can be encoded by a single scalar value.

Concerning *percussive* sounds, their main characteristic is a broadband energy envelope. Although percussions are often described as “higher” or “lower” (which is affected by the central frequency of the broadband noise-like disturbance in the spectrum), they lack any harmonic structure. However, percussions greatly affect the *rhythmic* structure of a music piece, *i.e.,* the timing relationships of the various music events. The rhythm of a piece describes all the aspects that relate to the temporal succession and periodical accent of various musical events, as well as the physical duration of pitched sounds.

An automatic system concerned with the analysis and understanding of musical pieces should represent, at least, the above characteristics of harmonic and percussive instruments. In particular, the different components that are necessary for music characterization, for instance in the context of a music transcription system, can be broadly presented in three groups: (a) meter, (b) melody line and (c) bass line. The *meter*, or hierarchical beat structure, represents the fundamental temporal structure of the various music events. The analysis of this structure if an essential part in understanding music signals, and it is an analysis step that even untrained listeners intuitively perform, as indicated by reactions such as “foot-tapping”. The meter is a hierarchical structure of pulses at different levels [Klapuri et al., 2006]. The level “temporal atom” or *tatum* denotes the shorter pulse period that is not incidentally encountered. Most event coincide with the pulses in this level. The next level, called *tactus*, is the most prominent one and it is the one that is commonly called *beat*. The rate of the beats in the tactus level determines the tempo of a piece. Finally, the pulses in the *measure* level define the rhythmic patterns of a piece. The analysis and understanding of the meter of music is fundamental for many applications, and several related tasks are defined within MIREX, as for instance tempo estimation, beat tracking, and onset detection.

The next two components that are necessary for the characterization of a music signal, are mainly related to harmonic instruments. The *melody* and *bass* lines are both temporal trajectories of a series of single tones. One of the differences between the two is the spectral region in which they are normally encountered, as the bass line is located in the lowest parts of the spectrum. In addition, the melody line is often characterized as the predominant line, as it is heard more distinctively than the rest. The detection of both lines is of interest in various applications, and several tasks defined within the MIREX are concerts with related topics, as for example the multiple fundamental frequency estimation and tracking, and audio melody extraction.

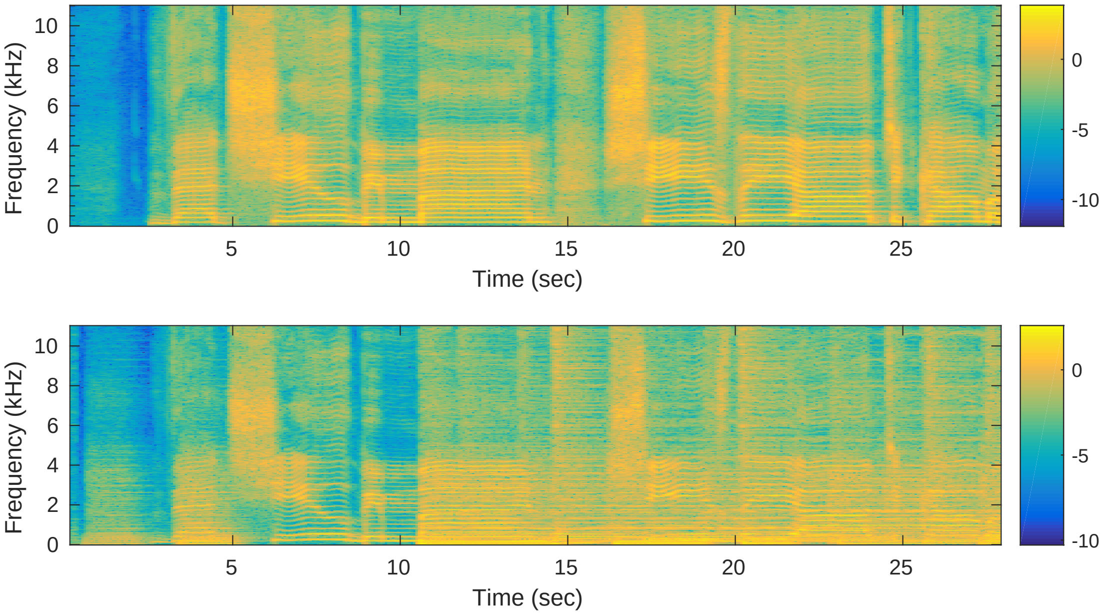


Figure 2.14: The spectrograms of a clean singing voice (top) and the same voice mixed with various instruments (bottom).

In the context of this work, it is interesting to emphasize that when analysing a singing voice signal, music is most commonly included, at a relatively high volume. Because of its characteristics, such as the complex beat and harmonic structures, music complicates several singing voice analysis tasks. An an example, in Figure 2.14 we present the spectrogram of a clean singing voice excerpt, and the spectrogram of the same signal mixed with a variety of background instruments. Notice that in the mixture the harmonic structure of the singing voice is much less evident, particularly after the first 10 seconds when the loudness of the background instruments increase. An automatic method built to detect the f_0 of the singing voice, in the first case only needs to understand the spectral structure and determine the f_0 among a set of harmonics. In the second case, the same method should be able to detect multiple f_0s, coming for the various harmonic instruments, and make a decision on which of the detected f_0s belongs to the singing voice.

The above description covers only a few of the aspects that concern the very rich scientific area which investigates and analyses music signals. The contents, and related concepts are very wide to be treated here comprehensively, but several interesting books and reviews are available, for instance [Casey et al., 2008, Downie, 2008, Klapuri and Davy, 2006, Schedl et al., 2014].

2.5 Conclusions

Acoustic signal processing is a very rich scientific area, concerned with the understanding of diverse acoustic signals. In the core of this field, time-frequency analysis sets the basis for building complex systems, which do not need to assume that acoustic signals are stationary. Time-frequency distributions study time-varying signals by transforming them into two-dimensional representations, which simultaneously describe the temporal evolution of the various frequencies. In a very common scheme for an acoustic signal processing application, this two-dimensional representation will be further exploited into a set of features that feed a decision making unit. The attributes and behaviour of the selected time-frequency representation spread to the extracted features, thus affecting the subsequent decision making and final result of the system. Therefore, the selection of a time-frequency representation, upon which a system is built, should be done with a careful consideration of (a) the characteristics of the particular acoustic signal the system is processing and (b) the exact specifications of the addressed task.

Some of the most interesting characteristics of the acoustic signals this thesis is concerned with have also been discussed in this chapter, making evident certain challenges that related signal processing methods may face. Acoustic signals and in particular speech and singing voice (or similarly other melodic musical instruments) have a very important structure simultaneously on the temporal and spectral domains. The various frequency components, their intensities and their variabilities in time generates meaningful acoustic signals, that carries a big amount of information in a very efficient way. From the moment of the generation of an acoustic signal, the acoustic environment starts degrading its quality.

Chapter 3

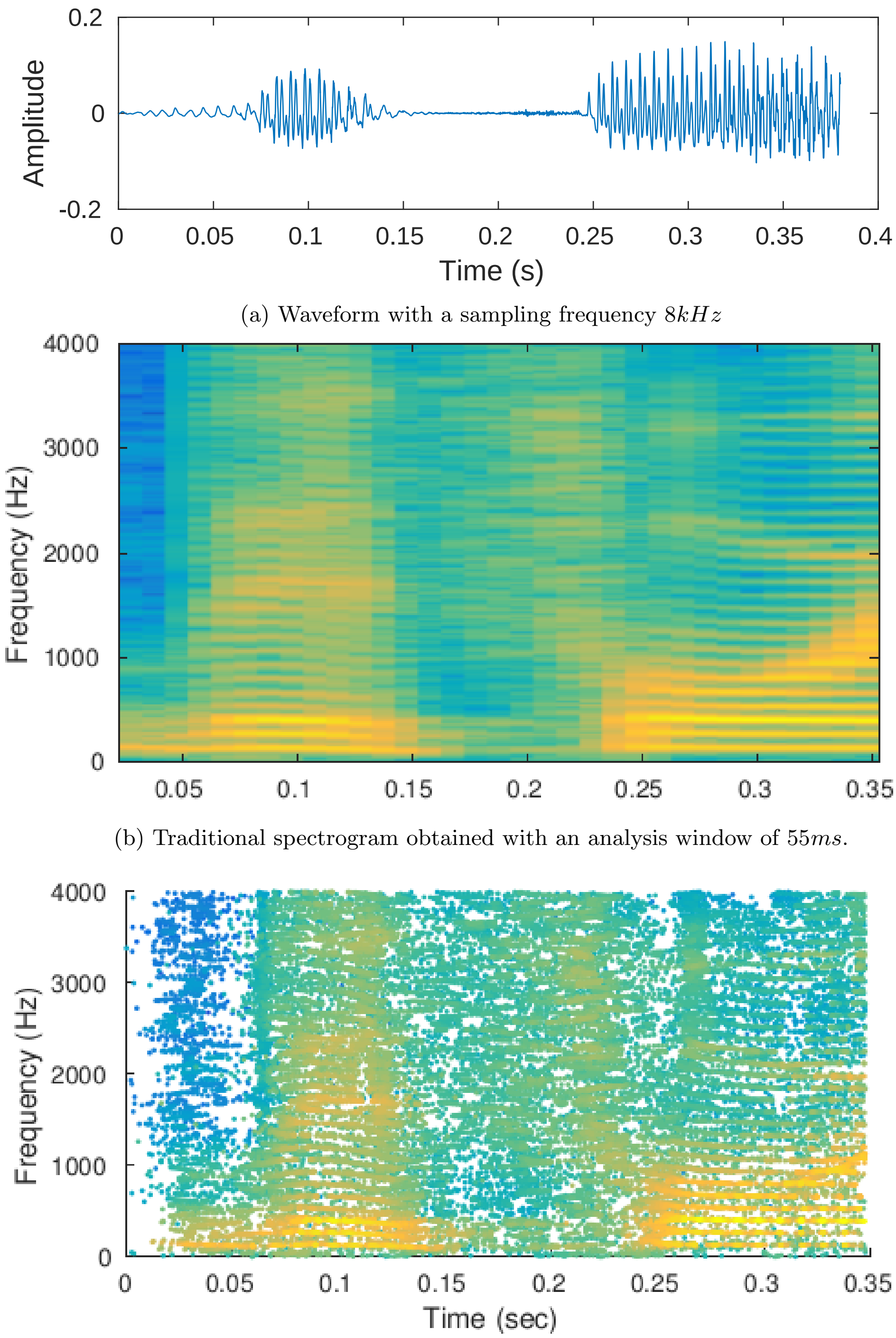
Time-frequency reassignment

In this chapter we set the theoretical basis for the use of the time-frequency RS in signal processing applications. A general introduction and a motivating example for the use of the RS is presented in Section 3.1. In Section 3.2 we present the mathematical formulation which leads to the time-frequency RS. In Section 3.4 we overview various applications that use the RS as the time-frequency representation of the input data. Finally, in Section 3.5 and Section 3.6 we introduce two novel representations exploited in this work, namely the reassigned cepstrum, and the dominance RS.

3.1 Motivating example

The time-frequency RS is a visualization of the instantaneous frequencies of the line components of multicomponent signals. This visualization is conceptually different from the traditional spectrogram, which visualizes the energy distribution in the time-frequency plane. The RS is often described as a *sharpened* version of the spectrogram but this statement can be misleading. Rather than seeing the reassigned spectrogram as an improvement upon the traditional one, it is important to study it as a distinct time-frequency representation of acoustic signals.

As a motivating example, in Figure 3.1 the traditional and the RS of the same speech test utterance are depicted. In the traditional spectrogram, the use of a relatively long analysis window ensures a good visualization of the harmonic components of the vowel regions. Nevertheless, there is a clear smearing of the time boundaries between consecutive phonemes. In the RS the onsets of the phonemes are visualized much better, while the structure of the harmonic content is maintained. In general, the RS yields a time-frequency image that is particularly precise in the representation of components and impulses. This is achieved by discarding information about the bandwidth of each component, information which is anyway already distorted by the short-time Fourier procedure.



(a) Waveform with a sampling frequency $8kHz$

(b) Traditional spectrogram obtained with an analysis window of $55ms$.

(c) Raw RS obtained with an analysis window of $55ms$.

Figure 3.1: The traditional (b) and the reassigned (c) spectrograms of a short utterances spoken by a male speaker.

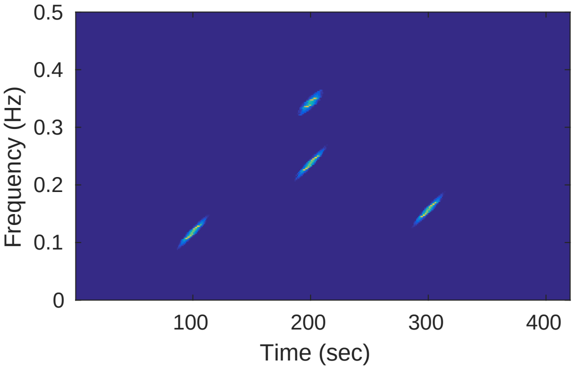


Figure 3.2: The RS of a multi-components signal, where each component has a Gaussian amplitude and a linear frequency modulation. For a comparison of the RS and other time-frequency representations see Figure 2.4.

From this plotting of the raw reassigned data in Figure 3.1, two important characteristics of the RS emerge. First, the RS is only defined at the regions where the analysed signal has a significant amount of energy. Second, the raw data obtained from the process of reassignment can be disappointingly noisy. Additionally, the RS is a positive representation, so each time-frequency point can be interpreted as an energy density. It satisfies the time and frequency shift invariance property [Auger and Flandrin, 1995, Plante et al., 1998]. Similar to the Wigner-Ville distribution, it encompasses a perfect localization of impulses, pure tones and chirp signals, and is not affected by the length of the analysis window as much as the traditional spectrogram. In addition, it does not suffer from the appearance of cross-components. In Figure 3.2 we present the RS of the same signal used in Figure 2.3 and Figure 2.4, where we can clearly observe the perfect localization of each auto-component and the fact that no cross-components appear. It is also interesting to mention that the Rènyi entropy of this RS is 6.31 bits, compared to 11.57 bits for the WVD presented in Figure 2.4 and 10.08 bits for the spectrogram.

Nevertheless, unlike the spectrogram, and more general the Cohen's class of TFDs, the time-frequency reassignment is not a bilinear representation.

3.2 The method of reassignment

The time-frequency reassignment was first introduced in [Kodera et al., 1976] and further discussed in [Kodera et al., 1978], with the name *modified moving window method (MMWM)*. The MMWM was originally described as a means of improving the readability of the spectrogram, specifically by plotting each time-frequency point at the center of gravity of the energy distribution rather than at the center of the analysis window. The proposed analysis pointed out that the spreading of the STFT magnitude can be compensated using phase information, which normally is discarded.

Later, the method of reassignment was redefined and the scope of the technique was extended beyond the case of the spectrogram [Auger and Flandrin, 1995, Flandrin et al., 2002]. The original formulation is very interesting in order to emphasize the particular relation between the time-frequency reassignment operations and the IF and GD of the analysed signal. The generalization however, expresses the time-frequency reassignment as an operation over the WVD and evidences a relation to techniques such as ridge and skeleton [Guillemain and Kronland-Martinet, 1996], synchrosqueezing [Auger et al., 2013], differential spectral analysis and IF density [Friedman, 1985]. We discuss first the generalized formulation of the method of reassignment, and then the original MMWM highlighting the relation to the IF and GD.

3.2.1 Formulation

As discussed in Section 2.1.4, and regarding the spectrogram as a member of the Cohen's class of bilinear TFDs, each spectrogram value is calculated by the summation of a whole distribution of values, and it is assigned to the geometric center of the time-frequency domain. However, this is a rather arbitrary point, which, except from the case of a homogeneous distribution, has no physical interpretation. A better choice is to assign the total energy to the *center of gravity* of the time-frequency distribution. The method of reassignment performs this step. Therefore, at each point (t, ω) of the original spectrogram two additional quantities are computed:

$$\hat{t}_x(t,\omega) = \frac{1}{S_x^h(t,\omega)} \int\int_{-\infty}^{+\infty} \tau W_x(\tau,\nu) W_h(\tau - t, \nu - \omega) \frac{d\tau d\nu}{2\pi} \tag{3.1}$$

$$\hat{\omega}_x(t,\omega) = \frac{1}{S_x^h(t,\omega)} \int\int_{-\infty}^{+\infty} \nu W_x(\tau,\nu) W_h(\tau - t, \nu - \omega) \frac{d\tau d\nu}{2\pi} \quad . \tag{3.2}$$

The point $\left(\hat{t}_x(t,\omega), \hat{\omega}_x(t,\omega)\right)$ defines the *local centroid* of the Wigner-Ville distribution W_x, observed through the window W_h centered in (t, ω). The RS is then defined as

$$\hat{S}_x^h(t,\omega) = \int\int_{-\infty}^{+\infty} S_x^h(\tau,\nu) \delta\left(t - \hat{t}(\tau,\nu), \omega - \hat{\omega}(\tau,\nu)\right) \frac{d\tau dv}{2\pi} \quad . \tag{3.3}$$

Based on the above formulation, the method of reassignment smooths the Wigner-Ville distribution using as a smoothing kernel the Wigner-Ville distribution of the analysis window. The obtained distribution is then refocused to the true regions of support of the components of the signal.

3.2.2 Relation to the instantaneous frequency and group delay

Time-frequency reassignment can be interpreted as estimating the instantaneous frequency and group delay for each time-frequency point. This direct relation of the method of reassignment to the IF and the GD of the analysed signal is better illustrated in the original work by [Kodera et al., 1978], *i.e.,* the MMWM. The MMWM builds upon the classical moving window method [Dziewonski et al., 1969], in which a signal $x(t)$ is decomposed into a set of coefficients defined as follows

$$\begin{aligned} \epsilon(t,\omega) &= \int x(\tau)h(t-\tau)e^{-j\omega[\tau-t]}d\tau && (3.4)\\ &= e^{j\omega t}\int x(\tau)h(t-\tau)e^{-j\omega\tau}d\tau && (3.5)\\ &= e^{j\omega t}X(t,\omega) && (3.6)\\ &= X_t(\omega) \quad . && (3.7) \end{aligned}$$

From (3.6) it can observed that the amplitude of the transform is the same as the amplitude of the STFT, while their phases differ by the linear frequency term:

$$\begin{aligned} M_t(\omega) &= M(t,\omega) && (3.8)\\ \phi_t(\omega) &= \omega t + \phi(t,\omega) && (3.9) \end{aligned}$$

In the moving window method the signal $x(t)$ is reconstructed by the squared magnitude of the above decomposition, which therefore is equivalent to a reconstruction from the STFT. On the other hand, in the MMWM reconstruction the phase information is used as well

$$\begin{aligned} x(t) &= \int\int X_t(\omega)h^*_\omega(\tau-t)d\omega d\tau && (3.10)\\ &= \int\int X_t(\omega)h(\tau-t)e^{-j\omega[\tau-t]}d\omega d\tau && (3.11) \end{aligned}$$

Making use of the observations (3.8) and (3.9), (3.11) is rewritten as

$$x(t) = \int\int M_t(\omega)h(\tau-t)e^{j[\phi_\tau(\omega)-\omega\tau+\omega t]}d\omega d\tau \quad . \tag{3.12}$$

A phenomenon known as *principle of stationary phase* states that only regions of slow phase variation contribute constructively in the above integral. For impulsive signals, *i.e.,* signals concentrated in time, the phase variation with respect to frequency is slow only near the time of the impulse. On the other hand, for periodic or quasi-periodic signals the phase variation with respect to time is slow in the vicinity of the frequency of the periodic oscillation. In general, the phase stationarity condition is satisfied when

$$\frac{\partial}{\partial \omega}\left[\phi_\tau(\omega) - \omega\tau + \omega t\right] = 0 \tag{3.13}$$

$$\frac{\partial}{\partial t}\left[\phi_\tau(\omega) - \omega\tau + \omega t\right] = 0 \quad . \tag{3.14}$$

Therefore the contribution to the integral of (3.12) is maximum around the point with coordinates $(\hat{t}(\tau, \omega), \hat{\omega}(\tau, \omega))$ defined by

$$\hat{t}(\tau, \omega) = -\frac{\partial \phi(\tau, \omega)}{\partial \omega} \tag{3.15}$$

$$\hat{\omega}(\tau, \omega) = \omega + \frac{\partial \phi(\tau, \omega)}{\partial \omega} \quad , \tag{3.16}$$

a point which is considered to be the center of gravity of the distribution. The quantities $\hat{t}(\tau, \omega)$ and $\hat{\omega}(\tau, \omega)$ are related to the GD and the IF respectively. Particularly [Kodera et al., 1978] showed that the $\hat{\omega}(\tau, \omega)$ is exactly equivalent to the IF of the most dominant component at that time and frequency. In addition, $\hat{t}(\tau, \omega)$ is equal to the time at which an impulse that lies within an analysis window $h(t)$ takes place, in relation to the beginning of the window.

3.2.3 Pruning of the reassigned spectrogram

The quantities defined in (3.1) and (3.2), *i.e.,* the time and frequency reassignment operators are only meaningful when the point (t, ω) has significant energy, that is $S_x^h(t, \omega) \gg 0$. When there is no energy at all the whole notion of reassignment is not valid and if the energy is very low reassignment results to be a seemingly random operation. As a result, the RS suffers from speckled, low energy noise [Nelson, 2002]. However, a very simple and intuitive step can help to reduce this noise.

The window used at the spectral analysis step emphasises the signal energy that is near the geometric center of the window. The process of reassignment remaps this energy to the center of support of the analysed signal. A large time or frequency reassignment indicates that the particular analysis window does not represent well the corresponding area of the signal. Therefore, data that produces large time or frequency reassignments can be discarded from the final representation [Fitz and Haken, 2002, Gardner and Magnasco, 2005] assuming that the unreliable data will be better represented in a neighbouring frame of the STFT.

Another interesting aspect demonstrated by [Gardner and Magnasco, 2005] relates to the consensus among neighbouring frequency estimates, *i.e.,* the rate of change of their reassigned frequencies. According to this study, a high degree of consensus, that is a slow change of the reassigned frequencies, indicates the quality of the local frequency estimates. The consensus, defined as $\frac{\partial \hat{\omega}(t,\omega)}{\partial \omega}$ can be rewritten using (3.16) and the above observation is then expressed as

$$\frac{\partial \hat{\omega}(t,\omega)}{\partial \omega} = 1 + \frac{\partial^2 \phi(t,\omega)}{\partial t \partial \omega} \approx 0 \quad . \tag{3.17}$$

In [Nelson, 2001, 2002] is was further demonstrated that a high degree of consensus among neighbouring time estimates characterises impulsive components

$$\frac{\partial \hat{t}(t,\omega)}{\partial t} = -\frac{\partial^2 \phi(t,\omega)}{\partial t \partial \omega} \approx 0 \quad . \tag{3.18}$$

In terms of implementation, (3.17) and (3.18) are rewritten as

$$1 - \frac{\partial^2 \phi(t,\omega)}{\partial t \partial \omega} < A \quad , \tag{3.19}$$

and

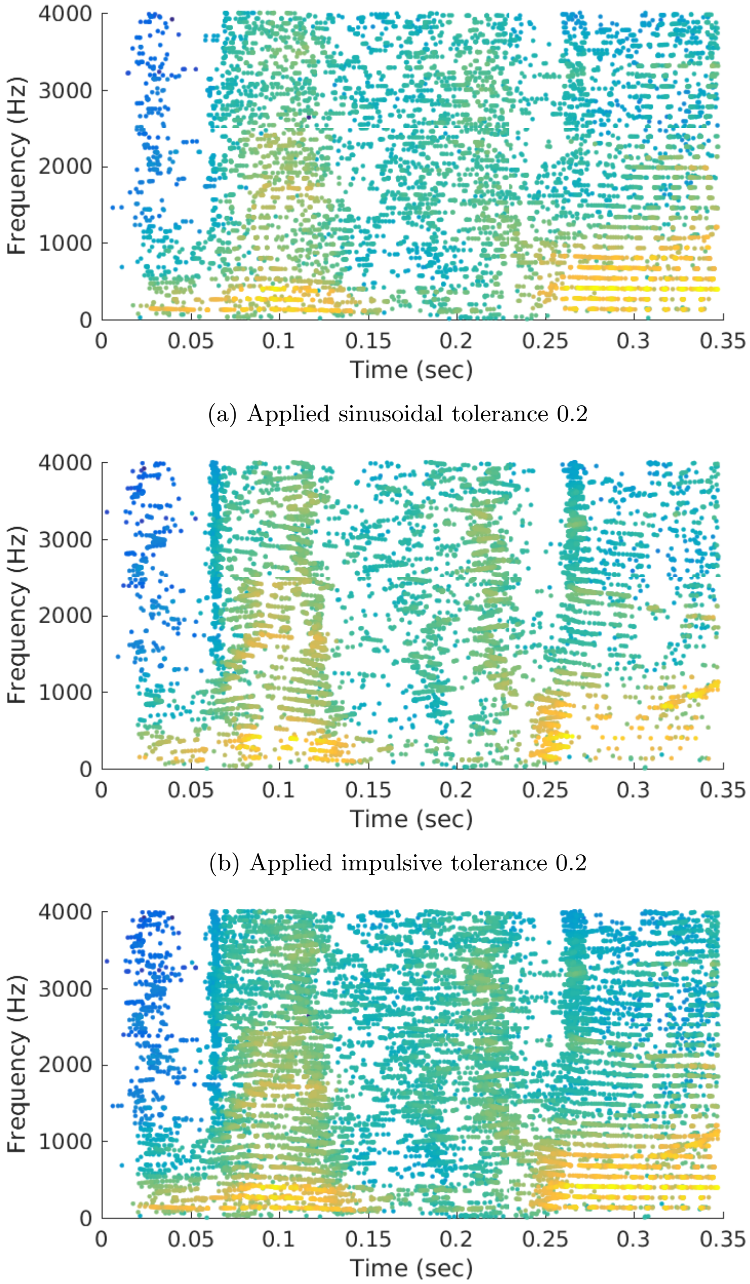
$$\frac{\partial^2 \phi(t,\omega)}{\partial t \partial \omega} < A \quad , \tag{3.20}$$

where A is a tolerance factor, which defines the maximum acceptable deviation of a spectral component from a pure sinusoid, and the maximum acceptable deviation of an impulse from the Dirac function. For speech signals, reasonable values reported for A are in the range $[0.2, 0.4]$. By discarding the points of an RS that do not meet the condition in (3.19), we obtain a visualization of the strongly sinusoidal components, as in Figure 3.3a. For the speech signal used in this example, analysed with a relatively short analysis window, these components are related to the vocal tract resonances. On the other hand, by discarding the points of an RS that do not meet the condition in (3.20) we obtain a visualization of the impulsive components, as in Figure 3.3b, which are related to the individual glottal pulses. In Figure 3.3c both tolerances are applied, which yields a "de-speckled" RS, comprising both sinusoidal and impulsive components.

3.3 Implementation aspects

3.3.1 An efficient implementation

The time and frequency reassignment operators, as defined in (3.15) and (3.16), cannot be directly implemented. This explains the limited exploitation of this representation, until the description of the efficient implementation of [Auger and Flandrin, 1995]. In



(a) Applied sinusoidal tolerance 0.2

(b) Applied impulsive tolerance 0.2

(c) Both tolerances have been applied, leading to a pruned version of the RS.

Figure 3.3: The pruning process based on thresholding the MPD of the phase

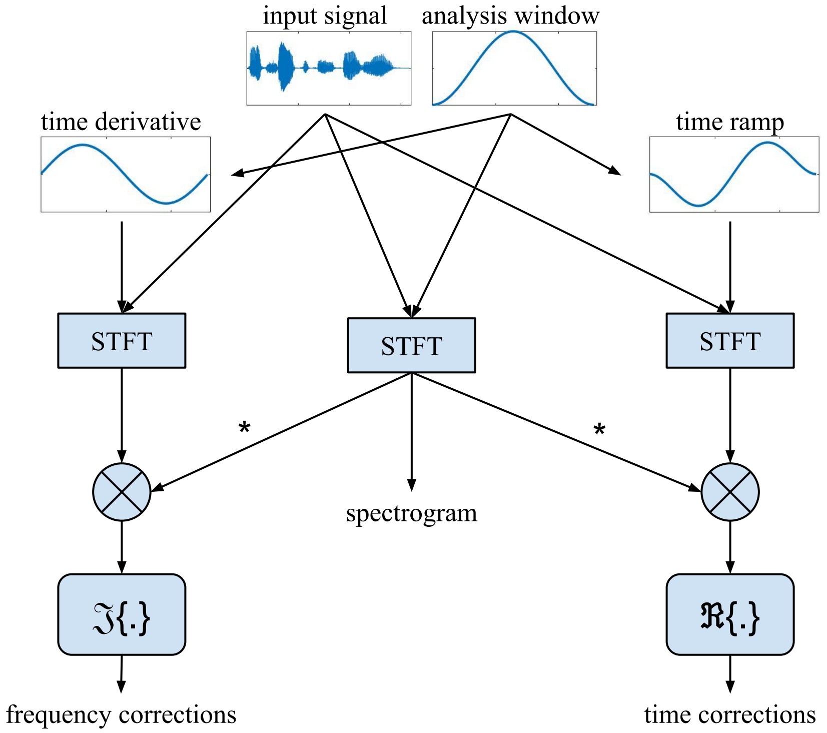


Figure 3.4: An efficient implementation of the time-frequency reassignment operations.

this work, the authors demonstrated that instead of using the derivatives of the STFT phase, the time and frequency reassignment operators can be computed from a set of STFTs, calculated with different analysis windows. Therefore, the time and frequency reassignment vectors are calculated as

$$\hat{t}(t,\omega) = t - \mathfrak{Re}\left\{\frac{X_{Th}(t,\omega)X^*(t,\omega)}{|X(t,\omega)|^2}\right\} \tag{3.21}$$

$$\hat{\omega}(t,\omega) = \omega + \mathfrak{Im}\left\{\frac{X_{Dh}(t,\omega)X^*(t,\omega)}{|X(t,\omega)|^2}\right\} \tag{3.22}$$

where $X(t,\omega)$ is the STFT computed with an analysis window $h(t)$, $X_{Th}(t,\omega)$ is the STFT computed with analysis window $h_t(t) = th(t)$, which is a time weighted version of $h(t)$, and $X_{Dh}(t,\omega)$ the STFT computed with an analysis window $h_D = \frac{d}{dt}h(t)$, which is the time derivative of $h(t)$. The necessary signal processing steps are demonstrated in Figure 3.4, and as shown the time and frequency corrections can be computed as a set of algebraic operations, completely skipping a calculation or an approximation of the phase derivative.

Concerning the calculation of the mixed partial derivatives (MPD) used for separating sinusoids from impulses, in the original work Nelson used finite differences for the computation. Nevertheless, using the above derivations of the reassignment operations it can be shown that the MPD can be computed directly from Fourier transforms by

$$\frac{\vartheta^2 \phi(t,\omega)}{\vartheta t \vartheta \omega} = \Re\left\{\frac{X_{TDh}(t,\omega)X^*(t,\omega)}{|X(t,\omega)|^2}\right\} - \Re\left\{\frac{X_{Th}(t,\omega)X_{Dh}(t,\omega)}{|X(t,\omega)|^2}\right\} \tag{3.23}$$

where $X_{TDh}(t,\omega)$ is the STFT of $x(t)$ computed using a window $h_{TD}(t) = t\frac{d}{dt}h(t)$, which is the window used to compute $X_{Dh}(t,\omega)$ multiplied by a time ramp.

3.3.2 Re-quantization

According to the uncertainty principle, the WVD has an intrinsic quantization grid, upon which the time-frequency points of the distribution can be defined. Following the definition of the RS as a smoothing operation over the WVD, it becomes clear that there is no single quantization grid upon which all the time-frequency reassigned points are defined. This fact comes in contrast with the classical view of a time-frequency representation as a energy distribution over an underlying Wigner-Ville grid. In addition, the raw time-frequency reassigned points cannot be visualized and easily used for further processing unless they are quantized into a well defined grid.

For the above reasons, many algorithms re-quantize the energy from the reassigned points back to the closest STFT grid point [Plante and Ainsworth, 1995]. This type of processing is the re-quantization approach available with the most commonly used toolboxes for computing the RS, as for instance in [Auger et al., Fitz, 2007]. As expected, this operation re-introduces some of the smearing that the reassignment operators have removed. However, it makes plotting functions much easier to implement and faster to execute.

3.4 Applications of the reassigned spectrogram

Time-frequency reassignment has been somewhat ignored in the literature, with only a few applications utilizing it as the time-frequency representation of the input signal. In addition, many initial studies that demonstrated benefits stemming from the use of the RS were not followed up with subsequent activities.

Speech signal analysis and visualization is one of the most important application areas for the RS, as it is very useful in representing simultaneously the temporal, *i.e.,* onsets of plosive sounds, and the spectral features, *i.e.,* harmonic structure of vowels, of speech signals [Fulop, 2011]. In addition, the suitability of the RS in visualizing individual vocal chord pulsations [Fitz and Fulop, 2009, Fulop and Fitz, 2006]. In [Plante and Ainsworth, 1995, Plante et al., 1998] the method of reassignment was applied in the context of speech formant analysis, and the notion of requantizing the RS points at the STFT grid centers was introduced. In [Meyer et al., 1997] the RS was utilized in a "double-vowel" identification task which was showing improvements over the recognition based on the traditional spectrogram. In a slightly different group of applications in the area of speech signal analysis, the RS has been exploited for speaker identification. The concept was first introduced in [Fulop and Disner, 2007] and further discussed in [Fulop and Kim, 2013].

Another field, within which the RS has been somewhat exploited so far, is the area of music analysis. In this context, the RS is again a competitive time-frequency distribution of the input, since it can represent at the same time the rich harmonic structure of melodic instruments and the musically important beat structure of the piece. In [Hainsworth, 2003, Hainsworth et al., 2001] the RS was exploited in transcribing classes of objects, such as sinusoids, transients and noise, within music signals. In [Hainsworth and Wolfe, 2001] a method of piano notes onset detection using time reassignment was presented. In [Khadkevich, 2011, Khadkevich and Omologo, 2013] the RS was used for chord recognition, and beat structure analysis. Finally, music synthesis using a reassigned version of the spectrogram has been proposed in [Fitz and Haken, 2002, Fitz et al., 2000].

Finally, the RS has been applied in other tasks both related to acoustic signal processing, for instance audio coding [Peeters and Rodet, 1999] and sinusoidal modelling [Ito and Yano, 2007], and other not related tasks, such as seismic data analysis [Odegard et al., 1997].

3.5 Reassigned cepstrum

The various characteristics of the cepstrum, and its ability to model the spectral envelope of the vocal tract, make it a particularly appealing representation for input speech. As described, the real cepstrum, commonly used to extract acoustic features, is calculated as the inverse DFT of the logarithm of the DFT of the input signal. Furthermore, it is a common practice to apply further processing steps before the inversion, as for instance the applying filters designed to emphasize certain characteristics of the acoustic signal. The cepstrum can be extended to the case of the RS, as the inverse DFT of the logarithm of the time-frequency reassigned spectrum of the input. The corresponding processing steps are visualized in Figure 3.5. As shown there, after the application of the logarithm the obtained representation, which is defined in the continuous time-frequency domain must be re-quantized before inverted with the inverse Fourier transform.

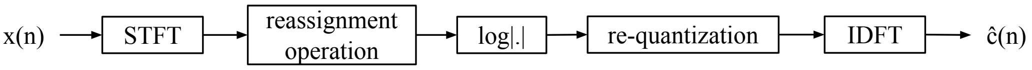


Figure 3.5: The processing steps for obtaining the reassigned cepstrum.

3.6 Dominance reassigned spectrogram

In a multi-component signal, the STFT can be used to estimate the amplitudes and phases of the individual components. However, when different components are located very close in frequency, and their IF is not changing with frequency, only one single component dominates the spectrum [Fulop and Disner, 2007]. In these regions, all nearby spectral data are pulled to the frequency of the dominant sinusoid, a fact that leads to a very low variation of IF over frequency. Therefore, the STFT points that coincide with the IF of a component result in the minimum amount of frequency reassignment in the vicinity of the component. This feature has been exploited in the literature for pitch extraction in what is called *fixed-point analysis* in systems such as YIN and PreFest [De Cheveigné and Kawahara, 2002, Goto, 2005, Kawahara et al., 1999].

The same property emerges in the RS, as shown in Figure 3.6, for a clean singing voice signal. The frequency reassignment shows a minimum value around the f0 and its integer multiples, since these frequencies dominate the spectrum. As described, the way that the spectral energy of each time-frequency bin of the spectrogram is reassigned to a new time-frequency reassigned (TFR) point is governed by the derivatives of the spectral phase at this bin, the same derivatives that theoretically result in the IF and local group delay of the analysed signal. Both of these quantities have been exploited in terms of pitch and melody extraction. For example, in [Rajan and Murthy, 2013] a set of modified group delay functions are used for melody extraction, since the presence of harmonic components corresponds to their local maximization. On the other hand, in fixed-point analysis, the IF is detected and then further used to perform f0 estimation assuming that the points that correspond to a minimum distance between the IF and the spectral bins, *i.e* fixed-points, are indicators of the presence of fundamental frequencies in this spectral region. Furthermore, as shown in Figure 3.6 for the same signal, a higher concentration of TFR points is expected around the same frequencies. The reason behind this is simple. The f0 components dominate the spectrum in terms of energy, but in the STFT calculation this energy is spread in the surrounding bins. The process of reassignment brings the energy back to the region of support of the dominating component, which results in a higher concentration of TFR points around the f0.

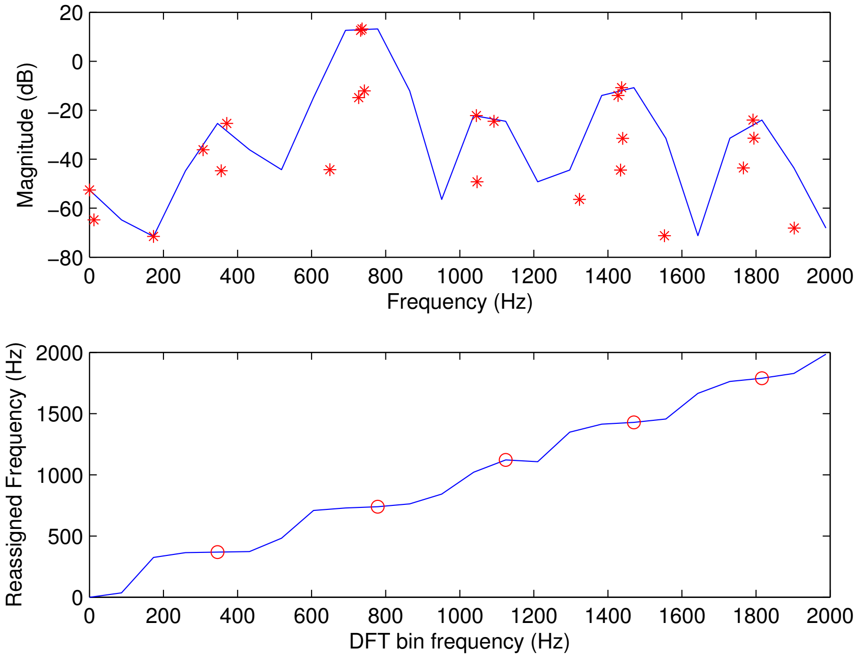


Figure 3.6: Frequency reassignment of a single frame for a fragment of clean singing voice. Top: The continuous line corresponds to the conventional spectrum, while the stars represent the reassigned one. Bottom: The continuous line is the mapping of the Fourier transform bin center frequencies to reassigned frequencies, and the circled red points show the bins that have the minimum frequency reassignment in their vicinity.

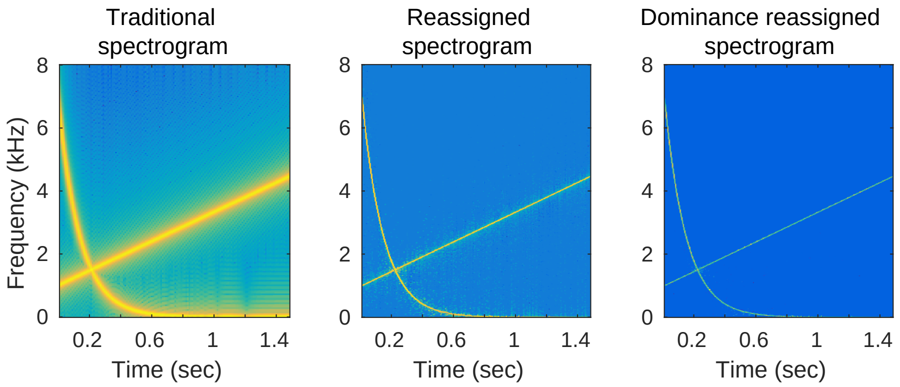


Figure 3.7: Comparison of different versions of the spectrogram of the mixture of a linear and a logarithmic chirp. Notice the improvement of the visualization in the reassigned and dominance reassigned spectrograms.

Summarizing, we can exploit the two interesting properties of the TFR points outlined above, namely that (i) more TFR points are found around the areas of high energy and (ii) the minimum frequency reassignment is observed for the TFR points around predominant components, and define a new representation called DRS. This representation aims at adding further salience in the harmonic components of the analysed signal, while suppressing impulsive and noisy points. The DRS is defined as

$$D(\hat{t},\hat{\omega}) = \left(\frac{X(\hat{t},\hat{\omega})}{\omega - \hat{\omega}}\right)^2 \quad , \tag{3.24}$$

where $X(\hat{t},\hat{\omega})$ is the power RS. The difference $(\omega - \hat{\omega})$ is expected to be minimum in the region around dominant components, leading to a maximization of the DRS. The use of the square alters the dynamics, adding further salience to the most dominant TFR points. Similar to the dominance spectrum that was introduced in [Nakatani and Irino, 2004], the DRS assigns to each TFR point a degree of dominance, which represents its importance in terms of harmonic content. In order to better visualize the effect of the dominance weighting on the RS, Figure 3.7 presents a comparison for a synthetic chirp signal. Apart from an improved visualization of the IF, we observe less random noise and better separability around the region that the two components meet. It is noted here that the Rènyi entropies of these three representations are, 19.72 bits for the spectrogram, 18.91 for the RS, and 17.88 for the DRS.

To better demonstrate the power of the DRS in describing the harmonic content of the predominant source, in Figure 3.8 a single frame of the re-quantized RS and DRS of a polyphonic music signal are depicted. The signal comprises a singing voice (predominant source) and a mixture of piano and bass (background). The two are mixed together in

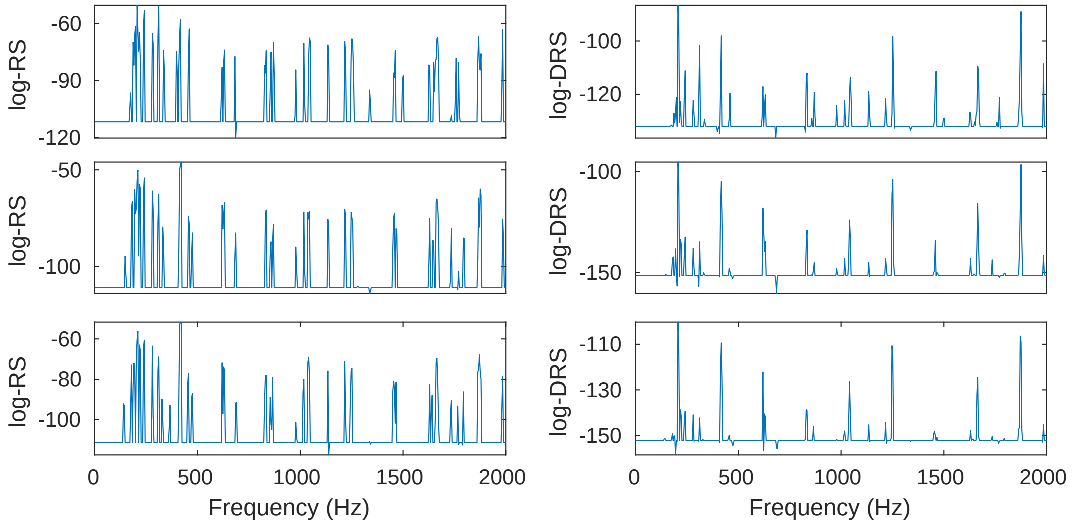


Figure 3.8: The log RS (left) and log DRS (right) of a single frame of a polyphonic music signal, which comprises a singing voice and 2 harmonic musical instruments (piano and bass). Each row corresponds to a different mixing ratio: -4dB (top), 0dB (middle) and 4dB (bottom).

decreasing voice-to-background ratio, shown in the three rows of Figure 3.8. It is observed that in increasing degree of signal complexity, the DRS is more adequate in emphasizing the fundamental frequency of the predominant source, and it is less affected than the RS by the background frequencies.

3.7 Conclusions

The RS is a time-frequency representation that can offer infinite resolution in the time-frequency domain. In practice, the RS remaps the spectral energy of each spectrogram bin, from the point at which it was computed to a new time-frequency point which is closest to the true region of support of the analysed signal. In this chapter we presented in detail the various theoretical aspects that are related to the method of reassignment, and discussed its main limitations, which are, first, the appearance of random-like noise in areas where there is no energy to reassign and, second, the need for a re-quantization step. Following this background information, we proposed two distinct representations stemming from the RS, namely the reassigned cepstrum and the DRS. A very initial investigation, particularly concerning the DRS, shows the potential of this representation in the analysis of acoustic signals. In the next chapters, we investigate in depth how the RS can be exploited in the context of different systems, each time focusing further on one of these two proposed representations.

Chapter 4

Speech segmentation and recognition

In this chapter, we extend on the reassigned cepstrum representation introduced earlier, and we propose the TFRCC. This novel set of speech features offers an improved representation of the speech structure, and is exploited within two applications, namely speech segmentation, and speech recognition. The remainder of this chapter is organized as follows. In Section 4.1 we overview the application areas and outline the main directions of relevant research. In Section 4.2 a detailed description of the TFRCC computation steps is presented. Following that, in Section 4.3 we describe the corpora that we exploit for various experimental activities. In Section 4.4 the proposed features are exploited in the context of speech segmentation and the corresponding experimental activities and results are presented. The use of the TFRCC features for speech recognition is described in Section 4.5, along with experimental activities and results. The chapter is concluded in Section 4.6.

4.1 Related work

The first systematic efforts towards a complete ASR system begun in the 50s, and already in 1952 the first isolated digit recognition system was based on recognizing formant patterns in the power spectrum of the speech signals [Davis et al., 1952]. During the next years, several systems addressed similar tasks [Forgie and Forgie, 1959, Fry, 1959, Olson and Belar, 1956] but did not manage to extend beyond the recognition of a vocabulary of around 10 words, spoken by a single speaker. In the 60s, Japanese laboratories started experimenting with hardware solutions, as for instance the hardware vowel recognizer proposed in [Suzuki and Nakata, 1961]. In the meanwhile, several fundamental ideas in speech recognition, for example feature normalization and dynamic programming were proposed in the same period [Martin et al., 1964, Vintsyuk, 1968].

Continuous speech was targeted for the first time in the late 60s [Reddy, 1966], diminishing the need of a pause after each uttered word, and setting the basis for the recognition of natural speech. In the late 1970s and early 1980s, the field of ASR was undergoing a change in emphasis: from solutions using simple pattern recognition methods to those exploiting complex statistical frameworks. Initially, approaches were based on the temporal alignment of speech patterns, *i.e.,* via dynamic time warping (DTW) and spectral distance measures [Myers et al., 1980, Rabiner and Juang, 1993].

However, a statistical framework, namely the hidden Markov model (HMM), which until recently constituted the standard approach, was developed in 1980s [Jelinek, 1997, Jelinek et al., 1975, Rabiner, 1989]. The underlying assumption of this statistical framework is that a speech signal can be modelled using a Markov state diagram in order to characterize the temporal properties, and a Gaussian mixture model (GMM) in order to characterize the spectral properties of speech. According to this framework, the decoding, *i.e.,* the process of computing the most likely spoken utterance given a speech signal, is based on the Viterbi algorithm, which is a dynamic programming algorithm that searches and finds a optimum solution for a statistical problem. At around the same period, the neural networks [Waibel et al., 1989] started being re-introduced, after their first appearance in the 1950s.

Nowadays, speech recognition is largely based on techniques developed during the last five years, such as training and optimization principles regarding deep neural network (DNN) [Hinton et al., 2012]. DNNs have been shown in multiple occasions to outperform GMM-based ASR solutions, largely due to their ability to model non-linear dynamics through the intricate connections on which they operate [Bellegarda and Monz, 2016, LeCun et al., 2015].

In the following, we review in more detail certain topics that are related with speech recognition, and in particular the application scenarios that we will assume later in this chapter in order to evaluate the use of the proposed TFRCC features.

4.1.1 Speech segmentation

Since the first efforts for ASR, a fundamental task has been the accurate segmentation and labelling of speech into phone units. A database comprising a complete acoustic-phonetic transcription of the speech utterances is useful for many purposes related to ASR, as for example the initialization of speech recognizers and the evaluation of their performance. In addition, speech segmentation is of interest in other fields, for instance to create databases for concatenative text-to-speech systems and tools that support phoneticians in their studies. The most accurate method of creating time-aligned phonetic labels is to employ an expert human annotator. This approach however, is expensive and requires an excessive amount of time, which has been measured as much as 400 times real time [Godfrey et al., 1992], or 30 seconds per phone [Leung and Zue, 1984]. Moreover, the variability in human annotations results into subjective and unreproducible segmentation choices [Cosi et al., 1991]. Therefore, the design and implementation of automatic methods for phone-level segmentation of speech is of great interest.

Many different approaches have been exploited for addressing the task of automatic alignment, with most being based on either HMM or DTW. The latter primarily uses fixed templates, while in general HMM based approaches are characterised by more flexibility and provide superior results [Hosom, 2009]. Therefore, HMM is the dominant technique in automatic segmentation of speech. In such systems, the acoustic signal and the phone transcriptions are used as input to a phone HMM-based forced alignment system. In other words, the Viterbi algorithm is used for a constrained search of the phoneme boundaries inside the utterance, given the corresponding phonetic transcription. A main drawback of the HMM-based forced alignment is that phone boundaries are not represented in the model. Opposite to the manual segmentation, where phonetic boundaries are placed at specific acoustic landmarks [Stevens, 2002], in forced alignment the boundaries are derived from the alignment between phone states and frames. To address this, different directions have been made, for instance boundary correction using sub-band energy changes [Kim and Conkie, 2002], SVM classifiers to group frames into boundary and non-boundary ones [Lo and Wang, 2007], and neural networks to refine boundaries [Toledano, 2000].

The segmentation results are evaluated as the percentage of correctly aligned boundaries, within different thresholds of tolerance. Because in continuous speech boundary positioning is an inherently subjective task, the goal of automatic phone alignment is often described as achieving the agreement between different human annotators. Within a tolerance of 20ms, the automatic methods have reached the 93.49% of inter-annotator agreement that has been reported in [Hosom, 2009] for TIMIT dataset [Garofolo et al., 1993c]. Nevertheless, when a lower tolerance is considered, the performance of automatic methods is still far from the corresponding inter-annotator agreement, that has been reported as high as 63% within 5ms for a dataset of German sentences [Wesenick and Kipp, 1996]. A reason that contributes to the loss of accuracy with lower tolerances is related to the features used in the forced alignment systems. MFCC and PLP, are currently the most popular choice [Brugnara et al., 1993, Yuan et al., 2013]. Both sets of features are obtained from the power spectrum as computed by the windowed speech signal. However, the application of the STFT can be considered as a source of uncertainty as it suffers from the smearing effect discussed earlier and causes an unavoidable trade-off between temporal and spectral resolution.

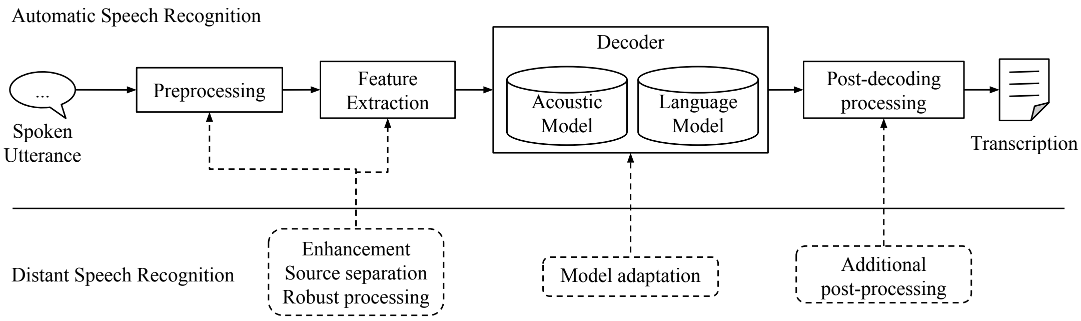


Figure 4.1: Block diagram of a statistical speech recognition system. In the top part, the building blocks of the ASR are shown. In the lower part, the various methods that are applied in order to address a distant talking scenario are also shown.

4.1.2 Automatic speech recognition

As discussed, ASR is based on statistical analysis of speech, performed by complex frameworks, with a general architecture as shown in the top part of Figure 4.1. Before fed to the core of the recognition framework, the acoustic input signal is *preprocessed*, a step that compensates for various variabilities introduced by the acoustic environment [Ephraim and Malah, 1984, Lim and Oppenheim, 1979]. In the next step, namely the *feature extraction*, the acoustic signal is represented in compact form through sets of parameters, as described in Section 2.3

The core of the statistical framework which addresses the recognition of spoken utterances lies in the *decoder*. At this stage, a search of the best match between the sequence of acoustic observations, *i.e.,* sets of features, and a sequence of words takes place. In more detail, the recognition problem can be regarded as computing

$$\arg\max_i P(w_i|\mathbf{O}) \quad , \tag{4.1}$$

where $\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, ...$ is the sequence of speech observations, and w_i the i-th word of the vocabulary. This maximization problem is not solvable directly, but according to Bayes' rule it can be rewritten as

$$P(w_i|\mathbf{O}) = \frac{P(\mathbf{O}|w_i)P(w_i)}{P(\mathbf{O})} \quad . \tag{4.2}$$

Given a set of prior probabilities $P(w_i)$ it becomes evident that the most probable uttered word only depends on the likelihood $P(\mathbf{O}|w_i)$. This term is determined by the *acoustic model*, which based on the assumption that the sequence of observations emitted by the spoken words is generated by a Markov model, is commonly represented by an HMM. The relationship between HMM states and the acoustic input has been represented for years with the power of GMM [Juang et al., 1986]. In practice, the GMMs are used in order to determine how well each state of each HMM fits a frame of acoustic features. Using sufficiently large number of Gaussian distributions, GMMs can model accurately most probability distributions, while they are relatively easy to train, with the expectation-maximization algorithm. Despite all their advantages, and the huge amount of research put into finding ways to improve their accuracy, GMMs are statistically insufficient to model data that lie on or around a non-linear surface, as for instance this of a sphere [Hinton et al., 2012]. An alternative way to evaluate the fit of a window of frames to the state of a HMM, is a feed-forward neural network that takes as an input several frames of coefficients, and has more than one layer of hidden units, *i.e.,* a DNN. In a very compact description, each hidden unit j maps its total input x_j to a state y_j, typically with the use of the sigmoid function

$$y_j = \frac{1}{1 - e^{x_j}} \quad . \tag{4.3}$$

The mapping is performed as

$$x_j = b_j + \sigma_i y_i w_{ij} \quad , \tag{4.4}$$

where b_j is the bias of unit j, i is the index of the units in the layer below, and w_{ij} is the weight of the connection between units j and i. The literature in the topic of DNN for ASR is vast, and detailed reviews of DNN-based ASR can be found in [Hinton et al., 2012, Yu and Deng, 2014].

All the problem formulation so far, concerned the recognition of an isolated word. In continuous speech the prior probability of a word sequence W is expressed as

$$P(W) = \prod_n P(w_i|w_1 w_2 \ldots w_N), \quad n = 1 \ldots N \tag{4.5}$$

where N is the number of words in the sequence W. This probability is given by *language model*, the second statistical part of the recognizer. The objective of this part is to provide information related to the most likely sequence of words to appear in a certain language, and to guide the search among the alternative word hypotheses during recognition.

The parameters of both, the acoustic and the language models are estimated during an initial training phase from a set of spoken utterances, the acoustic features, and their transcriptions [Bishop, 2006, Rabiner, 1989]. When training the acoustic models, the knowledge about acoustics and phonetics is encoded in the acoustic model parameters, based on the relations between the acoustic features and the corresponding phonetic units in the annotated training corpus. In the training phase of the language model, the elements that need to be learnt are the vocabulary and the relations between sequences of n words in this vocabulary.

Finally, the last stage in a recognition system is the *post-decoding processing*, an optional step that aims at the refinement of the decoding output, for instance through confidence measures [Jiang, 2005, Wessel et al., 2001].

4.1.3 Distant speech recognition

Despite the extensive efforts that have been made for reliable ASR, the performance of many voice interaction based systems is still inadequate under certain conditions. Particularly in a *distant talking* scenario, where there is no intrusive body- or head-mounted microphone to record the spoken utterance, challenges are introduced by a variety of reasons. Some examples are the presence of reverberation, obstacles that may exist in the path between the speaker and the microphone, the background noise, and the overlapping speakers. All these factors contribute to acoustic variabilities that degrade the performance of a DSR system [Gong, 1995, Wölfel and McDonough, 2009]. In order to overcome these limitations the numerous strategies that have been adopted are summarized in the bottom part of Figure 4.1, and can be broadly categorized into three groups.

In the first group, different methods attempt to reduce the variabilities introduced by the environment, and enhance the quality of the signal, or feature set, processed by the recognizer [Benesty et al., 2005, Droppo and Acero, 2008, Huang et al., 2008]. Such an improved acoustic signal can be achieved through source separation techniques which aim at suppressing acoustic sources that are overlapping with the target speech signal [Pedersen et al., 2007]. However, source separation can be done only in specific controlled conditions while, on the other hand, speech enhancement techniques can be used for more general purposes. Speech enhancement aims at suppressing, or attenuating environmental noise and improvingthe quality of the acoustic signal. In the same group, various techniques aim at extracting features sets that are robust in the noise and reverberant conditions, making the effect of such degrading factors less relevant in the subsequent recognition step [Hermansky and Morgan, 1994, Kenny, 2012].

In the second group, the processing is focused on the recognition process. A very common approach is the adaptation of the trained models to the noisy, or reverberant models [Droppo and Acero, 2008, Leggetter and Woodland, 1995, Wölfel and McDonough, 2009]. Other prominent examples that target the reduction of the mismatch between the assumed and observed acoustic scenarios material are the multi-condition training, and contaminated speech based training [Matassoni et al., 2002, Ravanelli and Omologo, 2015]. Keeping in mind that the core of ASR system is a statistical pattern recognition problem, this step is critical to significantly improve recognition performance.

The third group of methods uses additional information stemming from the recognition process in order to post-process the output and improve the recognition performance. Prevailing examples here are the methods Recognizer Output Voting Error Reduction (ROVER) [Fiscus, 1997] and Confusion Network Combination (CNC) [Evermann and Woodland, 2000, Mangu, 2000]. The combination of two or more of the aforementioned approaches in a single recognition system is a very common approach.

A very common practice that leads to improvements to DSR solutions, and facilitates many of the above mentioned techniques, is the use of multiple microphones in order to record the same speech signal. This action results in many instances of the same spoken utterances, *i.e.,* redundant information that can be exploited in several ways to improve recognition performance, as for example CS that will be discussed in detail in Section 6. Multi-microphone input also facilitates other techniques such as spatial filtering, and delay and sum beamforming among others.

4.2 Time-frequency reassigned cepstral coefficients

The common goal of the various approaches to the parametrization of speech, is to produce a compact set of values that describe the spectral shape of short segments of speech. Such segments are usually around $25ms$ long and are updated with a rate of around $10ms$. Within each segment, the speech signal is assumed to be stationary, a fact that enables, in core of the most commonly used feature sets, *i.e.,* MFCC and PLP, the use of the STFT for the estimation of the spectral content of the speech. The STFT enables the summarization of the speech content and the periodical update of the extracted parameters. Nevertheless, various alternative time-frequency distributions, that have been studied in the context of speech processing, can be exploited for the parametrization of speech as well.

Among these time-frequency distributions, the time-frequency reassignment is a method that can improve the representation of the speech spectral content, as it represents simultaneously the temporal, *i.e.,* onsets of plosive sounds, and the spectral features, *i.e.,* harmonic structure of vowels, of speech signals [Fulop, 2011]. In addition, when the recognized speech signal is impinged by reverberation, its spectral envelope, and therefore the MFCC features that describe this envelope, are smoothed and carry less information. The RS, obtained from the method of time-frequency reassignment, is a sharpened version of the traditional spectrogram and the reassignment operation mitigates these smoothing disturbances that are introduced by the reverberation.

The proposed TFRCC features are based on the RS and the various stages followed for the extraction of the TFRCC features are depicted in Figure 4.2. The following sections summarize thee computation steps.

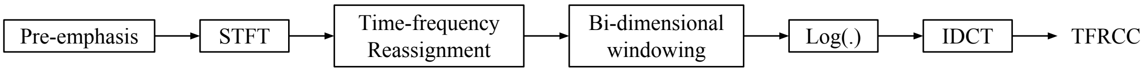


Figure 4.2: Block diagram of the TFRCC extraction steps.

Pre-emphasis The processing begins with the application of a pre-emphasis filter. As already discussed, the goal of this step is to model the input speech according to the human ear sensitivity and to account for the dependency of frequency and perceived loudness.

STFT calculation The discrete STFT is calculated in order to obtain a complex spectrum. In the following, X_h denotes the discrete STFT of a signal, calculated with the use of an analysis window $h(n)$, that is shifted in time with a certain step.

Time-frequency reassignment In the case of the discrete STFT, the reassignment operations in (3.1) and (3.2) cannot be directly computed. Nevertheless, in [Auger and Flandrin, 1995] it is shown that the reassignment operations can be performed with the use of two auxiliary windows, as follows

$$\hat{t} = t - \Re\left\{\frac{X_{\dot{\mathcal{T}}h} X_h^*(t,\omega)}{|X_h|^2}\right\} \tag{4.6}$$

$$\hat{\omega} = \omega + \Im\left\{\frac{X_{\dot{\mathcal{D}}h} X_h^*(t,\omega)}{|X_h|^2}\right\} \quad , \tag{4.7}$$

where $X_{\mathcal{T}h}$ is the discrete STFT computed using an analysis window, which is a time weighted version of $h(n)$, and $X_{\mathcal{D}h}$ is the discrete STFT computed using an analysis window, which is a frequency weighted version of $h(n)$. In practice, (4.6) and (4.7) reallocate spectral energy from the coordinate (t,ω) to the coordinate $(\hat{t},\hat{\omega})$ which can be formulated as

$$X(\hat{t},\hat{\omega}) = |X_h(t,\omega)|^2 \quad , \tag{4.8}$$

with $X(\hat{t},\hat{\omega})$ defined in the continuous time-frequency domain. As a result, the estimates of the spectral energy distribution of the input speech signal are more precise.

Bi-dimensional windowing The representation in (4.8), defined in the continuous time-frequency domain, cannot be directly used in the subsequent processing. In order to obtain a discrete version of $X(\hat{t},\hat{\omega})$ in a new time-frequency domain, a bi-dimensional window is applied. Since $X(\hat{t},\hat{\omega})$ is defined only at the points where there is energy to

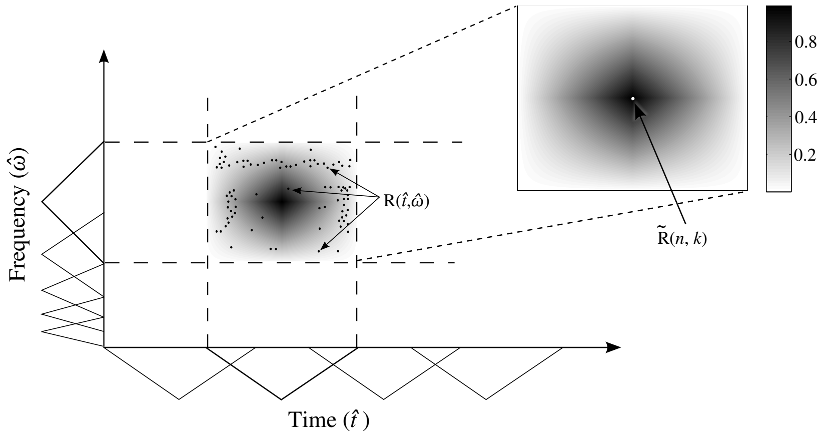


Figure 4.3: The bi-dimensional processing proposed for the re-quantization step required due to the use of the RS.

reassign, this new representation can be expressed as

$$S_w(m,k) = \sum_{(\hat{t},\hat{\omega})} w_k(m-\hat{t},\hat{\omega})X(\hat{t},\hat{\omega}) \quad , \tag{4.9}$$

where $S_w(m,k)$ strongly depends on $w_k(\hat{t},\hat{\omega})$, which is a bi-dimensional window defined in the continuous time-frequency domain, m denotes the generic time instant in the new discrete time domain, and k denotes the index of a frequency range. This processing is shown in Figure 4.3.

Different weighting schemes can be exploited for the design of the window, which becomes more evident when it is expressed as

$$w_k(\hat{t},\hat{\omega}) = l(\hat{t})g_k(\hat{\omega}) \quad . \tag{4.10}$$

In the above notation, $l(\hat{t})$ can be viewed as a continuous time window, that is shifted with a certain step, and $g_k(\hat{\omega})$ as a set of bandpass filters, for example a Mel-scale filter-bank, as the one used in MFCC, but defined in the continuous frequency space. The time resolution of the new time domain is determined by the length and the advance step of the time window $l(\hat{t})$, which should not be confused with the length and the advance step of the window $h(n)$ used for the calculation of the initial STFT. The frequency resolution is determined by the total number of filters in the filter-bank $g_k(\hat{\omega})$.

Compression The discrete $S_w(m,k)$ is logarithmically compressed. The output of this step is essentially equivalent to the log mel-scale filter-bank output of MFCC, but it offers a better localization of the energy distribution of the signal. In Figure 4.4 the log scale output of the Mel filter-bank applied on the traditional, and the reassigned spectra of the

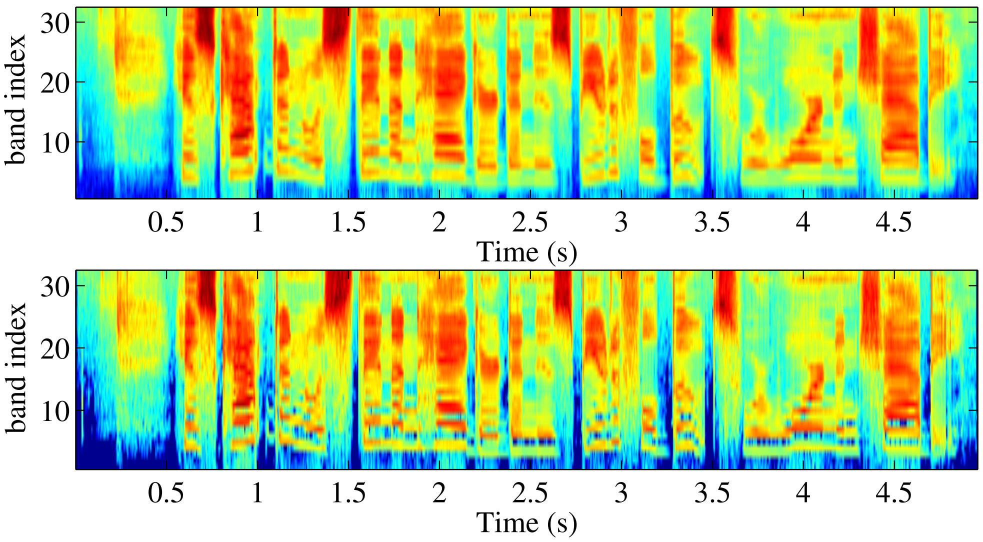


Figure 4.4: The traditional (top) and reassigned (bottom) output of the application of the Mel-scale filterbank, comprising 32 filters, on the spectrum of a speech utterance. Notice the sharper representation of the lower resonant frequencies.

same speech utterance are depicted.

Cepstral mapping The resulting representation is mapped into the cepstrum domain with the application of the IDCT, as typically done with MFCC.

Finally, common techniques, such as the augmentation of the vectors with time derivatives and the normalization of the cepstrum coefficients, can be applied to the TFRCC features.

4.3 Datasets

Here we present the different datasets that have been used for the experimental activities reported in Section 4.4 and Section 4.5. Part of the same data is used in activities detailed in the following chapters as well.

4.3.1 TIMIT

Speech segmentation experiments were performed with the use of TIMIT dataset. The TIMIT Database [Garofolo et al., 1993c] has been created in the late 1980s by Texas Instruments (TI) and Massachusetts Institute of Technology (MIT) The total of 10 utterances, each read by 630 native American English speakers of different dialects and accents, were derived from different corpora as for instance a phonetically rich one (sentences annotated with “sx” in the original TIMIT). TIMIT is annotated with a set of 61 phone units, which is intended to represent an intermediate level between phonetic and acoustic transcription. The transcriptions were produced by expert phoneticians, following waveform and spectrum analysis, along with a predefined set of rules [Garofolo et al., 1993b].

4.3.2 DIRHA-English

On the other hand, ASR, DSR and, as discussed in the next chapter, channel selection (CS) experiments were performed within the DIRHA Project framework [Cristoforetti et al., 2014, Ravanelli et al., 2015][1]. The DIRHA project was focused on the development of voice-enabled automated home environments based on distant-speech interaction, and one of the directions taken was the acquisition of real and simulated acoustic corpora for training, development and test purposes. To this end, some of the datasets within DIRHA are based on simulations realized combining clean speech signals, measured IRs and, optionally, multichannel background noise. Alternatively, some corpora were recorded under real world reverberant, and in cases noisy, conditions. From the whole set of DIRHA-English the following clean speech, IRs and real recordings were used either as provided, or, in cases, as a basis to create new simulations for our experimental purposes.

Clean speech The clean speech data corpus, comprising close-talk recordings, was acquired in a recording studio in the Fondazione Bruno Kessler (FBK), with the use of professional equipment. Native English speakers, both American and British, read material from various sources; in this work we used two distinct sets, as follows

1. Wall Street Journal - **wsj**: The text of this material is taken from the original WSJ0-5k corpus [Garofolo et al., 1993a].

2. Phonetically Rich - **phrich**: The text of this corpus comprises sentences designed to have a large phone coverage and phonetic context, taken from the Harvard corpus[2] .

Simulated data Simulated acoustic corpora was created with the use of the clean speech data described above, and real IRs measured in the acoustic enclosure called DIRHA. The DIRHA room corresponds to the “living-room” of the real environment used within the DIRHA framework. This room is studied here as a simulated realistic scenario created with measured IRs [Cristoforetti et al., 2014, Ravanelli et al., 2012]. The dimensions of the DIRHA room are $4.83m \times 4.51m \times 2.74m$ and the T_{60} has been measured around

[1] http://dirha.fbk.eu.

[2] `www.cs.columbia.edu/~hgs/audio/harvard.html`

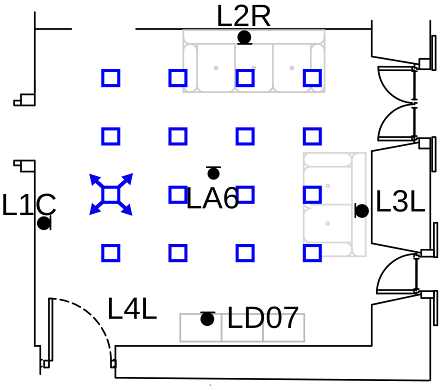


Figure 4.5: DIRHA room setting. Black dots indicate the microphone locations, and blue squares show the various positions of the speaker. The blue arrows are the four orientations which the speaker can adopt.

0.75*sec*, which a quite high value. The average distance between the speaker and the microphones can fluctuate in the range 1-4 meters, a distance that results in a significant degree of distortion, that can greatly degrade the quality of the simulated signal.

Real data Real data corpus comprises real recordings acquired by different subjects, positioned in a particular position in the DIRHA room, and reading utterances from the specified material, *i.e.,* the datasets described for the clean speech case. All the recorded channels were time-aligned at a post-processing step. From the various real data recordings of the DIRHA project, we use those that correspond to the WSJ and phrich datasets, called real-wsj and real-phrich, respectively. In addition, the close-talk signals of the real data were also captured by a head-set worn by the speaker during the recording sessions. An ideal voice activity detection is assumed to be applied over the real data, *i.e.,* ground truth boundaries were used.

Training material Finally, it is noted here that more material was used for the training of the acoustic models. In particular, a subset of the WSJ0-5k training set, comprising 7138 utterances, was used as source material for training. For DSR experiments, this material was contaminated with IRs taken from the DIRHA framework.

4.4 Forced alignment using TFRCC features

As described in Section 4.1.1, in a speech segmentation system when low tolerances are considered for the evaluation of forced-alignment results, the performance is still far from the corresponding inter-annotator agreement. A reason that contributes to the loss of accuracy with lower tolerances is related to the features most commonly used in the forced alignment systems, *i.e.,* the MFCC and PLP sets. As described, both sets of features are obtained from the power spectrum as computed by the windowed speech signal. However, the STFT can be considered as a source of uncertainty as it suffers from a smearing effect and causes an unavoidable trade-off between temporal and spectral resolution. We therefore propose the use of the TFRCC as a set of acoustic features which improve the accuracy of boundary positioning in forced alignment. The reassigned spectrogram provides an estimation of the instantaneous frequency of the input signal and, therefore, a more accurate representation of the time-frequency distribution of the energy.

For the evaluation of the proposed features we performed a set of speech segmentation experiments using forced alignment. The Hidden Markov Model Toolkit (HTK)[Young et al.][3] was used to build phone HMMs, for which the probability estimates of the observations were modelled with GMM. The system was trained on the training partition of the TIMIT database (3696 read sentences, excluding the "sa" files) and tested in the full testing partition (1344 read sentences, excluding "sa" files). The complete set of 61 TIMIT phonemes was mapped into a set of 48 phonemes, as reported in [Brugnara et al., 1993]. The models were trained with the application of the Baum-Welch algorithm, with a total of 6 iterations over the data.

As a baseline configuration we used MFCC features, extracted with the following steps: (i) pre-emphasis of the frames with a pre-emphasis coefficient $\alpha = 0.97$, (ii) application of a 20ms Hamming window, (iii) computation of the power spectrum with an analysis step size of 5ms, (iv) frequency warping with a Mel-scale filter-bank comprising 32 filters, implemented as the default HTK filter-bank, *i.e.,* with logarithmic spacing and constant amplitude, (v) conversion to the logarithmic domain, (vi) application of the IDCT transform to obtain 12 cepstra coefficients and (vii) liftering of the cepstra to obtain a more narrow range of variances. CMN was applied and the log energy was added to the feature vector. TFRCC feature vectors were extracted with the same configuration as above for (i) the calculation of the power spectrum of the acoustic signal, (ii) the pre-emphasis of the signal, and (iii) the application of the IDCT. CMN was applied and the log energy was added to the vector. For the design of the bi-dimensional window in (4.10), the same Mel-scale filter-bank as for MFCC was combined with an overlapping triangular window.

It is noted here that alternative configurations were explored and were found to have a similar effect in both features sets. For example, in Table 4.1 we present the results for four different configurations by combining the application of the pre-emphasis before (configurations 1 and 3) or after (configurations 2 and 4) the framing operation, and

[3] `http://htk.eng.cam.ac.uk/`

		Tolerance			
		5ms	10ms	15ms	20ms
#1	MFCC	45.74	72.12	82.89	87.76
	TFRCC	49.82	73.26	82.86	87.40
#2	MFCC	45.43	72.76	83.65	88.76
	TFRCC	49.99	73.98	83.63	88.25
#3	MFCC	45.11	72.70	83.63	88.64
	TFRCC	49.41	73.55	83.30	87.92
#4	MFCC	45.10	72.66	83.56	88.67
	TFRCC	49.43	73.67	83.38	88.06

Table 4.1: Percentages of correctly positioned boundaries, for different configurations of the MFCC and TFRCC features sets. In sets #1 and #3 the pre-emphasis filter is applied on the whole signal before the framing operation, while for sets #2 and #4 the pre-emphasis filter is applied after the framing. Sets #1 and #2 are created with the traditional triangular Mel-scale filter-bank, while sets #3 and #4 with the filter-bank described in [Davis and Mermelstein, 1980].

the use of the HTK Mel-scale filter-bank described above (configurations 1 and 2) or the original Mel-scale filter-bank used in MFCC features described in [Davis and Mermelstein, 1980] (configurations 3 and 4). In similar experiments it was found that the shape of the time window does not significantly affect the result. On the contrary, changes in the analysis step size result into more notable fluctuations, as presented in Figure 4.6.

Comparative segmentation results are reported in Table 4.2. For these experiments, the bi-dimensional window is created with a triangular time window of length 20ms, advancing in time with a step of 5ms. This, along with the 32-band filter-bank, produces the same time-frequency grid as in the case of the baseline MFCC configuration. The first two rows of Table 4.2 correspond to log-power spectrum domain feature sets, formed by the output of the Mel filter-bank in the case of MFCC features and the bi-dimensional windowing in the case of TFRCC features. The ability of the reassigned spectrogram to offer a more detailed representation of the fine structure of the time-frequency distribution of the acoustic signal is translated into a higher percentage of correctly aligned boundaries, particularly regarding low tolerances.

The next two rows concern the results based on features derived from the application of the IDCT. MFCC-based segmentation presents improved results over all tolerance values. On the other hand, the TFRCC features demonstrate a different behaviour. In fact, a slight decrease of performance within 5ms indicates that the application of the IDCT is not the optimal choice for this step. Nevertheless, the boundary alignment improves when a tolerance higher than 10ms is regarded.

Finally, the last two rows of Table 4.2 are obtained by the extension of the feature

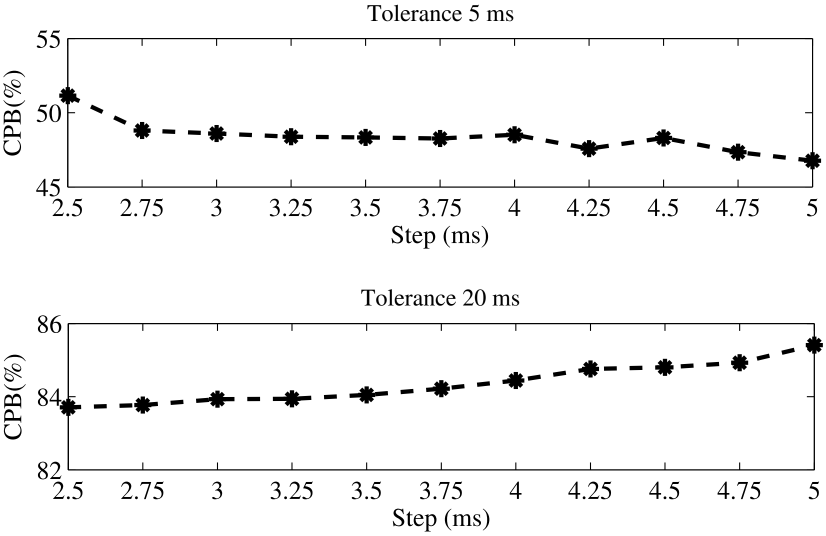


Figure 4.6: Percentage of correctly positioned boundaries (CPB) for increasing advance step of a triangular time window of length 20ms, given a tolerance of 5 and 20ms.

		Tolerance			
		5ms	10ms	15ms	20ms
Spectra	MFCC	36.22	64.63	78.42	84.43
	TFRCC	46.88	69.88	79.10	84.12
Cepstra	MFCC	37.55	65.21	79.12	85.09
	TFRCC	46.74	70.04	80.19	85.40
Δ, $\Delta\Delta$	MFCC	45.74	72.12	82.89	87.76
	TFRCC	49.82	73.26	82.86	87.40

Table 4.2: Percentages of correctly positioned boundaries, for different tolerances, using different feature sets. Notice that these results correspond to the configuration #1 presented in the previous table.

	vowel	stop	nasal	fric	liquid	all
vowel	**15.91**	52.52	**47.38**	40.56	14.02	39.32
	13.97	55.23	46.47	55.07	15.91	43.62
stop	42.82	42.18	29.54	37.06	29.21	39.73
	63.37	53.74	52.95	40.99	64.83	56.23
nasal	31.53	33.95	**20.00**	38.08	**28.41**	32.90
	51.57	34.02	17.50	44.31	27.84	42.49
fric	40.64	**50.12**	36.36	**32.76**	28.10	41.84
	55.37	49.14	53.11	29.80	53.60	52.16
liquid	**17.58**	45.66	**52.23**	36.20	19.08	23.33
	17.25	50.08	47.77	59.38	19.08	25.15
all	32.40	46.16	44.63	38.74	21.64	37.55
	44.23	51.66	47.00	49.07	36.97	46.74

Table 4.3: Percentage of correctly positioned boundaries per phonetic class within a tolerance of 5ms. For each transition pair, the first row corresponds to MFCC and the second to TFRCC. The transitions for which MFCC provide more accurate results (in bold) account for 24.7% of the testing material.

set with the first and second order derivatives, considering a total of 3 and 7 frames, respectively. Focusing on the strictest threshold of tolerance, we observe that in the case of MFCC a relative improvement of 21.81% is presented. The corresponding improvement for TFRCC is 6.52%. This is explained by the fact that the TFRCC features are changing more rapidly than MFCC. Moreover, the use of the same regression formula, which is optimized for the MFCC features, fails to model the dynamic properties of the TFRCC. Nevertheless, the TFRCC features perform better, given a tolerance of 5 and 10ms.

It is also interesting to analyse the results with respect to transitions between different phonetic classes. In Table 4.3, we consider five phonetic classes: vowels, stops, nasals, fricatives and liquids. Both segmentation techniques demonstrate certain limitations in locating the boundaries in transitions such as vowel-to-vowel and liquid-to-vowel. This is expected since no unique point can be defined as boundary in such cases. In fact, such transitions in TIMIT database have been annotated with heuristic rules [Garofolo et al., 1993b] which are not addressed in this experimental set-up. On the other hand, the TFRCC feature set presents an important improvement in better defined cases (36.5% relative improvement for any transition to vowel, 26.6% for any transition to fricative and 70.84% for any transition to liquid).

A final remark concerns the comparison of the results reported above to segmentation results reported in the literature, where results as high as 93.92% within a tolerance of 20ms have been reported in [Yuan et al., 2013] for the TIMIT dataset. The experiments presented in this section were designed to demonstrate the behaviour of the proposed features and compare them with MFCC. All the results can be improved, as in [Yuan et al., 2013], with the use of a more sophisticated HMM architecture, the use of context dependent models and the application of boundaries correction methods, not addressed in this work.

4.5 Speech recognition using TFRCC features

TFRCCs were proved particularly successful in detecting the boundaries between phones, when a very strict evaluation tolerance was considered. This can be attributed to the particularly good temporal resolution that can be achieved with the RS, without sacrificing the spectral resolution. Here, we further investigate the TFRCC features when used as a front-end for speech recognition. We target different ASR scenarios, and in particular recognition of close-talk sentence and recognition of simulated and real reverberant versions of these sentences. These recognition experiments were designed in order to investigate the behaviour of the TFRCC features compared to the MFCC features, under different acoustic conditions.

4.5.1 Recognition framework

All the recognition experiments are performed using the Kaldi speech recognition tool-kit [Povey et al., 2011], and the recipes are based on those described in [Ravanelli et al., 2015], adopted to the particular experimental set-ups addressed here. The decoding is based on the WSJ and phrich corpora from the DIRHA-English set. Concerning the training material it is noted here that decoding of the WSJ corpus is done with acoustic models trained on a subset of the clean WSJ (WSJ0-5k) [Garofolo et al., 1993a] training set, while the decoding of the phrich corpus is done with models trained on the TIMIT training set.

Feature extraction

Each recognition experiment is performed for both sets of acoustic features, *i.e.,* MFCC and TFRCC, extracted from analysis frames of 25ms long, with an overlap of 10ms. Both sets of features are augmented with their first and second order derivatives. The TFRCC feature extraction is implemented, in C++, within the Kaldi speech recognition tool-kit. The processing steps follow closely the MFCC extraction process.

Acoustic modelling

For our experiments, we consider 5 different acoustic models of increasing complexity. In the first level (*mono*), acoustic models represent 48 context independent phones. A three state left-to-right HMM is used to model each of the phones. The *tri1* acoustic models are based on simple triphone training, on features augmented with first and second order derivatives. After that, *tri2* and *tri3* acoustic models are trained on features transformed with linear discriminant analysis (LDA) and maximum likelihood linear regression (MLLR), with *tri3* models trained with speaker adaptive training. All these feature transformation techniques, leading from the baseline *mono* model to the more sophisticated *tri3* configuration, have been shown to be effective for DSR [Tachioka et al., 2013]. Furthermore, DNN running on top of the LDA-MLLR transformed features, were used. The DNNs were built according to Karel's recipe [Veselỳ et al., 2013] with a network architecture shaped by 6 hidden layers of 1024 neurons, with a context window of 11 consecutive frames (5 before and 5 after the analysis frame), and an initial learning rate of 0.008. Of course, the DNN acoustic models required relatively massive computational resources, compared to the previous models.

Language modelling

Concerning the language modelling, for the WSJ dataset we employ the baseline language model used in CHiME-3 [Barker et al., 2015], which is the standard WSJ-5k tri-gram. For the phrich dataset, in order to better focus on the behaviour of the proposed features in encoding acoustic information, we adopt a pure phone-loop as in [Ravanelli et al., 2015]. Although this decision yields a loss in overall recognition performance, we avoid certain non-linear behaviours due to the language modelling.

4.5.2 Close-talk performance

Here, we report the recognition results that were obtained for the close-talk sentences of each dataset, as these were recorded in the FBK recording studio. The recognition results for the clean WSJ test set are presented in Table 4.4. Concerning the acoustic models, as expected the use of more complex models, from *mono* to *DNN* based ones, results in significant improvements on the recognition performance. In addition, we observe the consistent improvements that the TFRCC features yield, compared to the MFCC features, for all the studied acoustic model types. In particular, the relative WER reduction rates, shown in the last row of Table 4.4 indicate that TFRCC features are further supported by the additional feature transformation techniques implemented in *tri2* and *tri3* acoustic models, and the statistical modelling power of DNN models.

The relatively low reduction rate observed in the cases of *mono* and *tri1* models, may be attributed to various non-linear decisions stemming from the use of a tri-gram language model. For this, it is interesting to study ASR results that use a simple phone loop.

Features	mono	tri1	tri2	tri3	dnn
MFCC	22.9	11.1	10.4	6.3	3.7
TFRCC	22.7	11	10	5.8	3.5
	Relative WER reduction (%)				
TFRCC to MFCC	0.8	0.9	3.5	8.6	5.7

Table 4.4: Recognition WER results (%) for the clean WSJ dataset.

Features	mono	tri1	tri2	tri3	dnn
MFCC	47.3	42.8	40.2	32.9	28.1
TFRCC	47.3	41.9	39.2	32.1	27.2
	Relative WER reduction (%)				
TFRCC to MFCC	0	2.1	2.4	2.4	3.3

Table 4.5: Recognition PER results (%) for the clean phrich dataset.

Table 4.5 reports the results for the close-talk recordings of the phrich utterances. First, we observe a general decrease in the recognition performance, which is expected due to the use of a phone-loop as opposed to a language model. Nevertheless, the improvement of the recognition performance with the use of more complex acoustic models is still evident in this experiment. Finally, also for this dataset the TFRCC features result in improved recognition performances.

Pruning the RS As discussed in Section 3.2.3, the RS can be pruned by applying a threshold in the mixed partial derivatives of the phase. This process results in different versions of the same RS which emphasizes either the harmonic or the impulsive components of the input signal. When both thresholds are applied, then the result is an improved, de-speckled representation. In the case of speech, this process has been exploited by phoneticians, for example for the study of various spectral characteristics of different phonetic categories [Fitz and Fulop, 2009].

In the experiments reported here, we explore how pruning affects the TFRCC features, in case emphasizing discrimination cues among different phonetic categories. We derive a set of 36 different configurations of the feature vectors, by using six different values for each of the sinusoidal and impulsive thresholds. For each set of features, a *tri3* acoustic model is trained and tested with matching conditions, *i.e.,* the same feature configuration is used on training and testing material. The results are presented in Figure 4.7, in a 3-dimensional representation, where the color represents the obtained phone error rate (PER). First, we observe that better results are obtained around the diagonal of the visualization, *i.e.,* when the two thresholds are assigned similar values. This is reasonable, as in these feature sets the amount of energy attributed to sinusoid and impulsive components is balanced, as would be expected in any parametrization of speech signal. Similarly, speech recognition

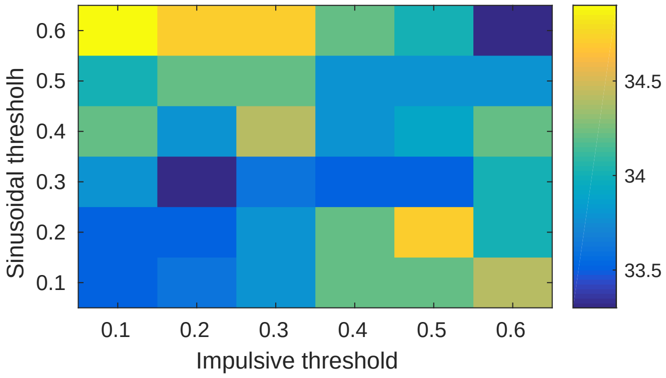


Figure 4.7: Recognition PER results (%) for the clean phrich datasets, and variable tolerances applied on the second order derivatives of the RS used for the extraction of the TFRCC features. The results are obtained from the use of the *tri3* acoustic models.

results seem to deteriorate when a bigger imbalance between the two parts is introduced, as shown by the higher error rates in the diagonally opposite areas of the visualization. Nevertheless, it is relevant to highlight that in any of the presented cases the PER is higher than the corresponding result when no threshold is applied (see Table 4.5). This fact supports the use of no pruning, although this topic deserves further investigation.

4.5.3 Performance under reverberation

Here, we study the performance of the proposed features in reverberant conditions. For the decoding of reverberant speech stemming from the WSJ datasets, the models were trained on contaminated speech utterances stemming from the WSJ0-5k training set. For the decoding of the phrich dataset, the training is based on the contaminated training part of the TIMIT corpus. In [Ravanelli and Omologo, 2014] it was shown that, for the contamination of the training material, the accurate selection of IR is not a critical aspect, and therefore we used IRs referring to the LA6 microphone, installed on the ceiling of the ITEA apartment, as shown in Figure 4.5. The decoding is performed for the full set of five microphones shown in Figure 4.5. At this point, the single distant microphone (SDM) experimental set-up is not used for any further multi-microphone processing, but it provides a better characterization of the effect of reverberation in the specific experimental scenario.

In Table 4.6 the reported results correspond to the simulated WSJ corpus (sim-wsj) and in Table 4.7 to the real WSJ corpus (real-wsj). These corpora are detailed in Section 4.3.2. As expected, the presence of reverberation drastically reduces the recognition performance for both cases. Nevertheless, we still observe that, overall, the use of TFRCC features

Mic	mono	tri1	tri2	tri3	dnn
L1C	65.5	42.3	36	24.8	16.1
L2R	63.9	41.2	35.4	24.4	15.5
L3L	65.2	41.9	35.9	24.8	16.2
L4L	67.5	43.4	37	24.9	16.2
LA6	68.5	44.3	38.9	26.3	17.1
Avg	66.12	42.62	36.64	25.04	16.22

(a) Results using MFCC based front-end

Mic	mono	tri1	tri2	tri3	dnn
L1C	63.9	41.3	34.7	23.9	15.6
L2R	64.2	39.5	35.1	24.2	15
L3L	63.3	40.2	34.4	23.5	15.7
L4L	65.2	41.4	35.4	24.2	15.9
LA6	66.1	42	37.1	25.4	16.5
Avg	64.54	40.88	35.34	24.24	15.74

(b) Results using TFRCC based front-end

Table 4.6: SDM WER results (%) for the recognition of the sim-wsj dataset

Mic	mono	tri1	tri2	tri3	dnn
L1C	66.7	40.9	33.9	23.1	14.5
L2R	68.1	43.1	37	24.1	16.7
L3L	64.5	40.6	33.6	22.8	15.1
L4L	64.4	41.9	34.1	23.3	15.4
LA6	66.2	42.4	35.7	22.9	15.4
Avg	65.98	41.78	34.86	23.24	15.42

(a) Results using MFCC based front-end

Mic	mono	tri1	tri2	tri3	dnn
L1C	65.1	40.5	34.2	22.8	14.9
L2R	67.5	42.9	35.8	24.1	16.5
L3L	64	38.9	33.2	22.3	14.4
L4L	64.2	41	33.4	22.7	14.2
LA6	65.4	40.6	33.6	22.6	14.4
Avg	65.24	40.78	34.04	22.90	14.88

(b) Results using TFRCC based front-end

Table 4.7: SDM WER results (%) for the recognition of the real-wsj dataset

results in improvements of the performance, compared to MFCC features.

In Table 4.8 we present the relative WER reduction rate, when the TFRCC features are used. First, we compare these relative WER reductions to the corresponding results when clean speech is used (see Table 4.4). We notice that, for the *mono* and *tri1* acoustic models the reverberated conditions lead to an increase of the performance gap between the TFRCC and MFCC features. For the remaining acoustic models, the inclusion of reverberation leads to similar relative WER reduction rates, with TFRCC outperforming MFCC features in most cases. We note here that although MFCC result in a maximum

Mic	mono	tri1	tri2	tri3	dnn
L1C	2.24	2.36	3.61	3.63	3.73
L2R	**-0.47**	4.13	0.85	0.82	3.23
L3L	2.91	4.06	4.18	5.24	3.09
L4L	3.41	4.61	4.32	2.81	1.85
LA6	3.5	5.19	4.63	3.42	3.51
Avg	2.39	4.08	3.55	3.19	3.08

(a) sim-wsj

Mic	mono	tri1	tri2	tri3	dnn
L1C	2.4	0.98	**-0.88**	1.30	**-2.76**
L2R	0.88	0.46	3.24	0	1.2
L3L	0.78	4.19	1.19	2.19	4.64
L4L	0.31	2.15	2.05	2.58	7.79
LA6	1.21	4.25	5.88	1.31	6.49
Avg	1.12	2.39	2.35	1.46	3.5

(b) real-wsj

Table 4.8: TFRCC to MFCC relative SDM WER reduction (%) for the two datasets. Numbers in bold highlight the cases that MFCC features result in better recognition accuracy.

Features	mono	tri1	tri2	tri3	dnn
MFCC	69.5	64	62.5	60.9	54.9
TFRCC	69.1	63.6	61.6	57	52.4
	Relative WER reduction (%)				
TFRCC to MFCC	0.5	0.6	1.5	6.4	4.6

Table 4.9: Recognition PER results (%) for the reverberant phrich dataset.

of relative 2.76% better recognition rate in a single case, the TFRCC features consistently result in better WERs, even up to 7.79% relative.

In the last set of experiments we study the effect of reverberation in the recognition results of the phrich dataset. In Table 4.9 we present the PER of the recognition output for this dataset. First, similar to the recognition of reverberant WSJ data, we observe a significant reduction in the recognition accuracy, compared to the corresponding close-talk experiment. Moreover, as already noted, the lack of a language model yields a further increase in the average recognition error rate. Nevertheless, note that also in this complex case, the TFRCC features still outperform the MFCC features, for any type of acoustic model.

4.6 Conclusions

In this chapter, we presented a new set of features that can be used as a front-end for phone segmentation, as well as for speech recognition and other similar tasks. The proposed features result from the time-frequency RS of the speech signal. In the experimental activities in the area of speech segmentation, the TFRCC features were shown to perform equally well with the traditional MFCC features, as far as more relaxed tolerance thresholds are concerned. On the other hand, they outperform MFCC features, with strict thresholds of tolerance. The power of the proposed feature set lies in the ability of the method of reassignment to offer a much sharper representation of the energy distribution of the speech signal. The experiments also indicated that further improvements are possible in the proposed analysis: in fact, both the application of the IDCT and the extension of the features with time derivatives do not yield an improvement as high as expected based on the behaviour of the forced alignment with MFCC features.

In addition, we presented a set of experimental results for the recognition of speech signals represented with the TFRCC features. We found that these features consistently lead to improvements, compared to the use of the MFCC features. These results were confirmed with the use of clean, and reverberated material taken from two different corpora. In addition, we studied the effect of using different acoustic models, from simple monophone based models to state-of-the-art DNNs, as well as the effect of using a triphone language model as opposed to a simple phone loop. In the case of reverberated data, for one of the used corpora we presented results for 5 microphones installed in the experimental set-up, setting the basis for a multi-microphone technique to improve DSR recognition accuracy, which will be introduced in Section 6 .

Chapter 5

Objective quality measures

In this chapter we discuss the use of the RS as the time-frequency representation upon which objective signal quality measures are computed. As discussed in more detail in the following, such measures can be exploited in a wide range of speech analysis applications. We are particularly interested in how objective quality measures can be exploited in order to characterize reverberation, and subsequently to improve DSR recognition performance, as discussed in Chapter 6. In order to move to this direction, we study the relation of speech quality measures with specific parameters that describe a reverberant scenario. In Section 5.1 we provide some background in the area of objective signal quality measures, and present some of the most widely exploited ones. In Section 5.2 we revisit these measures, using the RS as the time-frequency representation upon which they are built, and propose a new subjective quality measure based on the RS. In Section 5.3 we study objective quality measures, with an emphasis on their adequacy in characterizing reverberant conditions[1]

5.1 Related work

The evaluation of the quality of speech signals is of interest in a wide range of applications. Traditionally, the first attempts to evaluate speech quality targeted the evaluation of the distortions introduced by speech codecs or communication channels [Furui, 1991, Malfait et al., 2006, Rix et al., 2001, Wang et al., 1992]. Later, speech quality measures, also known as distance or distortion measures, have been used for the evaluation of various automatic systems that aim at the improvement of speech quality. In particular, systems that perform speech enhancement, noise reduction, and dereverberation are known to introduce certain types of distortions [Hu and Loizou, 2008], which can degrade the quality of the

[1] The analysis presented in this section represents an extension of the collaboration with Cristina Guerrero.

input speech. Objective quality measures are commonly exploited for the assessment of the introduced distortions.

Before going to further details on how speech quality can be evaluated, the notion *speech quality* itself should be discussed. In [Loizou, 2013], quality if defined as "*one of the many attributes of speech signals, [...] which is highly subjective, and difficult to evaluate reliably*". Although not equivalent, the term quality is often used as a synonym to the *intelligibility* of a spoken utterance. The most accurate and reliable method for evaluating speech quality is through subjective listening tests, where human listeners judge the overall quality of a given speech utterance. However, in order to obtain repeatable results, subjective evaluation has to follow very strict rules [itu, 2003, Hu and Loizou, 2007, Quackenbush et al., 1988] and it is a costly and time consuming process. For these reasons, objective speech quality measures have been devised and exploited for many years in speech processing applications. Objective speech quality measures aim at assessing the quality, or intelligibility, of a distorted speech signal, and they are expected to incorporate different sources of knowledge, as for example psychoacoustics, and phonetics. In the ideal case, an objective speech quality measure should accurately match the results obtained with subjective listening tests.

5.1.1 Mathematical considerations

Consider a vector space χ, and two vectors $\mathbf{x}$ and $\mathbf{y}$ defined in this space. A metric d, also called distance function or simply distance, is defined as

$$d : \chi \times \chi \rightarrow [0, \infty] \quad , \tag{5.1}$$

and satisfies the following properties

1. $0 \leq d(\mathbf{x}, \mathbf{y}) < \infty \quad \forall \mathbf{x}, \mathbf{y} \in \chi$ (non negativity),
2. $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (identity of indiscernibles),
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry),
4. $d(\mathbf{x} + \mathbf{z}, \mathbf{y} + \mathbf{z}) = d(\mathbf{x}, \mathbf{y})$ (invariance), and
5. $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{x}, \mathbf{z})$ (triangle inequality).

The first three properties ensure that the given metric is easy to manage in a mathematical way, and are customary in the definition and study of distances. According to these properties, the distances take positive values with a distance equal to 0 signifying that a vector is compared to itself. Furthermore, the distance from $\mathbf{x}$ to $\mathbf{y}$ is equal to the distance from $\mathbf{y}$ to $\mathbf{x}$. In practice, this ensures that when comparing two sounds, the resulted distance is not affected by which of the two is used as a reference. The fourth property ensures that the addition of some distortion to both sounds will not change their overall distance. The final requirement, *i.e.,* the triangle inequality, is necessary for the distance measure d to be called a metric. It is interesting to note here that, in speech processing applications, there are situations that the above properties may not be necessary or even desirable [Flanagan et al., 2008, Gray and Markel, 1976, Loizou, 2013]. However, in general, distances that do respect the above attributes lead to results that are easier to interpret in an intuitive way. As far as objective speech quality measures is concerned, the condition of symmetry and the triangle inequality are relaxed, and the term distance is used more in analogy to the term dissimilarity rather than its strict mathematical definition.

Focusing further on distance measures for speech processing, when $\mathbf{x}$ and $\mathbf{y}$ are two speech vectors and $d(\mathbf{x}, \mathbf{y})$ the distance among these, two additional properties that have been described in [Gray and Markel, 1976, Rabiner and Juang, 1993] are the following.

1. It should be possible to efficiently evaluate $d(\mathbf{x}, \mathbf{y})$, and
2. $d(\mathbf{x}, \mathbf{y})$ should be physically meaningful and have a valid interpretation in the frequency domain

The first of these criteria, concerns the amount of calculation for its evaluation, an aspect that is not so relevant any more. The second, however, is a practical one and it can ensure that a distance measure is usable in real applications. In addition, it suggests that a particular value of the measure should be correlated with the subjective distance judgement, as measured through listening tests.

Objective speech quality measures can be categorized according to the domain on which they operate and have a particular interpretation. The characteristics and some well known examples of the common groups are discussed in the following.

5.1.2 Time domain measures

Historically, the first distance measures were described in the time domain, and were exploited by coders that aimed at reproducing the waveform of the input signal. The most typical time domain measures are the signal to noise ratio (SNR) and segmental SNR (SNRseg). The correlation of SNR with subjective quality has been measured quite poor, making it of little interest as a general objective measure of speech quality [Hansen and Pellom, 1998]. Instead, SNRseg is calculated in a segmental manner [Tribolet et al., 1978], but it can be affected by extended silent regions, as the reduced signal energy in these regions will bias the whole calculation. In addition, it is not possible to incorporate perceptual information in the calculation of SNRseg, as for instance could be done by weighting more the distortions that appear in frequencies that are perceptually more relevant.

5.1.3 Spectral domain measures

Due to the various problems related to the time domain measures the first spectral domain alternatives were introduced. These are less sensitive in silent regions, time misalignments and delays [Quackenbush et al., 1988], while from their definition they have a particular spectral interpretation. A common spectral alternative for a time domain measure is the frequency-weighted SNRseg, defined as

$$d_{fwSNRseg}(S,\tilde{S}) = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} w_j \log_{10}\left[S^2(j,m)/\left(S(j,m)-\tilde{S}(j,m)\right)^2\right]}{\sum_{j=1}^{K} w_j} \tag{5.2}$$

where K is the number of frequency bands, M is the total number of segments (frames), m is the current segment index, w_j is the weight assigned to the j-th frequency band, $S(j,m)$ is the clean signal spectrum and $\tilde{S}(j,m)$ is the evaluated signal spectrum.

Another natural choice for distortion measures between S and $\tilde{S}$, is the set of L_p norms defined as

$$d(S,\tilde{S})^p = \int_{-\pi}^{\pi} |V(\omega)|^p \frac{d\omega}{2\pi} \quad , \tag{5.3}$$

where $V(\omega)$ is the difference on the two spectra in the log spectrum domain

$$V(\omega) = \log S(\omega) - \log \tilde{S}(\omega) \quad . \tag{5.4}$$

From (5.3), and for $p = 1$ we obtain the log spectral distance and for $p = 2$ the root mean square (RMS) log spectral measure. The L_p measures are linear, satisfy the symmetry and positive definiteness properties, and have a strong mathematical basis. In addition, they can be easily related to decibel variations in the log spectral domain. However, L_p norms can be quite irregular, as shown in [Rabiner and Juang, 1993]. Furthermore, they are particularly affected by the contents of the evaluated sound signal, and when versions of the same signal are compared the values are considerably smaller, than when comparing versions of different sounds. Finally, one of the main problems of the L_p measures is their computational efficiency.

5.1.4 Cepstral domain measures - cepstral distance

As demonstrated in [Gray and Markel, 1976], the CD measure is an efficient method for the computation of the RMS log spectral measure, when the cepstral coefficients c_k are computed recursively from the LP coefficients a_k, as shown in (2.31). With the application of the Parseval's theorem on the L_2 measure, we obtain

$$d_2^2 = \sum_{-\infty}^{\infty}(c_k - \tilde{c}_k)^2 = (c_0 - \tilde{c}_0)^2 + 2\sum_{1}^{\infty}(c_k - \tilde{c}_k)^2 \quad , \tag{5.5}$$

where c_k and $\tilde{c}_k$ are the cepstral coefficients of the two evaluated signals. The infinite number of terms makes this definition of no practical use. For this reason, the d_2^2 measure is normally truncated in a smaller set of L terms, which yields the CD measure

$$d_L^2 = (c_0 - \tilde{c}_0)^2 + 2\sum_{1}^{L}(c_k - \tilde{c}_k)^2 \quad . \tag{5.6}$$

The practice of keeping only a small set of cepstral coefficients, leads to a measure which is closely related to the rms distance between smoothed versions of the cepstra.

5.1.5 Likelihood ratios and perceptual measures

Several other distortion measures are built upon the log spectral difference $V(\omega)$. The Ikatura-Saito (IS) [Itakura and Saito, 1970] measure is defined as

$$\begin{aligned} d_{IS}(X, \tilde{X}) &= \int_{-\pi}^{\pi}[e^{V(\omega)} - V(\omega) - 1]\frac{d\omega}{2\pi} && (5.7)\\ &= \int_{-\pi}^{\pi}\frac{X(\omega)}{\tilde{X}(\omega)}\frac{d\omega}{2\pi} - \log\frac{\sigma^2}{\tilde{\sigma}^2} - 1 \quad , && (5.8)\end{aligned}$$

where σ^2 and $\tilde{\sigma}^2$ are the one-step predictor errors of $X(\omega)$ and $\tilde{X}(\omega)$, respectively which are defined as

$$s^2 = \exp\int_{-\pi}^{\pi}X(\omega)\frac{d\omega}{2\pi} \quad . \tag{5.9}$$

As discussed in [Buzo et al., 1980], the IS measure satisfies a form of the triangle inequality, it is subjectively meaningful and results from the standard LP formulas. In particular the IS measure can be expressed as

$$d_{IS}(\mathbf{a}_x, \mathbf{a}_{\tilde{x}}) = \frac{\sigma^2}{\tilde{\sigma}^2}\frac{\mathbf{a}\mathbf{R}_{\tilde{x}}\mathbf{a}_x^T}{\mathbf{a}_{\tilde{x}}\mathbf{R}_{\tilde{x}}\mathbf{a}_{\tilde{x}}^T} + \log\frac{\mathbf{a}\mathbf{R}_{\tilde{x}}\mathbf{a}_x^T}{\mathbf{a}_{\tilde{x}}\mathbf{R}_{\tilde{x}}\mathbf{a}_{\tilde{x}}^T} - 1 \quad , \tag{5.10}$$

where $\mathbf{a}_x$ and $\mathbf{a}_{\tilde{x}}$ are vectors created with the LP coefficients of the clean and distorted speech signals respectively, and $R_{\tilde{x}}$ is the autocorrelation matrix of the distorted speech signal.

However, as in general the ratio measures, the IS does not respect the symmetry property. In fact, for large distortions the asymmetry is substantial [Rabiner and Juang, 1993]. The log likelihood ratio (LLR), also referred to as the Itakura distance, is defined in a similar fashion as the IS measure, but does not incorporate the estimation of the gain using the variance of the predictor error. The LLR measure is found as follows

$$d_{LLR}(\mathbf{a}_x, \mathbf{a}_m) = \log \frac{\mathbf{a}\mathbf{R}_{\tilde{x}}\mathbf{a}_x^T}{\mathbf{a}_{\tilde{x}}\mathbf{R}_{\tilde{x}}\mathbf{a}_{\tilde{x}}^T} \quad , \tag{5.11}$$

In practice, the LLR represents the ratio of energies in the LP residuals of the clean and distorted signals. In [Hu and Loizou, 2008] LLR was found to correlate well with subjective evaluation of signal quality.

A last set of objective quality measures have been inspired by psychoacoustic studies, and the understanding of human auditory system. The Perceptual Evaluation of Speech Quality (PESQ) measure, is a very computationally complex measure, which is recommended by the ITU-T for narrow-band speech quality assessment [Recommendation, 2001, Rix et al., 2001]. In order to compute the PESQ score, the signals are first modelled according to a standard telephone headset. Then they are time aligned, equalized and transformed with an auditory filter. Two distortion parameters, one symmetric disturbance d_{sym} and one asymmetric disturbance d_{asym}, are extracted from the difference of the transforms of the clean and distorted signals. The PESQ score is then computed as a linear combination of the average disturbance values. High correlations with subjective listening tests were reported by [Hu and Loizou, 2008, Rix et al., 2001].

5.2 Reassigned objective quality measures

Despite the long literature in the field of objective quality estimation, and the extended efforts to achieve results similar to those obtained from listening tests, there are still open problems in this area. For example, as discussed in [Hu and Loizou, 2008], although objective measures predict well overall quality, their performance is not so satisfactory in predicting background distortion. Motivated by the improved time-frequency representation offered by the RS, and the additional approaches that can be exploited given the continuous nature of the RS data, we study here objective quality measures defined on time-frequency reassigned spectral data. The distance measures defined earlier, *i.e.,* fwSNRseg, LLR, and CD, can all be redefined in the reassigned time-frequency domain by replacing the traditional short-time spectrum with the time-frequency reassigned one. This results in measures such as the RCD, the reassigned fwSNRseg, and the reassigned LLR which all share the same physical interpretations as their original counterparts. Before using the RS for the calculation of such measures the representation has to be re-quantized, which, as already discussed in Section 3.3.2, leads to information loss and smearing of the time-frequency distribution. Nevertheless, this re-quantization results in a positive two dimensional representation, very similar to the spectrogram, which can ensure that each of these distances will follow the same set of mathematical properties as the original one.

Reassigned point wise distance (RPWD) Here, we attempt to avoid altogether the problematic re-quantization step. To achieve this, we devise a distance measure which operates directly on the continuous time-frequency data and compares the geometries of specific time-frequency regions. In practice, this measure is the average Euclidean distance between a set of time-frequency reassigned points in the clean and the distorted signals, as follows

$$d(\mathbf{c}_{\hat{x}}, \mathbf{c}_{\hat{m}}) = \frac{1}{N} \sum_{\forall(t,\omega)\exists T\Omega} \sqrt{(\hat{t}_c - \hat{t}_m)^2 + (\hat{\omega}_c - \hat{\omega}_m)^2} \quad . \tag{5.12}$$

where $T\Omega$ are time-frequency regions. Although this approach is not affected by the reintroduction of smearing due to re-quantization, noise can be a problem. As discussed in Section 3.2, the RS suffers from random like noise which results, mainly, from forcing the reassignment operation in regions where there is no energy. Particularly in the higher frequencies, where the STFT has in general less energy the noise like time-frequency reassigned data can drastically affect this type of reassigned euclidean distance. For this, we limit the calculation of this distance in the region 30Hz-2000Hz, a step that ensures less noisy results without discarding any relevant spectral areas. A final remark here concerns the set of mathematical properties of distance measures. As mentioned, RPWD can be viewed as the average Euclidean distance within two sets of points, and as such respects all the mathematical properties of a proper metric. Nevertheless, it is important to note that the RS itself is a non linear representation. The manner that the RS points are reassigned when even small amounts of the same distortion are added in different signals is not always the same. Therefore, the invariance property will not hold.

5.3 Experiments and results

In the following experiments, we study objective speech quality measures, particularly as far as their ability to characterize reverberant conditions is concerned. For this, we focus on the behaviour of the measures when the reverberation parameters, DRR and T_{60} (see Section 2.4.1) are varying. From the long list of distance measures presented so far, here, we concentrate on the CD, RCD and RPWD distances. The experiments reported in this section are performed in a synthetic environment called SQUARE room, shown in Figure

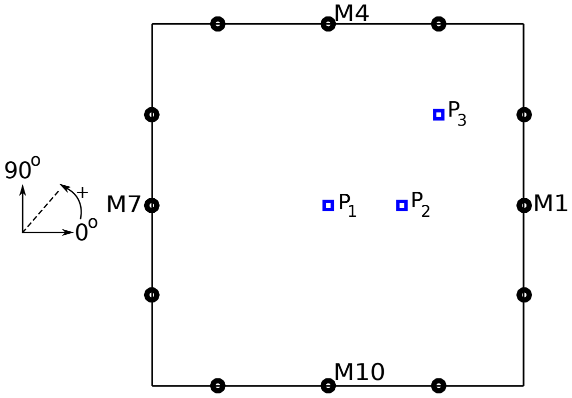


Figure 5.1: SQUARE room setting. Black dots indicate the microphone locations, and blue squares show the various simulated positions of speaker. For each position, the speaker may assume 36 possible orientations. Orientations are given as depicted on the polar coordinate system on the left.

5.1. The dimensions of the room are $4.80m \times 4.80m \times 2.7m$ and it is simulated using IRs generated with our IM tool, which offers the possibility to set the orientation of the source with a given acoustic directivity pattern. Moreover, it gives a fine control of several other parameters as, for example, T_{60}, which we vary in the range $0.2sec - 0.9sec$. The full set of IRs comprises the 3 positions shown in Figure 5.1, and for each position a set of 36 orientations. Each position and orientation is simulated for a set of four microphones[2].

For the experimental activities in the SQUARE room reported in the following, we use a data set which includes 30 sentences, and was created with a random selection of 5 utterances for each of the 6 speakers included in the WSJ0-5k DIRHA-English corpus (see Section 4.3.2).

5.3.1 The effect of the amount of speech

Number of utterances As mentioned, from a mathematical standpoint, the distance measures should follow the invariance property, which means that when adding the same amount of distortion to different signals, the distances of each distorted version from the original should be equal. In the case of reverberation the distortion is not additive, so the property is not exactly held, but in any case, the distance measure should not be critically affected by the amount of speech used, and the exact contents.

In a preliminary experiment we tested the dependency of the distance measures on the number of utterances and the utterance length. In Figure 5.2 we present the CD, RCD and RPWD measures between clean and reverberant signals, for increasing values of T_{60}. Each experiment is repeated for a different number of total test utterances. From this figure several observations can be made. First, we observe that, in general, all the investigated

[2] As the microphone names hint the original set comprises more than 4 microphones, which are exploited in other works, as for instance in [Guerrero, 2016].

measures behave in a way consistent with the reverberation parameters. In particular, for increasing T_{60} the various distance measures are increasing as well, indicating the increase of the amount of reverberation. This aspect will be further studied in the following section.

A second observation from Figure 5.2 is that, for all distance measures, and across all T_{60} values, the number of utterances does not largely affect the results. The use of a single utterance may lead to an over- or underestimation of the CD and RPWD distances, but with the use of 25-30 utterances measures start converging. In fact, the calculation of the measures in a per-frame basis and the averaging over all the frames of an utterance leads to this behaviour, and ensures results that do not vary largely among sentences. Particularly for the RCD notice that the various curves are closer to each other compared to the case of CD, indicating a better convergence, even with a reduced set of utterances.

Behaviour of RPWD Finally, in relation to this experiment it is interesting to demonstrate how the behaviour of the RPWD of a single utterance changes for different values of the T and Ω parameters, which define the time-frequency range within which each average value of the measure is computed. This effect is shown in Figure 5.3, where the RPWD between clean and reverberant utterances is shown, for different values of these two parameters. We observe, as expected, that small values lead to significantly noisy results, and, on the other hand larger values result in heavily smoothed output. Based on this type of analysis, and a more detailed investigation, we have selected the configuration $T = 0.3$ and $\Omega = 30Hz$ as the default values for our experiments.

Utterance length A more accurate analysis of the variations of the distance measures according to the utterance length is performed next. In Figure 5.4, we show the CD, RCD and RPWD measures as a function of the utterance length, for two T_{60} values. We observe that CD and RCD behave in a very similar fashion, and in general, the variations of the distances decrease for utterances longer than 7-8 sec. The results of the RPWD are not quite similar, as we observe generally larger variations, and, surprisingly, a lower variation of the measure for the larger reverberation time. This can be an indication that this measure is more affected by the contents of the utterance, *i.e.,* the spectral phonetic structure which gets smoother in more reverberation.

5.3.2 Relation with the reverberation

In the next experiment, we investigate the ability of the CD, RCD and RPWD distances to evaluate reverberant speech. Assuming various values for the T_{60} for the generation of the IR, we compute the average CD, average RCD and average RPWD between the clean and the reverberated signals. The results are presented in Figure 5.5, where it is shown

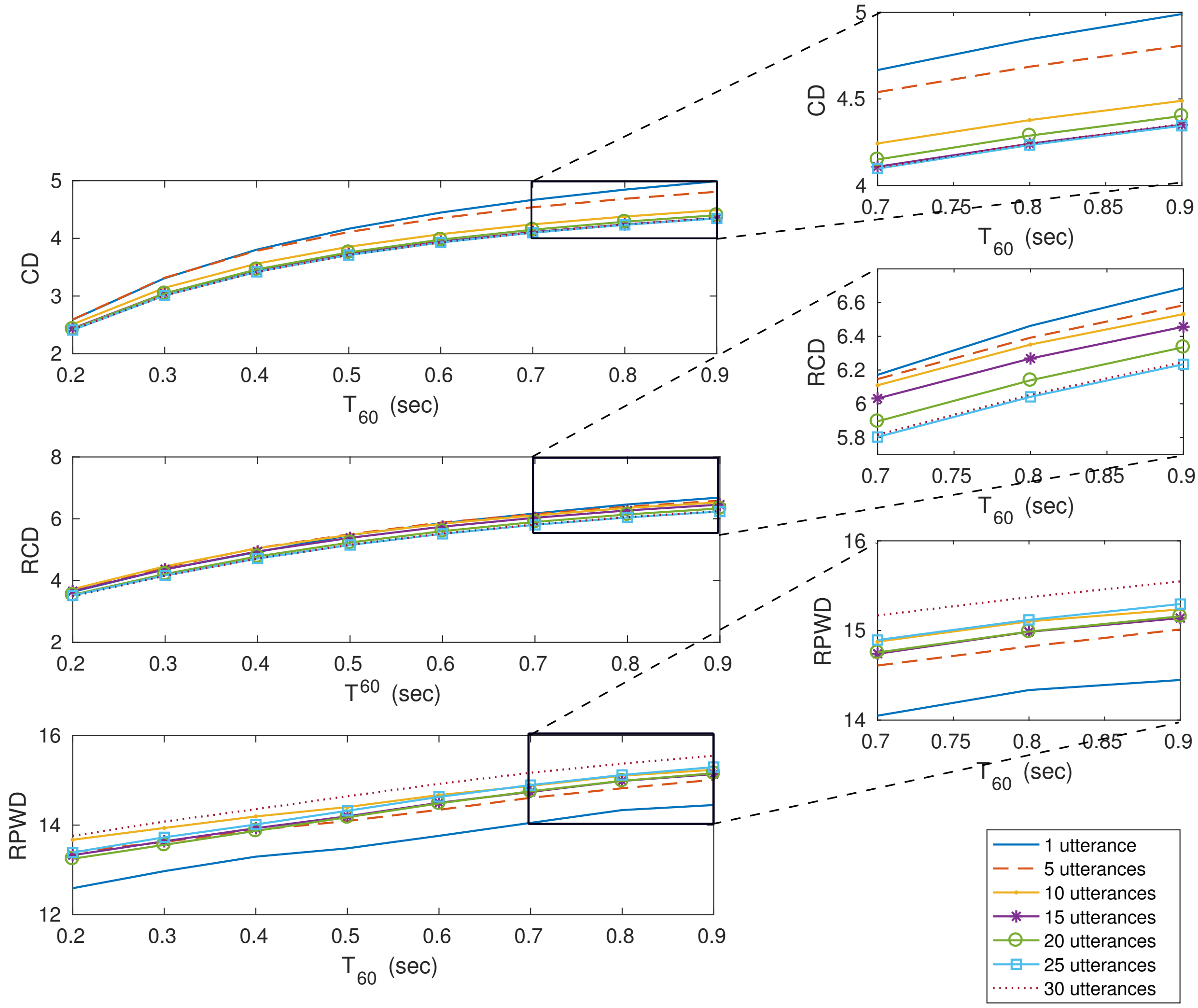


Figure 5.2: The CD (top), RCD (middle) and RPWD (bottom) measures between clean and reverberant signals, as a function of T_{60}. Each curve corresponds to a different number of utterances, as shown in the legend. The speaker is located at P_1 of the SQUARE room, adopting the orientation 0.^{o} The considered microphone is M1.

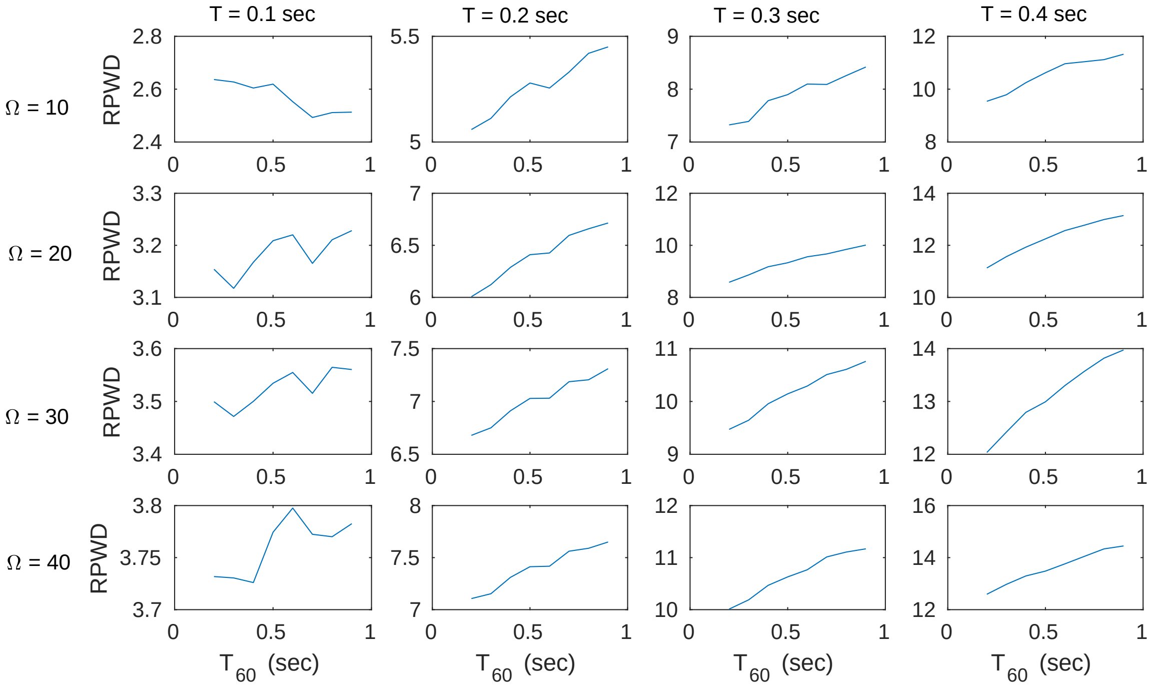


Figure 5.3: The RPWD between a reverberant and a clean speech utterance, for increasing T_{60}, and various values for the parameters T and Ω.

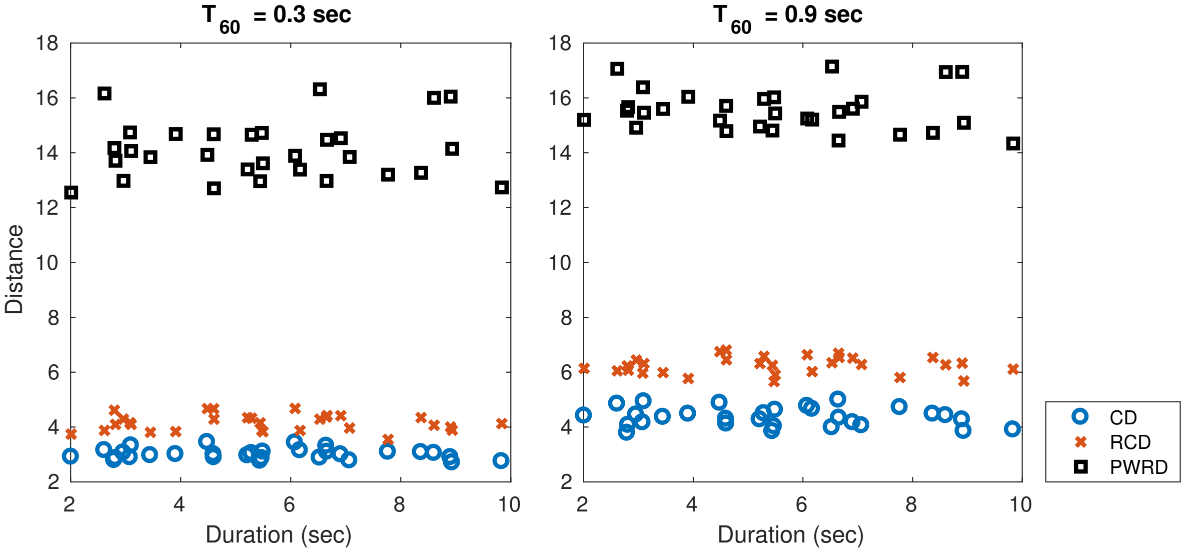


Figure 5.4: The investigated distances between clean and reverberant signals, as a function of the utterance length. The speaker is located at P_1 of the SQUARE room, adopting the orientation $0.^o$ The considered microphone is M1.

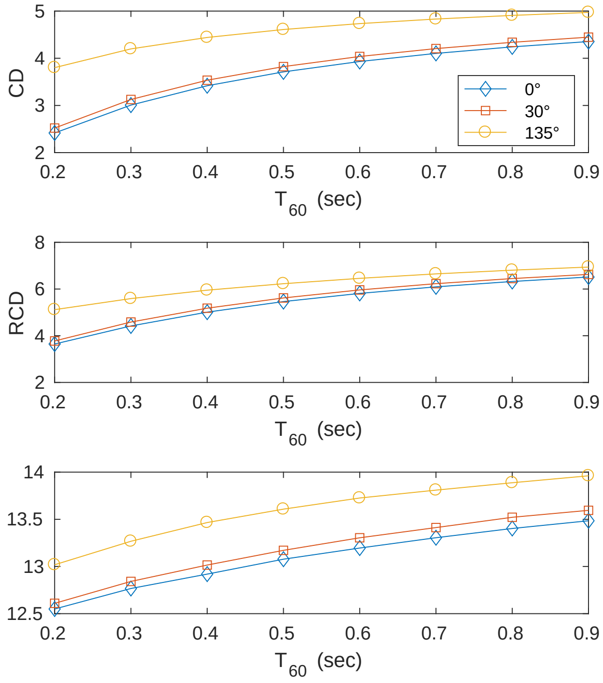


Figure 5.5: The CD (top), RCD (middle) and RPWD (bottom) between a reverberant to a close-talk signal, in terms of increasing reverberation time. The speaker located at P_1 of the SQUARE room, adopting three different orientations. In this experiment we consider only the microphone M1. Notice the better discrimination offered between 0° and 30° by RPWD, compared to CD and RCD.

that these measures have a behaviour very similar with the DRR. First, we notice that less directive cases result in larger measured distances between the clean and the reverberant signals. The RPWD seems to offer a slightly better discrimination between the most directive cases. Second, we observe that the average distances monotonically increase along with the increasing T_{60}. Taking into account the findings of Figure 2.12 the objective quality measures are expected to follow the DRR changes in an inverse fashion. Although not presented here, we found that this finding is valid for other objective speech quality measures discussed in the literature, such as fwSNRseg and LLR, and their reassigned versions.

In order to investigate the previously outlined relation between the reverberation parameters and objective speech quality measures, in the following experiments we report the variations of these measures under different positions and orientations of a speaker located in the SQUARE room setting, with a fixed T_{60} equal to $0.7sec$. The CD, RCD and RPWD measures between the clean and reverberated signals as a function of different orientations adopted by the speaker are presented in Figure 5.6. The DRR of the IRs used to reverberate the corresponding utterance is also illustrated. As expected, it is observed that when the speaker is oriented towards the microphone under consideration,

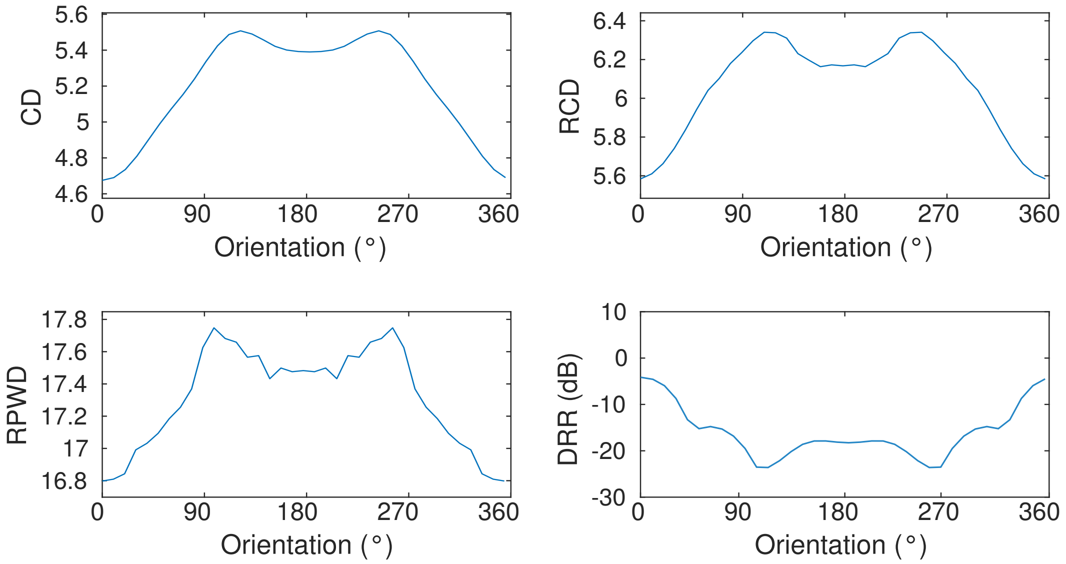


Figure 5.6: Objective quality measures, and DRR, as function of the speaker orientations, for an utterance simulated with the speaker located at P_1. Results are presented here for the microphone M1.

i.e., orientation 0^o, the minimum CD, RCD and RPWD values and maximum DRR value are measured. In addition, there is a clear inverse behaviour between the DRR and each of the objective measures evaluated here. Comparing the CD and RCD curves, we can see that the two measures behave in a very similar fashion, with RCD providing slightly a less smooth curve, which, however, should not pose any problems. From the RPWD curve, a limitation concerning this measure becomes evident. Although this measure offers a very good discrimination among orientations less than ±100°, its values become too noisy outside this region.

Next, we perform the above experiment for a different position, and additional microphones. We report only the results obtained with the CD and RCD measures as we showed that RPWD can be quite noisy. The set of CDs and RCDs between the close-talk signal and four reverberated instances, *i.e.,* microphones M1, M4, M7 and M10, as a function of different orientations are shown in Figure 5.7. We also present the DRRs of all the corresponding IRs. These results confirm the previous insights concerning the relation between CD/RCD and DRR.

In addition, this case illustrates how these parameters vary under more complex conditions, for instance when the speaker is oriented towards a microphone, but at a considerably larger distance. As an example, in Figure 5.7, we can compare the curves computed from signals captured by microphones M4 and M7, which represent cases of a near and a far microphone, respectively. When the speaker is oriented at around 130^o, *i.e.,* direct

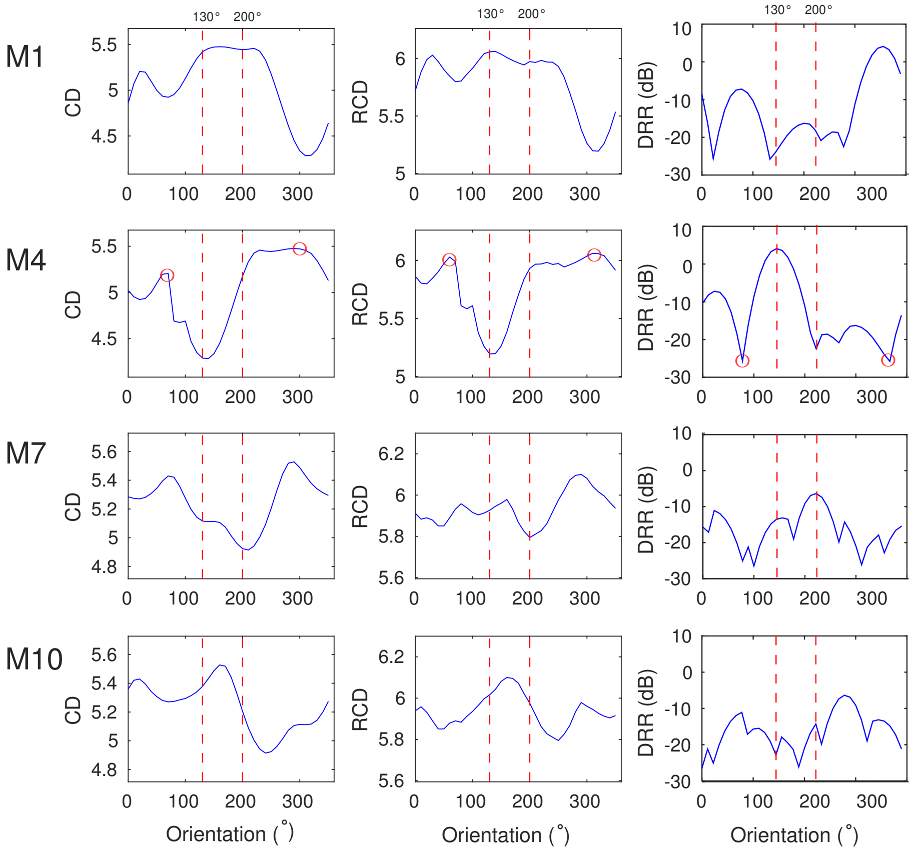


Figure 5.7: CD (left), RCD (middle) and DRR (right) as a function of different orientations for an utterance simulated with speaker located at P_3. Results are presented, from top to bottom, for the microphones M1, M4, M7 and M10. Lines in red emphasize two orientations of interest in the related discussion.

towards M4, in the related sub-figures there is a clear distinction of the lowest CD/RCD and highest DRR both over all orientations and over all microphones. On the other hand, when the speaker is oriented at around 200^o, *i.e.,* direct towards M7, although the curves are characterized by a minimum CD/RCD and a maximum DRR, the distinction of these values is not so clear. This can also be related to the average DRR decrease, and exposes a complex non discriminative scenario for the identification of the least distorted channel, even with the exploitation of prior information.

A final remark concerns the comparison between CD and RCD measures. Similar to the previously illustrated case (speaker position at P_1), the two measures produce very similar results. There are however some interesting variations. Notice for example the maximum values obtained in CD and RCD for microphone M4, which are marked with a red circle in the figure. According to our findings so far, these two orientations should correspond to the two minimum values marked in the DRR curve. We observe that, for the RCD this holds, as the marked maxima are the largest ones. In the case of CD this analogy is not valid. Similar observations can be done for other orientations and microphones.

The effect of thresholds on RS In Section 3.2.3 we discussed the possibility to prune the RS in order to emphasize either the harmonic, or the impulsive components of the input signal. This is performed with the application of a threshold on the mixed partial derivatives of the phase. In Section 4.5.2 we demonstrated how the choice of different values for these two parameters can significantly affect ASR results obtained with the use of the TFRCC features, *i.e.,* features extracted exploiting the RS of speech signals.

Here, we are interested to study the effect of these two parameters in the ability of the RS to characterize a reverberant environment. To study this aspect, we calculate the RCD between clean and reverberant for various values assigned to the sinusoidal and impulsive thresholds, used to prune the RS of the data. In Figure 5.8, we present these results assuming the SQUARE room setting, and for a speaker located at position P_1. First, we observe that, in general, changes of the sinusoidal threshold affect more the RPWD curves than corresponding changes in the impulsive threshold. For instance, notice the very low variations among the different curves presented in the first column of results, and compare to the extended changes between each consecutive pair of results in the first row. This can be probably attributed to the fact that, in speech, the impulsive content is relatively less than the harmonic. This fact means that any changes in this part affect less the overall result, which is obtained as an average over the whole duration of each test utterance.

Another finding from this type of analysis, is the indication to the best configuration

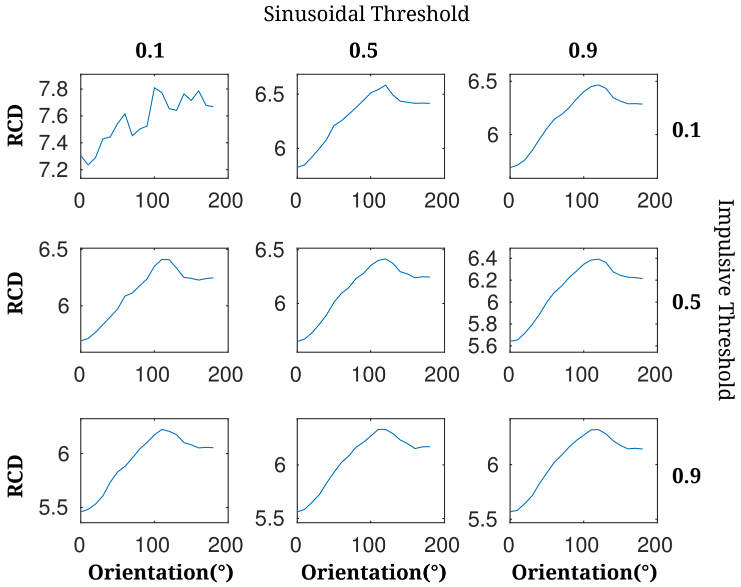


Figure 5.8: The RCD for different sinusoid and impulsive thresholds. The speaker is located at the P_1 position of the SQUARE room. The x axis corresponds to different speaker orientations, here in the range 0° - 180°.

of the RS when this is used for evaluating reverberated data. In particular, higher values for both thresholds lead to curves that characterize the various orientations in a way more consistent to the DRR parameter, which, according to our discussion so far, this is a useful property for a speech quality measure.

5.4 Conclusions

In this chapter, we presented various objective speech quality measures, and focused on their ability to characterize the amount of distortion due to reverberation, in a way that is consistent to common reverberation parameters. In particular, we investigated the CD measure, in its traditional and reassigned version, as well as a new measure defined on the RS of a signal. The findings from the presented experiments are several. First, we investigated how each measure is affected by the amount and length of the test utterances used, showing that all measures converge with a use of relatively small number of utterances. Moreover, we found that the behaviour of all studied measures is very similar to the behaviour of the DRR when different T_{60} values are assumed. At the same time, we found that the CD/RCD and RPWD are all good choices for discriminating among three different orientations assumed by a speaker, located in a room of increasing T_{60}. In the next experiments, we focused on a reverberant enclosure with T_{60}= $0.7sec$, and investigated closely the behaviour of CD, RCD and RPWD for multiple orientations of the speaker. From this, we verified the inverse relation between these measures and DRR, and identified certain cases which can be problematic in terms of characterizing the amount of distortion through an objective quality measure. Finally, we observed that in such a configuration RPWD can be disappointingly noisy, while CD and RCD operate in a similar manner, with RCD offering some improvements for certain orientations.

Chapter 6

Channel selection for multi-microphone DSR

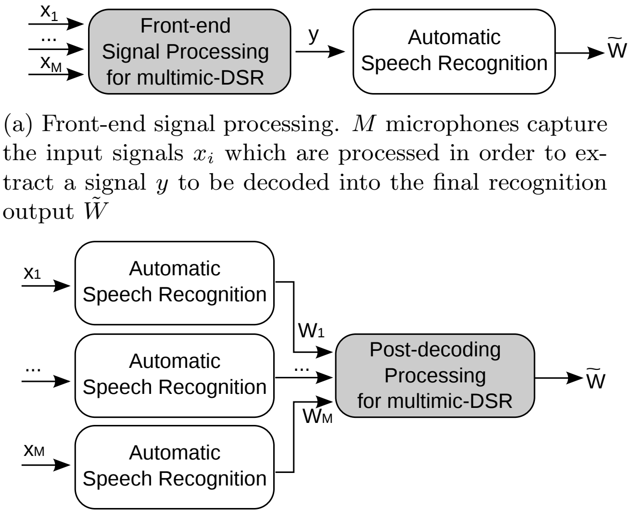
In the previous chapter, we demonstrated that objective speech quality measures offer a meaningful basis to characterize a reverberant signal, as they behave in a way consistent to the DRR and T_{60}, two commonly exploited reverberation parameters. In this chapter, we build upon this finding and incorporate objective quality measures in a DSR system, and in particular, a multi-microphone solution based on CS. First, in Section 6.1.1 we present some background on multi-microphone DSR, and in Section 6.1.2 we discuss the use of CS in this context. In Section 6.2 we propose a CS method based on CD[1], or, equivalently on the RCD. Following this, in Section 6.3 we present experimental activities in the SQUARE room setting, with a focus on demonstrating the behaviour of the proposed methods for a range of possible positions and orientations. In Section 6.4 we present CS results, for the same task presented in Section 4.5.3, exploiting both MFCC and TFRCC features.

6.1 Related work

6.1.1 Multi-microphone DSR

The use of multiple information sources is advantageous in mitigating the challenges introduced in speech recognition under distant-talking conditions. Inspired by the human auditory system, where two sensors (*i.e.,* left and right ears) are used in parallel for the understanding of the various acoustic stimuli, multi-microphone DSR solutions make use of multiple acoustic instances of the same signal, acquired by more than one sensors (*i.e.,* microphones) placed in the acoustic environment. These sets of microphones can adopt different forms, as for instance in the cases of *microphone arrays* and *distributed*

[1] This CS method stems from a collaboration with Cristina Guerrero.



(a) Front-end signal processing. M microphones capture the input signals x_i which are processed in order to extract a signal y to be decoded into the final recognition output $\tilde{W}$

(b) Post-decoding processing. Each signal x_i is decoded, and then individual recognition outputs are processed to extract the final output $\tilde{W}$.

Figure 6.1: Typical architectures for multi-microphone DSR.

microphone networks. Microphone arrays are compact placements of sensors, set up according to the demands of a particular signal processing mechanism that performs signal combination. Details such as the sensor characteristics, the geometry of the array, and the spacing among the sensors are all very important and well studied attributes that can affect the success of multi-microphone approaches employed after the signal acquisition. [Alvarado, 1990, Brandstein and Griebel, 2001, Flanagan et al., 1985, Huang and Benesty, 2007, Rabinkin et al., 1996]. On the other hand, distributed microphone networks comprise a limited number of sensors, which are not subject to any geometry constraints and can be placed on objects or mounted on the walls. All the microphones are connected to a computing system that ensures the synchronized capturing of audio signals. Compared to microphone arrays, the distributed microphone networks offer a broader spatial coverage.

Multi-microphone solutions do not depend only on the characteristics of the microphone networks, but also on the techniques used for processing the multiple speech signals. In a multi-microphone DSR scenario the goal of the recognition system is to process the multiple inputs and derive a single recognition output for the spoken utterance. For achieving this, different architectures can be adopted [Kinoshita et al., 2013, Wölfel and McDonough, 2009], with processing modules that operate either at front-end or at post-decoding processing level, as depicted in Figure 6.1.

As shown in Figure 6.1a, front-end approaches process multiple instances of the same acoustic information and produce a single input for the subsequent recognizer. Signals, or features are combined through methods such as beamforming, speech enhancement and feature combination. CS based on scores computed from the signals, or acoustic features is another valid front-end solution. Finally, an effective practice consists in combining front-end processing approaches. As an example, [Kumatani et al., 2011] presented a system where a selection of multiple channels was performed for applying beamforming on a reduced set of signals. In general however, beamforming limits the scope of a method to scenarios that employ microphone arrays, which are characterized by a limited distance between adjacent microphones. Inter-sensor spacing generally affects the resolution of spatial sampling in any array processing application [Van Veen and Buckley, 1988]. In particular, this problem becomes critical in distant-speech applications, due to the broadband nature of speech [Flanagan et al., 1985, Ward et al., 1995].

Post-decoding processing approaches perform a combination of information at the last stage of the recognition system, as shown in Figure 6.1b. Renown methods, such as ROVER [Fiscus, 1997] and Confusion Network Combination [Evermann and Woodland, 2000], require an individual, parallel recognition of each input signal before applying their combination algorithms. Other methods, such as decoder-based CS [Obuchi, 2004] have also been explored. The complexity and resource demanding nature of post-decoding processing solutions increases with the number of captured channels, as each acquired signal has to be decoded independently, before any combination or selection method is applied.

A detailed review of multi-microphone approaches for DSR, along with an extensive set of related experiments can be found in [Guerrero, 2016].

6.1.2 Channel selection

CS methods share the objective to detect the least distorted channel among the available ones, assuming that a better match will result between the selected channel and the acoustic models of the DSR system. As mentioned, CS can be applied either at front-end or at post-decoding level, commonly referred to as *signal based* and *decoder based* approaches, respectively. In both cases, one relies on a specific measure which is optimized for the final selection. According to the type of information exploited for the computation of their measure, these methods can be further categorized into *informed* and *blind methods* [Guerrero et al., 2016].

Informed methods exploit measures computed with the use of prior information. Although not directly applicable in a real scenario, the study of these methods is particularly interesting because it offers an understanding of the effectiveness of the related measures under diverse reverberant conditions. In addition, such methods can be explored to derive an upper-bound performance for a blind method that uses a similar score [Wolf and Nadeu, 2010]. In particular, the *oracle CS* which exploits the word error rate (WER) of each recognized signal in order to identify the best channel, can also be seen as an informed decoder-based method to use for reference purposes. Although most of the measures described in the literature can be easily modified to be used in an informed way, very few authors have performed such a study. In [Wolf and Nadeu, 2009] measured IRs were used to verify the assumption that DSR can be benefited from IR based CS. In [Wolf and Nadeu, 2010] the SNR and the position/orientation of the speaker were used for computing informed measures.

On the other hand blind methods use no prior information, and the scores are devised from the waveforms of the signals, or a more sophisticated representation. Blind decoder based CS methods use information such as the likelihoods or posterior probabilities, to assess the quality of each channel. Therefore, it is not possible to apply decoder based techniques independently from the ASR process. Some representative examples of such decoder based methods can be found in [Obuchi, 2004, 2006, Shimizu et al., 2000, Wölfel, 2007]. A detailed review of this topic can be found in [Wolf, 2013]. Although there is the assumption that decoder based measures present a higher correlation to WER in DSR, this has not been so far proven in the literature [Wolf and Nadeu, 2014].

Blind signal based CS methods include, among others, the use of energy and SNR, cross-correlation between signals [Kumatani et al., 2011], and the modulation spectra of the original and the beamformed signals [Himawan et al., 2015]. One of the most successful measures described in the literature is envelope variance (EV) [Wolf and Nadeu, 2014]. EV based CS exploits the fact that the reverberation smooths the energy of speech signals. This is observed as a reduction in the dynamic range of the envelope in the speech portions of the input signal. For the calculation of the EV measure, the filter-bank energies (FBE) $X_m(k,l)$ in channel m, sub-band k and time frame l, are first normalized as follows

$$\hat{X}_m(k,l) = e^{\log X_m(k,l) - \mu_m(k)} \quad , \tag{6.1}$$

where the mean value $\mu_m(k)$ is calculated over the logarithm of the FBE of the entire speech utterance. The mean normalized sequence of FBE is then compressed with the application of a cube root function, and the variance $V_m(k)$ of each sub-band k, for each channel m, is extracted. EV based CS selects the channel that maximizes the average variance over all channels:

$$\hat{M}_V = \arg\max_m \sum_k \frac{V_m(k)}{\max_m(V_m(k))} \quad . \tag{6.2}$$

The application of a different weighting for each channel and sub-band in (6.2) was proposed in [Wolf and Nadeu, 2014]. However, to the best of our knowledge, no further elaboration of this concept has been described, and no experimental evidence has been

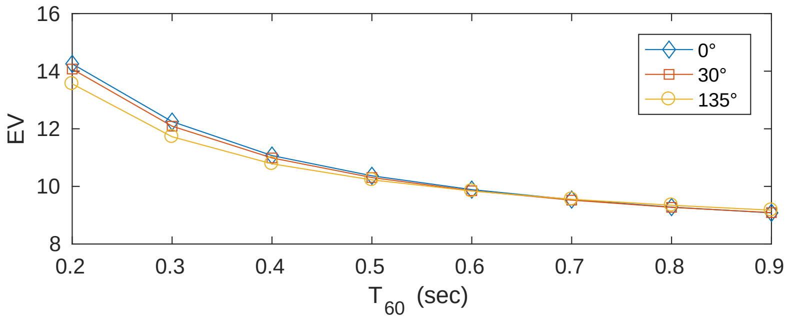


Figure 6.2: The average EV measure of artificially reverberated utterances, in terms of increasing reverberation time. The speaker is located at P_1 of the SQUARE room, adopting three different orientations. In this experiment we consider only the microphone M1.

derived to support the use of such a weighting scheme. In Figure 6.2 we present the EV of a reverberated signal, for increasing T_{60} values, and three different speaker orientations. Compared to Figure 5.5, notice the much lower discrimination power of EV among different speaker orientations, and the particular low performance for higher T_{60}.

6.2 CD and RCD based channel selection

In a multi-microphone scenario, with many microphones distributed in the room, DSR can be performed on any of the reverberated instances of the same utterance. Given the highlighted relation between reverberation parameters and objective quality measures, it is reasonable to assume that an objective measure will be advantageous in detecting the least distorted channel, in order to improve the recognition accuracy. Although we found that RPWD can, in cases, offer a better discrimination among reverberation conditions caused by different orientations (*e.g.,* see Figure 5.5, in the following approach we focus on the CD/RCD for various reasons. First, the CD is perhaps the most intuitive objective measure for signal quality, which as shown also applies well in cases of reverberation. Cepstrum-based comparisons are equivalent to comparisons of the smoothed log spectra of the signals [Rabiner and Schafer, 2011]. In this domain, the reverberation effect can be viewed as additive [Huang et al., 2001]. Furthermore, as discussed in [Rabiner and Juang, 1993], the CD has a particular frequency domain interpretation in terms of relationship between a set of signals and their geometric mean spectrum. These last attributes are exploited in the proposed CS method, as described in the following. It is noted here, that in the following we use the term CD to refer to any distance calculated in the cepstrum domain, *i.e.,* both to traditional and reassigned cepstra.

6.2.1 Informed channel selection

In the informed CS method we assume the availability of the close-talk speech signal, $x(t)$. Each distant microphone signal can be expressed as follows:

$$x_m(t) = x(t) * h_m(t) \tag{6.3}$$

where m is the microphone index, and $h_m(t)$ is the related IR. As previously pointed out and indicated in (6.3), in this work we are assuming that $x_m(t)$ is not distorted by environmental noise. Equivalently, in the STFT domain each distant microphone signal is expressed as

$$X_m(t,\omega) = X(t,\omega)H_m(t,\omega) \quad . \tag{6.4}$$

The *complex cepstrum* of $X_m(t,\omega)$ is defined as the inverse Fourier transform of its complex logarithm. In practice, as in many speech processing applications, the *complex cepstrum* is replaced here by the *real cepstrum*, which uses the logarithm of the magnitude of $X_m(t,\omega)$. This can be written as

$$\log|X_m(t,\omega)| = \log|X(t,\omega)| + \log|H_m(t,\omega)| \quad . \tag{6.5}$$

From this representation it can be inferred that the CDs $d(\mathbf{c}_x, \mathbf{c}_m)$ between the close-talk and the reverberant signals are more affected by the set of IRs than by the content of the spoken utterance.

Given the set of CDs $d(\mathbf{c}_x, \mathbf{c}_m)$, and assuming that the least distorted channel corresponds to the one *nearest* to the close-talk signal, the selection is performed as follows:

$$\hat{M}_x = \arg\min_m d(\mathbf{c}_x, \mathbf{c}_m) \quad . \tag{6.6}$$

6.2.2 Blind channel selection

In a real scenario the close-talk signal is not available, therefore we propose here a non-intrusive way to estimate CDs, from which CS is performed. The method relies on the assumption that one of the distant microphone signals is characterized by a higher DRR than the remaining ones. This typically occurs when the speaker is oriented towards that microphone and/or the speaker is located closer than the critical distance. The remaining channels are more affected by several degrading factors, for example attenuation effects due to the multiple reflections and to the head of the speaker.

Based on the above assumption, we proposed to compute a reference as the logarithm of the geometric mean of the signals $x_m(t)$, in the magnitude spectrum domain:

$$\hat{R}(t,\omega) = \log \prod_m |X_m(t,\omega)|^{1/M} \tag{6.7}$$

$$= \frac{1}{M} \sum_m \log |X_m(t,\omega)| \quad . \tag{6.8}$$

where $X_m(t,\omega)$ is the STFT of the signal captured by microphone m, and M is the total number of microphones.

The cepstrum computed from the reference is then used to calculate the distance between the reference and each microphone signal $d(\mathbf{c}_{\hat{R}}, \mathbf{c}_m)$. The least distorted channel can be selected as the one *furthest* from the reference:

$$\hat{M}_{\hat{R}} = \arg\max_m d(\mathbf{c}_{\hat{R}}, \mathbf{c}_m) \quad . \tag{6.9}$$

In order to better explain the proposed method, we elaborate on (6.8), which with the use of (6.5) can be rewritten as:

$$\hat{R}(t,\omega) = \frac{1}{M} \sum_m [\log |X(t,\omega)| + \log |H_m(t,\omega)|] \tag{6.10}$$

$$= \log |X(t,\omega)| + \frac{1}{M} \sum_m \log |H_m(t,\omega)| \quad . \tag{6.11}$$

The second term of (6.11) represents an estimation of the average reverberation that affects the multiple instances of the close-talk signal. Assuming to have a set of microphones uniformly distributed in space, with one characterized by a substantially higher DRR than the others, the resulting reference will be strongly influenced by the latter ones, i.e. it will be far from the former.

Of course, a favourable situation as the one previously outlined can not always be expected. For example, if all channels are equally impinged by reverberation, the selection of a specific channel is not relevant for improving the recognition performance. It is expected that in such cases the decoding of all the microphone signals will result in a similar recognition error rate. For this reason, we focus on scenarios in which CS is meaningful, *i.e.,* scenarios that feature the speaker at favourable positions and/or orientations in relation to at least one of the microphones.

6.3 Experiments in the SQUARE room

For the CS experiments in the SQUARE room, we use a data set which includes 120 sentences, referred to as WSJ120 data set. To create this data set we randomly selected 20 utterances for each of the 6 speakers included in the WSJ0-5k DIRHA-English corpus. Given the fact that each recognition experiment performed in this room is repeated for the whole data set at each position and orientation, a preliminary experiment showed that this is a sufficient number of utterances to consider.

6.3.1 Relation between speaker orientation and oracle channel selection

The proposed CS method relies on the assumption that one of the distant microphone signals is characterized by a higher DRR, and that this is closely related to the orientation of the speaker. Moreover we assumed that the selection and decoding of the channel that is less affected by reverberation, *i.e.,* has a higher DRR, can lead to improved recognition rates compared to SDM or a random CS. Before proceeding to the evaluation of the proposed CS methods, we study closer these assumptions, and the relation that exists between speaker orientation and recognition rates.

In Figure 6.3, we present oracle CS results obtained using WSJ120 data set in the SQUARE room setting. The recognition experiments are based on *tri3* acoustic models, which are trained on MFCC features. In order to better understand the oracle curve, let us associate it with the angles highlighted in Figure 6.4. We notice that the lowest error rates are achieved when the speaker is directly oriented towards one of the closer located microphones. Opposite to that, there are certain regions where an increase of WER is observed. These regions correspond to the following geometric conditions:

- the speaker is directed towards a corner of the room, and/or
- the speaker is directed towards a microphone that is clearly more distant than the remaining ones, and/or
- due to the symmetry of the geometrical problem (e.g. speaker in P2 directed towards M7) the microphone is impinged by a more significant contribution in terms of strong early reflections.

Table 6.1 presents a subset of the SDM recognition results. The first two rows correspond to cases in which the speaker is oriented towards M1. Notice how this condition is reflected into a much lower WER for the indicated microphone. The next set of orientations, around 60^o, corresponds to cases in which the speaker is oriented towards the top-right corner of the room. The last set of orientations, around 180^o, corresponds to

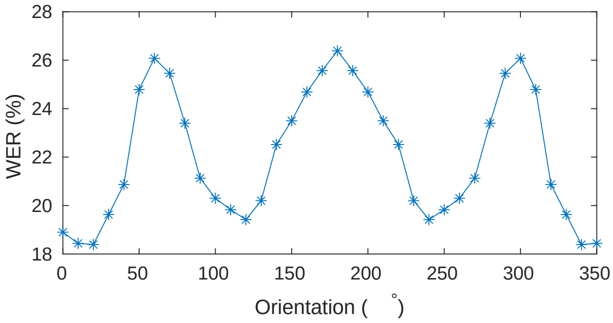


Figure 6.3: WER for the oracle CD when the speaker is located at the position P_2 of the SQUARE room, with microphones M1, M4, M7 and M10.

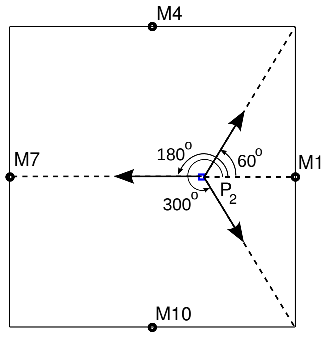


Figure 6.4: When the speaker is located at the position P_2 of the SQUARE room the orientations 60^o, 150^o, 210^o and 300^o correspond to the corners of the room. When the microphones M1, M4, M7 and M10 are considered the speaker is directed towards one of them at the orientations 0^o, 120^o, 180^o and 240^o, respectively.

cases in which the speaker is directed towards a more distant microphone. For the latter two angular regions, a slightly better performance is provided with M1 and M7, respectively. However, all the available channels produce very similar WER. Therefore, it can be argued that any type of CS, even the oracle one, is not relevant here.

6.3.2 Relation between CD/RCD and WER

The study presented in Section 5.3 provided an important basis for the use of objective quality measures, and in particular the CD and RCD, as a means for the selection of the least reverberant channel in a multi-microphone DSR scenario. Here, we further investigate the validity of our assumptions with a study on the relation between WERs and CD/RCD values between clean and reverberant signals.

In Figure 6.5 we present the scatter graph for these measures and WER value pairs.

orientation	M1	M4	M7	M10
0^o	19.4	33.1	35.6	33.1
10^o	20.7	26.6	35.4	39.8
...				
50^o	30.0	31.0	36.4	34.0
60^o	31.9	33.1	37.4	33.8
70^o	31.1	32.5	39.5	34.6
...				
170^o	35.2	32.4	30.0	36.8
180^o	35.1	34.7	29.8	34.7
190^o	35.2	36.8	30.0	32.4
...				

Table 6.1: SDM WER (%) for speaker position P_2 and microphones M1, M4, M7, M10 of the SQUARE room.

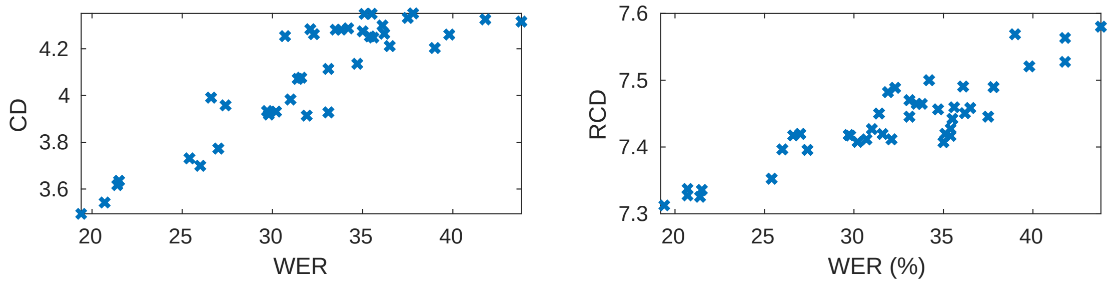


Figure 6.5: Distribution of the average CD (left) and RCD (right), between close talk and reverberant signals, with relation to the average WER achieved by the reverberant signals. Different points correspond to different channels, *i.e.,* M1, M4, M7, M10 of various orientations at position P_2. The acoustic models are trained on reverberated material.

Each point relates the average CD/RCD between the close-talk and the reverberated signals for the WSJ120 data set, and the average WER that results from the decoding of the reverberated signals, with *tri3* acoustic models and MFCC features. It is evident that CD and RCD are related to the recognition rate, as for both measures a clear trend can be observed that an increasing degree of signal distortion, corresponds to an increasing WER. Furthermore, we can observe that the RCD seems to correlate with the WER in a slightly more linear way.

A final remark from this experiment concerns the application of the proposed CS method for speech recognition, using acoustic models trained on reverberant speech. In the literature on CS, clean acoustic models are commonly used in order to evaluate the detection of the least distorted signal [Wolf, 2013]. Under such conditions, even an oracle CS results in a very low performance. However, the results reported in Figure 6.5 prove that the use of reverberant acoustic models, which guarantee a better overall performance, is a valid choice, as already shown in past work [Matassoni et al., 2002].

6.3.3 Relation between position/orientation and CD/RCD

Here, we examine closer the behaviour of CD and RCD based CS, for three positions and multiple orientations in the SQUARE room.

In Figure 6.6, we introduce a polar representation of CS experiments, in which the angle corresponds to the speaker orientation, and the radius to the rate, normalized to 1, at which each channel is selected. Horizontally, each row of polar graphs corresponds to a different position of the speaker, with the polar graphs showing the results of CD informed, RCD informed, and oracle CS respectively. In the latter one, it must be noticed that for some cases the same WER was achieved by more than one microphone. In such cases all the selected channels contribute equally to the rate represented in the polar graph.

Focusing first on the left column concerning the informed method, the results can be explained in a very intuitive way: the best channel corresponds to the microphone towards which the speaker is roughly directed. For example, at position P_1 the selected microphone changes every 90^o, with the region at which a microphone is selected centred around this microphone. When the speaker moves closer to M1 (position P_2) the region at which this microphone is selected is symmetrically expanded around it. An interesting observation results from position P_3, where the behaviour of the informed CS is different from the above cases, but can be related to reflections that arrive at the selected channel. For instance, for the orientation of 60^o, the selection of M1 can be explained by the first set of reflections that arrive at this microphone from the top wall.

In the next column, the polar graphs show the results of the informed RCD based CS method. Immediately, we notice a better agreement of the CS performed with the RCD method to the oracle CS, at least in the sense that, the selection is also affected from the contents of the utterance. Notice that the CD based method consistently selects the same channel, for each position/orientation, and results in very high selection rates for the majority of cases. On the other hand, the RCD and oracle methods select more than one channels for different sentences uttered on the same position and orientation.

6.4 Experiments in the DIRHA room

In this section we present the CS results for the multi-microphone DSR experiments performed in the DIRHA setting. It is noted here that, in contrast to the data sets addressed in the previous sections, each comprising the whole WSJ120 set simulated for a particular position and orientation, the corpora used here include a large number of mixed

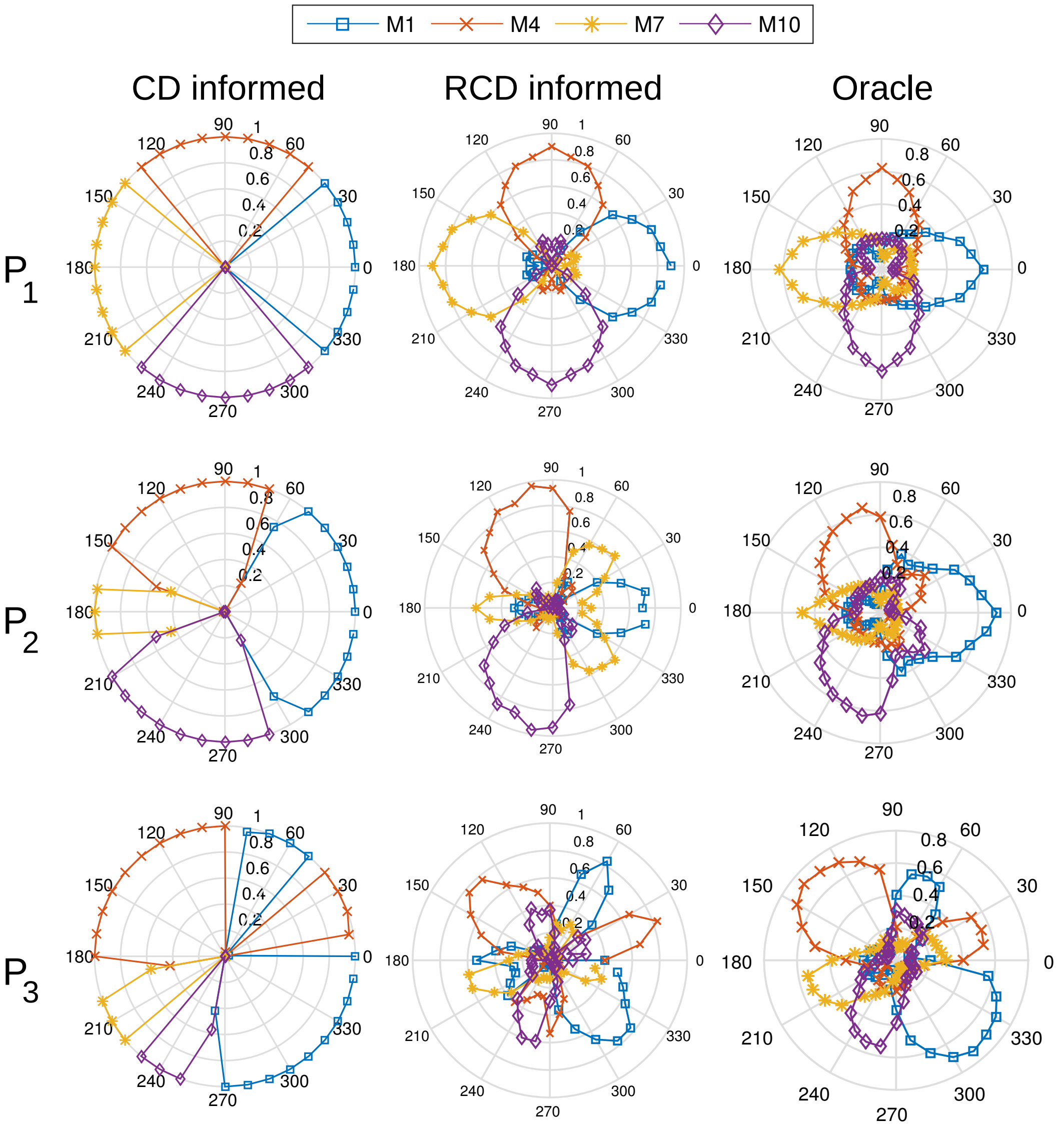


Figure 6.6: Channel selection with the informed CD (left), informed RCD (center) based methods, and oracle (right). The representation refers to multiple positions and orientations of the speaker, with the use of the microphones M1, M4, M7 and M10.

positions and orientations. These corpora are the sim-wsj and real-wsj sets, a simulated and a real version of the full wsj set. The corresponding SDM results have been presented in Section 4.5.3, where it was shown that the TFRCC features result in lower error rates for every type of acoustic model used, compared to the baseline MFCC features. In the following we report results based on *dnn* acoustic models, and compare the effect of each CS method with the use of MFCC and TFRCC features. In addition, we report the state-of-the-art EV method.

Table 6.2a contains the results for the sim-wsj dataset. First of all, we observe that, for

	MFCC	TFRCC	Rel.
SDM	16.22	15.74	2.9
oracle	10.16	9.57	5.8
CD informed	13.69	13.5	1.4
RCD informed	13.57	13.12	3.3
CD blind	14.71	14.28	2.9
RCD blind	14.63	14.1	3.6
EV	14.58	14.17	2.8

(a) real-wsj dataset

	MFCC	TFRCC	Rel.
SDM	15.42	14.88	3.5
oracle	9.67	8.49	12.2
CD informed	12.95	12.99	**-0.3**
RCD informed	12.53	11.75	6.2
CD blind	14.31	13.84	3.3
RCD blind	13.49	13.03	3.4
EV	13.74	12.87	6.3

(b) real-wsj dataset

Table 6.2: CS WER results (%) with DNN based acoustic models.

every CS method employed, the use of TFRCC features result in a better performance than the MFCC features, with a relative improvement ranging from 1.4% to 5.8%. Moreover, RCD based CS, both blind and informed, consistently leads to lower recognition errors, for both feature sets. The state-of-the-art CS method EV is found to slightly outperform both CD and RCD blind measures, for the MFCC features. However, when TFRCC are used for the recognition of the test data, RCD based CS closes this gap as well, and provides the best recognition rate among all blind methods. Table 6.2b presents the same experiment as above, for the real-wsj corpus, which contains real recordings that took place in the DIRHA room, and apart from reverberation there is some noise degrading the uttered speech. In these results, we observe that TFRCC based recognition achieves a relative improvement of 12.2% as far as the oracle CS is concerned. In addition, the use of the informed CD based CS leads to slightly improved recognition rate for the MFCC features. For this corpus the EV based CS method is found to outperform both CD and RCD based CS methods, for both types of acoustic models. With a closer look at the results, we found that a reason behind this is the particular set of positions and orientations in the real-wsj set, many of which exhibit a very low DRR. When the speaker is located far from the microphones or/and is directed away from the closest microphone, all channels are similarly affected by reverberation. This situation poses a limitation for the CD/RCD based method, and can be better addressed by the EV method.

In Table 6.3 we present some relative WER reduction rates for many of the results shown above. The most interesting finding from these numbers, is the increased performance that RCD demonstrated over CD informed and blind for the real-wsj, if compared to the sim-wsj. Since, as discussed, the real-wsj contains recordings with some environmental noise apart from reverberation, this can be an indication that the RCD is a more suitable measure in noisy situations.

	MFCC	TFRCC
oracle to SDM	37.3	39.2
CD blind to SDM	9.3	9.3
RCD blind to SDM	9.8	16.6
EV to SDM	10.1	9.9
RCD inf. to CD inf.	0.8	2.8
RCD blind to CD blind	0.5	1.3

(a) sim-wsj dataset

	MFCC	TFRCC
oracle to SDM	37.3	42.9
CD blind to SDM	7.2	7
RCD blind to SDM	12.5	12.4
EV to SDM	10.8	13.5
RCD inf. to CD inf.	3.2	9.5
RCD blind to CD blind	5.7	5.8

(b) real-wsj dataset

Table 6.3: Relative CD WER results (%) with DNN based acoustic models.

6.5 Conclusions

In this chapter, we investigated a multi-microphone DSR method, and in particular a CS technique that exploits objective speech quality measures. The focus was put on the CD and RCD measures, which were used for performing CS in an informed and a blind fashion. The contributions of this chapter are numerous. First, we proposed an effective approach to study CS for DSR. In particular, we designed a series of experiments that cover the possible source orientations, in a thorough way, under various speaker positions and microphone network configurations. From the corresponding results, CD/RCD measures were found to be well related to the recognition rate, as obtained by decoding reverberant signals. In addition, we showed that the informed CD/RCD measures resulted in an intuitive selection of the least reverberant channel. Finally, certain limitations of CS were outlined, for example when a clearly best channel is not available.

In the last set of experiments, we evaluated the use of CD and RCD based CS in combination with acoustic models trained on MFCC and TFRCC features. We found that the proposed CS method is a valid approach to improve recognition performance in a real multi-microphone setting. Overall, the proposed method produced results comparable to, or better than the state-of-the art EV method. In addition, the RS based CD led to improved CS results, even when MFCC based DSR was evaluated.

Chapter 7

Pitch and melody extraction

So far, we found the RS to be a powerful time-frequency representation of speech signals as it has been successfully exploited in different areas of speech signal analysis. One of the strengths of the RS is its ability to offer a much better visualization of the harmonic content of the signals, which can be further emphasized by the DRS introduced in Section 3.6. In this chapter we investigate further this representation, as we address the topic of pitch contour extraction for singing voice melody extraction. In the first part, Section 7.1, we provide some theoretical background on the topics of fundamental frequency, pitch and melody, overview various methods concerned with the extraction of these quantities and discuss the general underlying architecture as well as the importance of time-frequency representations in addressing such tasks. In Section 7.2 the proposed pitch contour extraction method, which is based on the RS of the input signal is presented in detail. Various experimental activities and related results in the context of melody extraction are described in Section 7.3.

7.1 Related work

7.1.1 Fundamental frequency, pitch and melody

As already described, in a complex pitched tone the frequency of each partial is an integer multiple of the fundamental frequency. Opposite to this objective definition of a quantity that can be measured, pitch is a perceptual attribute of sounds, which is used by listeners in order to characterize these sounds as high or low. According to the American National Standard Institute, pitch is defined as "*that attribute of auditory sensation in terms of which sounds may be ordered on a scale extending from low to high*" [ANSI]. This definition however, does not reveal the close relation between pitch and the frequency of the corresponding sound. Although the perceived pitch of a tone can be different from the perceived pitch of a sine wave with a f_0 equal to the one of the tone, the two quantities usually correspond very well [De Cheveigne, 2005, Klapuri and Davy, 2006]. An interesting exception is the case of the *missing fundamental*, when the pitch of a tone is perceived while there is no component with a similar f_0. This is because in pitch perception the auditory system processes as well periodicities which are implied by the relationships between the higher harmonics. More information on the relation between *fundamental frequency* and *pitch* can be found in [Hartmann, 1996, Rasch and Plomp, 1999, Terhardt, 1974].

A term which is closely related to fundamental frequency and pitch, particularly when music signals are concerned is the *melody*. Many different definitions can be found for this term, as for instance "*the dominant individual pitched line in a musical ensemble*" [Paiva et al., 2006] or "*the single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music, and that a listener would recognize as being the 'essence' of that music*" [Poliner et al., 2007]. Although any pitched instrument can attribute to the melody line of a song, it is the f_0 of the singing voice that, particularly in the pop culture, is identified more often as the melody line [Goto et al., 2010].

7.1.2 Pitch detection algorithms

Pitch detection from speech signals has a very long history, and although the clear distinction between pitch and fundamental frequency, most pitch detection methods actually perform f_0 estimation, since they do not take any perceptual models into consideration [Hess, 1983]. Similarly, in the context of this work we will not make a distinction and in the following, the terms PDA and f_0 estimation are used interchangeably. Many PDAs are described in the *time-domain* and they are concerned with the estimation of the period of a quasi-periodic signal, and its inversion to obtain the f_0. The use of the zero-crossing rate, firstly introduced in [Kedem, 1986], has been often exploited for pitch detection, for instance in [Rossignol et al., 1998]. Nevertheless, it has been indicated that this method can be problematic [Roads, 1996], in particular for noisy signals or for signals with a lot of energy concentrated around their f_0. Similar to the zero-crossing rate, there are methods based on the peak rate and the slope event rate. However, time-domain methods are usually not able to account for spectrally complex waveforms, such as those of speech and singing voice.

Another representation commonly exploited by PDAs is the *autocorrelation* function, which for periodic signals is itself periodic and its first peak indicates the period of the signal. Well known methods based on some type of cross-correlation are proposed in [Boersma, 1993, Secrest and Doddington, 1983] and are offered by various speech analysis tools, such as Praat[1] and the Speech Filing System[2]. With this set of methods problems appear when a complex waveform, with multiple harmonics is analysed and the first peak corresponds to a high order overtone, rather than the f_0. A popular method which attempts to address this problem in several ways in the YIN f_0 estimator [De Cheveigné and Kawahara, 2002]. For instance, YIN minimizes the difference between the waveform and its delayed duplicate, employs a cumulative mean function which de-emphasizes higher-period dips and includes a parabolic interpolation of the local minima.

Many methods operate on the frequency domain and use the spectral representation of the input in order to estimate the f_0 of a signal. One of the first methods operating on the spectral domain was based on the identification of the partials in the signal, using peak detection and a pair-wise comparison of these partials [Piszczalski, 1986]. Alternatively, the use of filters enabled the f_0 estimation by comparing the output of filters with different center frequencies, as when the passband of a filter lines up with a spectral peak the result is a higher output value. Various filter types have been used, such as Comb filters [Moorer, 1977] or narrow user-tunable band-pass filters [Lane, 1990]. Finally, the f_0 of a signal can be estimated from its cepstrum. The idea behind such methods is that the Fourier transform of a pitched signal presents regularly spaced peaks, which is the harmonic spectrum of the signal. In the log spectrum domain these peaks are reduced and a periodic waveform is produced, the period of which, *i.e.,* the distance between consecutive peaks, is related to the fundamental frequency of the original signal. The Fourier transform of this waveform, used for the cepstrum calculation, results in a peak at the period of the original waveform.

7.1.3 Melody extraction systems

The increasing interest in music related applications, for example the automatic transcription of audio recordings, the creation of karaoke files, and the music retrieval by singing or humming, has recently led to extensive research activities in the area of modelling the main melody of real world music recordings. According to the prerequisites of each specific application, the melody line has to be described in terms of a sequence of frequencies, transcribed into sung or played notes, or in the common case that the main melody is sung, expressed in terms of vocal effects, as for example tremolo and vibrato. All these closely related tasks demand different post-processing units for their solution, but they all have to incorporate a core module that detects and models the spectral regions of the input audio where the melody line concentrates its energy.

In the case of clean singing voice it can be argued that a speech based PDA will be

[1] `http://www.fon.hum.uva.nl/praat/`
[2] `http://www.phon.ucl.ac.uk/resource/sfs/`

a good candidate for f_0 estimation. However, more commonly than not, singing voice is part of more complex musical ensembles, comprising many harmonic and percussive instruments. The problem of detecting the pitch of singing voice in such polyphonic recording is addressed in melody extraction systems. Due to the similarities between PDA and melody extraction, the first successful solutions in the later were inspired by the extensive literature in the area of pitch extraction. In [Paiva et al., 2006] an auditory model is combined with the detection of peaks in the autocorrelation function and in [Marolt, 2004] the spectral modelling synthesis (SMS) harmonics plus noise model of speech is applied to music. However, the nature of music signals brings limitations in the success of such methods. First, a music signal may comprise many different instruments, with two or more notes from the same, or different, instruments sounding simultaneously. Furthermore, percussive sounds and inharmonicities may, in principle, take place at any moment and the vocal melody may interfere with partials of different sounds. For all these reasons, the use of pitch extraction methods, that are designed for speech or monophonic sound, does not produce the necessary results when singing voice in the context of polyphonic music is considered.

Among the numerous approaches that have been proposed for the melody extraction from polyphonic music signals, a general underlying architecture can be derived as depicted in Figure 7.1. The processing is usually separated into three distinct units. In the first unit, the spectral processing takes place. Based on the approach selected to process the input signal in the spectral domain, the systems are divided into two broad categories: the *salience-based* and the *separation-based*. Methods in the first category transform the input audio signal into a pitch salience signal, where each frequency is associated with a certain value of energy or salience. Sub-harmonic summation, firstly introduced by [Terhardt et al., 1982], is the most commonly adopted method for the creation of the salience signal, and it is used in [Jo et al., 2010, Ryynänen, 2008, Salamon and Gómez, 2012]. A diverse weighting of spectral peaks, based on a pair-wise comparison between peaks is proposed in [Dressler, 2009], but SHS is once again incorporated in the calculation of the partial peak weights.

On the other hand, *separation-based* approaches attempt to segregate the singing voice from the music accompaniment and perform the melody line tracking on the segregated vocal signal, assuming that the problems associated with the presence of the background music will be eliminated. In [Durrieu et al., 2009], the source/filter model is used to represent the singing voice, while the background is modelled as a sum of sources with distinct spectral shape. In [Tachibana et al., 2010], the Harmonic-Percussive Sound Separation (HPSS) algorithm is used, exploiting in this way the temporal variability of the melody compared to more sustained notes. Closely related to the above methods, a set of

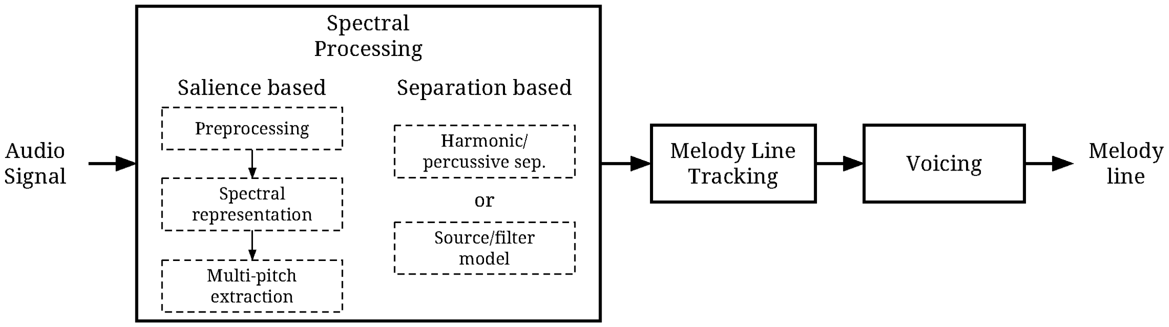


Figure 7.1: The general architecture of different approaches to melody extraction. In the spectral processing we find two distinct approaches, namely the *salience-based* and *separation-based* ones.

proposed solutions combine source separation techniques with the creation of a salience function. Such *hybrid approaches* have been proposed in [Hsu and Jang, 2010, Yeh et al., 2012], and the spectral processing is based on a salience signal extracted after an initial harmonic/percussive separation step.

In a comparison between *salience-* and *separation-based* techniques, the segregation of the vocal signal could, in principle, lead to superior results in the task of melody line tracking. However, the distortions that are introduced in the spectrum of the vocal signal cause significant loss during the melody tracking process. Currently, there are promising results with the use of *separation-based* approaches, but the best performing systems are *salience-based*, as can be seen in surveys and comparative studies, for example in [Poliner et al., 2007, Salamon, 2013]. On top of that, these types of systems are conceptually simpler and computationally more efficient. The above reasons support the focus of this work on *salience-based* approaches to melody extraction.

After the spectral processing in either a *salience-* or a *separation-based* melody extraction system, the second step concerns the tracking of the melody line. The goal of this step is to detect the spectral regions that are most likely to coincide with the melody line components. Different approaches adopted are dynamic programming [Hsu and Jang, 2010, Rao and Rao, 2010, Tachibana et al., 2010], tracking agents [Goto, 2004, 2005], score based clustering [Arora and Behera, 2013] and HMM [Ryynänen and Klapuri, 2006, Sutton et al., 2006, Yeh et al., 2012]. Streaming rules imposing time and frequency continuity in the melody lines, are implemented in [Dressler, 2009], while a combination of peak streaming rules and statistical characterization is applied in [Salamon et al., 2011].

Finally, the third step of melody extraction systems concerns the voicing detection. The goal is to determine the regions in time where the melody line exists as opposed to those that do not contain any melody. Inspired by the voice activity detection (VAD) in speech processing, many systems employ static or adaptive energy thresholds [Cancela and Magallanes, 2008, Dressler, 2009, Poliner and Ellis, 2005], while more sophisticated methods have also been described, as for example the incorporation of a silence model in the statistical framework used in [Ryynänen, 2008].

7.1.4 Spectral representations

The vast majority of melody extraction systems exploits the STFT for the transformation of the input signal into the spectral domain. The problems that arise from the use of the STFT, in the context of many diverse signal processing applications, appear in the case of melody extraction as well, and are extensively discussed in the literature. These concern the unavoidable trade-off between the time and frequency resolution, and the fact that the selected resolution is fixed over the whole spectrum. However, in melody detection a higher frequency resolution is desired in the lower areas of the spectrum, where more frequency components exist, while a higher time resolution can be of advantage in the middle frequencies where singing voice is normally located and presents relatively fast changes. Higher frequencies are not of interest as melody does not generally go over 1.5-2 kHz.

For these reasons, alternative representations have been proposed for the tasks of pitch tracking and melody extraction, as for example the Multi-Resolution FFT (MR-FFT) [Dressler, 2006, Hsu and Jang, 2010, Yeh et al., 2012], multirate filterbanks [Goto, 2005] and constant-Q transform [Cancela and Magallanes, 2008]. A different set of approaches attempts to improve the localization of the spectral energy after the calculation of the STFT, making use of frequency and time correction mechanisms as the ones discussed in [Keiler and Marchand, 2002]. Parabolic interpolation is incorporated in [De Cheveigné and Kawahara, 2002] and IF calculation in [Dressler, 2006, Salamon and Gómez, 2012].

7.2 A reassigned based melodic pitch extraction method

Similarly to the general architecture of melody extraction systems, depicted in Figure 7.1, in the first part of the proposed processing four main modules are incorporated: (i) preprocessing, (ii) spectral representation, (iii) multi-pitch extraction and (iv) post-processing. Each part is discussed in more detail in the following sections.

7.2.1 Preprocessing

The spectral preprocessing corresponds to the application of an equal loudness filter, which enhances the frequencies that are perceptually more relevant to the spectral areas where the melody line is normally found. The design of the filter is based on ReplayGain 1.0[3]

[3] http://wiki.hydrogenaudio.org

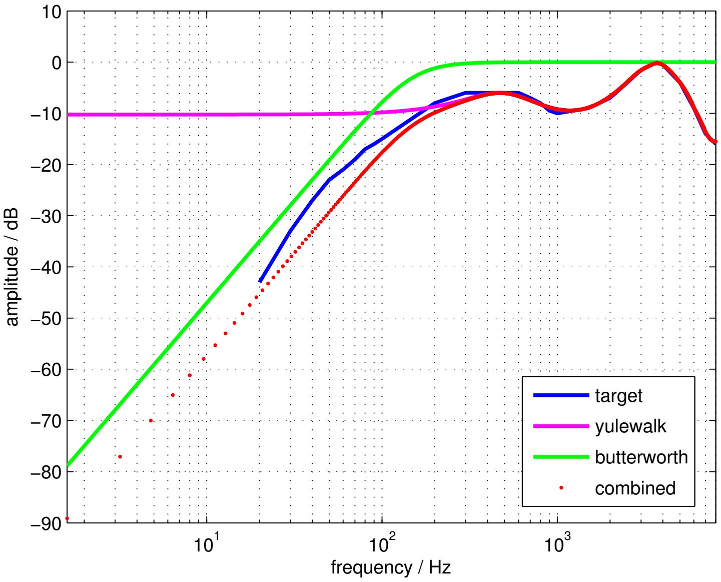


Figure 7.2: The frequency response of the applied filter. The final filter (red) results from the cascade application of a Yulewalk and a Butterworth filter. The ideal equal loudness filter is also shown.

specification. The use of this filter was motivated by the fact that it follows the human auditory system by enhancing the most perceptually dominant frequencies, and attenuates the rest [Robinson, 1958]. Furthermore, the resulting attenuation of the low-band frequencies, where instruments such as bass are found, is advantageous to the problem of tracking the singing melody. The use of perceptually inspired filters as a preprocessing step to the task of melody estimation is not a new concept, in fact it is preferred in most of the relevant methods. The filter is realized as a cascade application of a Yulewalk and a Butterworth filter as illustrated in Figure 7.2. After the application of the equal-loudness filter, the input signal is further processed with a low pass filter and downsampled.

7.2.2 Spectral representation

The way that the spectral energy of each time-frequency bin of the spectrogram is reassigned to a new TFR point is governed by the derivatives of the spectral phase at this bin, the same derivatives that theoretically result in the IF and GD of the analysed signal. It is worth mentioning here, that both of these quantities have been exploited in terms of pitch and melody extraction. For example, in [Rajan and Murthy, 2013] a set of modified group delay functions are used for melody extraction, since the presence of harmonic components corresponds to their local maximization.

With the RS at the core of the proposed processing, and instead of using the above mentioned quantities for the next steps, we propose the use of the dominance reassigned spectrogram to facilitate the extraction of the melodic pitch contour (MPC). The filtered input signal is transformed into the time-frequency domain with the STFT, and the reassigned time-frequency coordinates $(\hat{t}, \hat{\omega})$ of the TFR points are calculated. In order to count for the dominance of the components, we use the DRS as in (3.24).

7.2.3 Multi-pitch extraction

In this step, we are interested in creating a set of continuous pitch contours that characterize the melodic content of the music signal, starting from the set of dominance weighted TFR points, *i.e.* the DRS. To do so, we propose a tracking method that groups f0 candidates into pitch contours. Tracking is not a new concept, as grouping methods based on it are presented in [Cancela and Magallanes, 2008, Salamon, 2013]. However, in this work, there is an important difference at conceptual level. In the literature, grouping is normally incorporated in the melody line tracking step in order to combine, in time and in frequency, peaks that have been extracted from the spectrogram with some multi-pitch extraction method. On the other hand, the grouping is used in the proposed system as a method to detect regions of the RS where the TFR points are *connected*, and then, assign these points to unique pitch contours. The detected *connectivity* is an indication that in the corresponding RS region there is some underlying structure, which is related to the melodic components. This is a reasonable assumption to make because of the nature of the RS: in the areas that there is little or no energy to reassign, the TFR points are distributed in a noisy way.

Using the new DRS representation as a starting point, more TFR points, of higher dominance, are found around the melodic components. These areas, or neighbourhoods, of increased density around melodic components are detected and tracked in this processing block, forming MPC. In order to determine the set of candidate pitch contours, the iterative tracking Alg. 1 is implemented.

A neighbourhood N, of a central TFR point $(\hat{t}_c, \hat{\omega}_c)$ is defined as the spectral area that contains all the spectral points for which $|\hat{t}_c - \hat{t}| \leq \Delta\hat{t}$ and $|\hat{\omega}_c - \hat{\omega}| \leq \Delta\hat{\omega}$, where $\Delta\hat{t}$ denotes the maximum allowed time deviation from the center of the neighbourhood and $\Delta\hat{\omega}$ the maximum allowed frequency deviation. On the other hand, given a neighbourhood N, the central TFR point $(\hat{t}_c, \hat{\omega}_c)$ can be found as the *center of gravity* of it, as follows

$$(\hat{t}_c, \hat{\omega}_c) = \frac{1}{D_N} \left(\sum_{(\hat{t}_n, \hat{\omega}) \in N} D(\hat{t}_n, \hat{\omega}) \hat{t}_n, \sum_{(\hat{t}, \hat{\omega}_k) \in N} D(\hat{t}, \hat{\omega}_k) \hat{\omega}_k \right), \tag{7.1}$$

Algorithm 1 Multi-pitch extraction

```
1: **Input:** The dominance RS $D(\hat{t}, \hat{\omega})$, the RS $X(\hat{t}, \hat{\omega})$.
2: $E_{total} \leftarrow \sum_{\forall(\hat{t},\hat{\omega})} X(\hat{t}, \hat{\omega})$,    $E_{contours} \leftarrow 0$
3: **while** $E_{contours} \leq r_{min} E_{total}$ **do**
4:     Initialize a new pitch contour, $C$
5:     $N_0 \leftarrow \arg\max N \sum_{(\hat{t}_n,\hat{\omega}_n)\in N} D(\hat{t}_n, \hat{\omega}_n)$
6:     $i \leftarrow 0$
7:     **while** $|N_i| < N_{min}$ **do**
8:         $P_c(N_i) \leftarrow$ centerOfGravity($N_i$)
9:         Add $P_c(N_i)$ in $C$
10:        Remove $P_c(N_i)$ from $D(\hat{t}, \hat{\omega})$
11:        $N_{i+1} \leftarrow$ getNeighbourhood($P_c(N_i)$)
12:        $i \leftarrow i + 1$
13:    **end while**
14:    $E_{contours} \leftarrow E_{contours} + \sum_{(\hat{t},\hat{\omega})\in C} X(\hat{t}, \hat{\omega})$
15: **end while**
```

where D_N is the local dominance of N, calculated as

$$D_N = \sum_{(\hat{t}_n,\hat{\omega}_n)\in N} D(\hat{t}_n, \hat{\omega}_n) \quad . \tag{7.2}$$

At each outer iteration of the Alg. 1, the neighbourhood with the highest local dominance is selected as the starting point of a new pitch contour (see line 5, where the local dominance is maximized over all the different neighbourhoods N of the DRS). In the inner iteration, the *center of gravity* $P_c(N_i)$ of the neighbourhood under consideration is added to the current contour. The same point is used in order to update the neighbourhood before the following iteration, as described above, and then it is removed from the DRS. The tracking continues with the remaining points and it is exhaustive, meaning that a contour ends when the newest created neighbourhood is empty, or its cardinality $|N_i|$ reaches a certain threshold N_{min}, and both directions in time have been checked. The outer iteration stops when the energy of the created contours, $E_{contours}$, is more than a certain ratio, r_{min}, of the total energy, E_{total}, of the musical excerpt. The selection of adequate parameters for N_{min} and r_{min} is discussed in Section 7.3.1. Apart from creating pitch contours, this step acts as a de-noising operation, which ensures that the random noise, that the RS is known to suffer from, will be significantly reduced. This is also demonstrated in the experimental section.

7.2.4 Post-processing

After the extraction of a set of MPC, a post-processing step that detects and corrects harmonic sets is applied. The processing is based on the sub-harmonic summation matching theory of [Terhardt et al., 1982], which inspired a very successful pair wise evaluation of spectral peaks, proposed in [Dressler, 2011]. Here, we use the same idea of pair-wise comparison of pitch contours in order to detect harmonic sets and correct them by adding missing harmonic roots.

Each pair of MPC is controlled for existing harmonic relations. This is performed as follows. First, the harmonic number of the lower frequency MPC in the pair is computed as

$$h_{low} = \text{round}\left(\frac{af_{low}}{f_{high} - f_{low}}\right) \quad , \tag{7.3}$$

where f_{low} and f_{high} are the lower and higher frequencies of the pair, respectively. The coefficient a is defined as

$$a = \begin{cases} 1 & \text{for successive harmonics} \\ 2 & \text{for odd harmonics} \end{cases} \quad . \tag{7.4}$$

For a successive harmonic pair it holds that $h_{high} = h_{low} + 1$ and for an odd pair $h_{high} = h_{low}+2$. Each pair of MPC is ensured to belong to the same harmonic series, with numbers h_{low} and h_{high}, if the following criterion holds

$$1200 \log_2 \frac{f_{high}}{f_{low}} \log_2 \frac{h_{high}}{h_{low}} < 120 \quad . \tag{7.5}$$

If this criterion is ensured, meaning that the two contours are in a harmonic relation with a variance less than 120 cents from the exact harmonic interval and $h_{low} \geq 2$, the presence of the root MPC in the set of contours is checked. In the case this MPC is missing, it is added to the harmonic set.

7.3 Melody extraction

The experimental activities reported here concern, first, the optimal settings of some parameters of the algorithm, and, secondly, the system evaluation, which consists of three steps: (i) the extraction of pitch contour candidates, (ii) a comparative evaluation of the proposed method to the state-of-the-art system MELODIA[4] [Salamon, 2013], and (iii) a "glass-ceiling" analysis of the maximum possible accuracy. For the experimental activities we use datasets created in accordance with the MIREX guidelines for the task of melody line extraction, namely the ADC2004, MIREX05 and MIR-1k [Hsu and Jang, 2010] datasets.

Traditionally, the melody extraction systems are evaluated in terms of pitch and voicing accuracy. Over the past few years, through the annual evaluation of melody extraction

[4] `http://mtg.upf.edu/technologies/melodia`

systems, performed as part of the MIREX framework, it has been observed that the state-of-the-art is not improving in terms of these measures. Due to the complexity of the task, and the involvement of multiple steps in melody extraction systems, it is not always evident how each intermediate decision affects the final behaviour of a completed system. For this reason, in this work we are interested in the evaluation of the intermediate steps of a melody extraction system. Therefore, in order to create a set of measures that represent the quality of each intermediate step, we adopt evaluation measures similar to those introduced in [Keiler and Marchand, 2002] and also used, with minor changes, adapted for the evaluation of the salience function design steps in [Salamon et al., 2011] .

7.3.1 Statistical study

The proposed method uses a set of parameters, the settings of which can affect the final behaviour and performance. Concerning the spectrogram estimation, we use a window length of 30ms and a step of 5ms. These values are very commonly used for music related applications, and were experimentally proved to behave very well in the proposed system. For the calculation of the RS another important parameter is the frequency range, which starts at C_3 $(130.81Hz)$[5] and extends for four octaves, that in general include most of the energy carried by the melody line and its harmonics.

Finally, two parameters of Alg. 1, referred to as r_{min} and N_{min}, directly affect the behaviour of the proposed method.

r_{min}**:** This is the energy threshold which is used as a stopping criterion for the outer iteration. Figure 7.3 shows the average ratio of the total energy that is attributed to the melody line, and the first four harmonics of it (a subset of the MIR-1k dataset has been used for this analysis). As presented there, 0.5-0.7 of the total energy is attributed to the melody line and its first harmonic. When all four harmonics are considered, the total energy percentage can get as high as 0.95. Based on this, in the following experiments r_{min} is given values in the range 0.3-0.8.

N_{min}**:** This is the minimum number of TFR points that lie within a certain region and trigger the detection of a melodic component. The average number of TFR points in the regions of melody lines is shown in Figure 7.4. For the same data, the average number of points that lie outside the regions that are related to the melody line has been measured less than 10 points for the smallest meaningful distance under consideration (0.2 semitones). Further experiments showed that, in the range of 10 to 40 TFR points, higher quality MPC are extracted when 15 TFR points are

[5] Although it is possible for the melody line to be located below this value, this is not common in the used material, and therefore C_3 was selected for lower complexity.

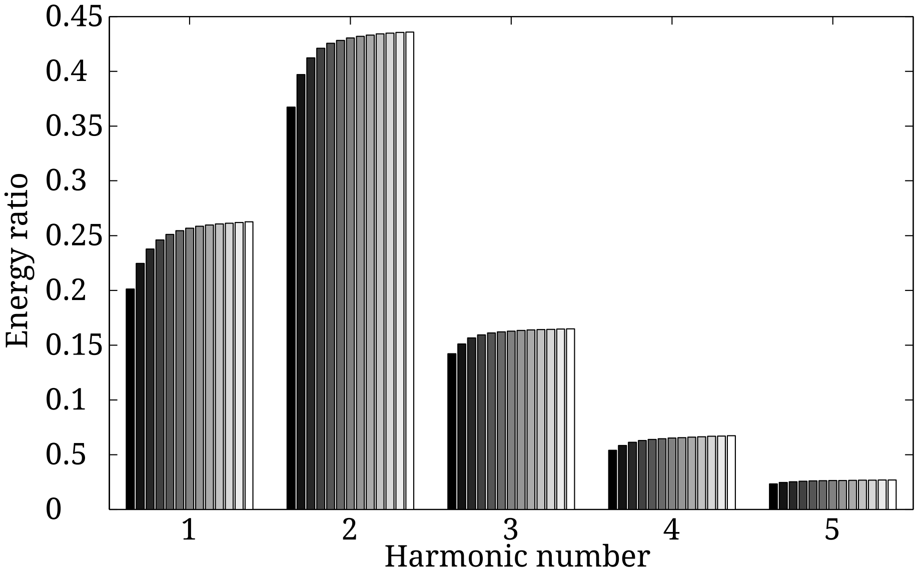


Figure 7.3: The ratio of energy that belongs to TFR points that are related to the melody line. Each group (2-5) corresponds to a different harmonic of the melody line (1). Each bar is calculated with a different confidence around the melody line, in the range from 0.2 (darkest bar) to 0.8 (lightest bar) semitones around the pitch.

considered as the upper threshold of regions that are not related to any melodic content. As before, a subset of the MIR-1k dataset has been used for this analysis.

7.3.2 TFR point-wise evaluation

Here, we study the ability of the proposed method to correctly identify the set of TFR points that are related to the melodic content of the piece, as opposed to the TFR points that are related either to the background music or to noise attributed to the operation of reassignment. For this, we study the following two approaches to TFR point discrimination:

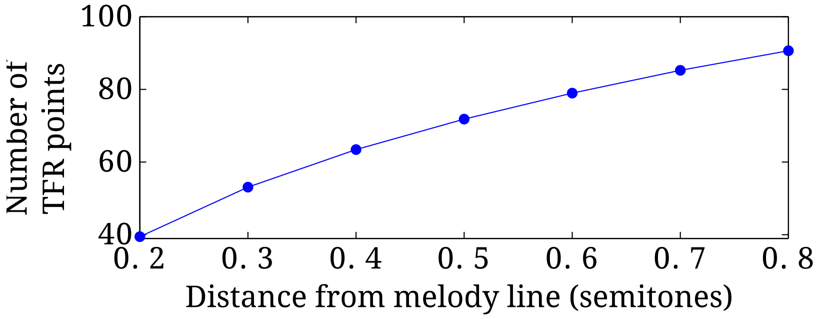


Figure 7.4: Number of TFR points in neighbourhoods of different sizes. Different f0 values have been used for this experiment.

Proposed method: Using the Alg. 1 we create two sets of TFR points. The first comprises all the points that have been assigned to an MPC, and are therefore considered relevant to the melody line. The second set contains all the remaining points.

MPD method: Inspired by the literature on reassignment, we impose the MPD criterion of [Fulop and Fitz, 2007], and create two sets of points. One comprises the points that meet the following condition

$$\left| \frac{\vartheta \phi^2(t,\omega)}{\vartheta t \vartheta \omega} \right| < A \quad , \tag{7.6}$$

where A is a tolerance factor that defines the maximum variation of an accepted component from the ideal sinusoid (relevant to the melody line), see Section 3.2.3 The remaining points belong to the second set. When the MPD method is used, the selection of A depends on the task, and usually it is experimentally found.

In order to quantify the results, we define the following evaluation metrics:

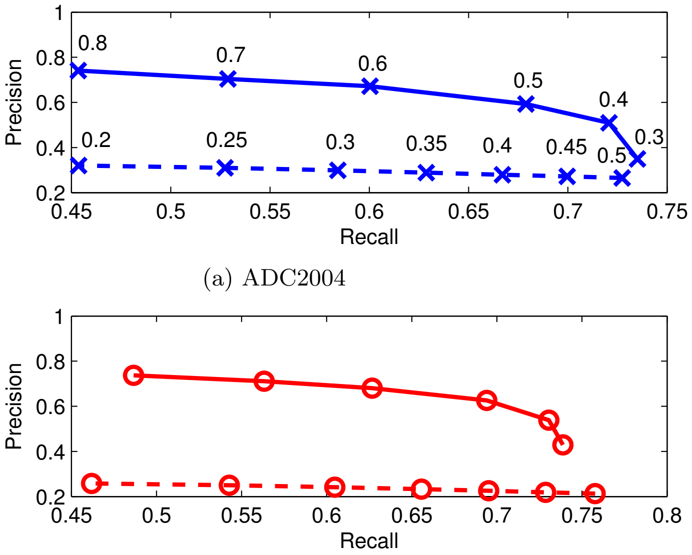
Point precision: The total number of relevant points retrieved, divided by the total number of points retrieved by the algorithm (the proposed or the MPD one).

Point recall: The total number of relevant points retrieved, divided by the total number of points in the ground truth.

Point f-measure: The geometric mean of the point precision and point recall.

Energy recall: The spectral energy sum for all the melody points tracked by the algorithm, divided by the total energy of the peaks in the ground truth. This is a measure relevant to the importance of the missed TFR points, as far as their total energy is concerned.

Starting from a music recording, to apply the measures introduced above, the ground truth is formed by the TFR points that lie within half semitone from the annotated melody line. The points are selected from the RS of the mixture and not the clean melody track recording as done in the literature, for example in [Salamon and Gómez, 2009]. The reason is that the use of the non-linear operation of the RS does not allow us to assume that the spectral points remain the same when the background music is added to the melody track. In fact, it has been experimentally found that the number of TFR points that remain the same is practically negligible. In this experiment, the post-processing that detects missing harmonics is not yet applied, since this concerns contours rather than sets of points that are evaluated here.



(a) ADC2004

(b) Mirex05

Figure 7.5: The Precision/Recall curves of the proposed (solid) and the MPD (dashed) methods in creating sets of TFR points related to the melodic content. As shown in the top figure, the points correspond to different r_{min} values for the proposed method, in the range 0.3 to 0.8, and different A values for the MPD method, in the range 0.5 to 0.2. The trend of the points is the same in the subsequent figures, but the labels are omitted for visualization purposes.

As commonly the case in information retrieval tasks, there is a trade-off between how precise (point precision) and how sensitive (point recall) each TFR point selection method is. In the proposed method, the exact behaviour in terms of precision/recall is controlled by the parameter r_{min}. In the MPD method this is controlled by the threshold value A introduced earlier. In the first set of experiments we are interested in the behaviour of the two systems in terms of precision/recall, which is depicted in Figure 7.5, for two different datasets. Although the two curves in these graphs correspond to different values of two different parameters, such a comparison is meaningful, since, in practice, each of them designates the strictness of the point selection method. From the curves, it is evident that the proposed method is much more precise in selecting TFR points that are related to the melody line.

In Figure 7.6 the point f-measure for the two datasets and methods is shown, again with a clear advantage for the proposed method. An interesting observation is that the f-measure is optimized when $r_{min} \approx 0.5$, which is in agreement with the initial study shown in Figure 7.3. In practice, the algorithm does not manage to correctly identify the melodic regions until the amount of energy attributed to the first two harmonics has been exhausted. This was expected since more energy was found in the region of the first harmonic that the region of the melody line.

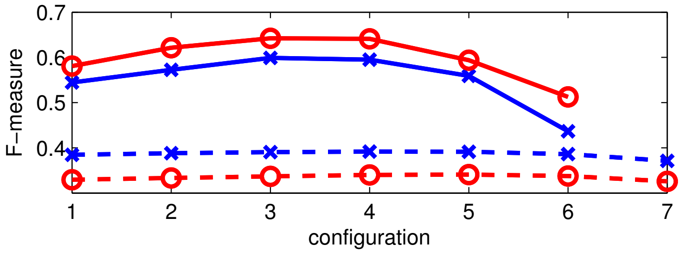
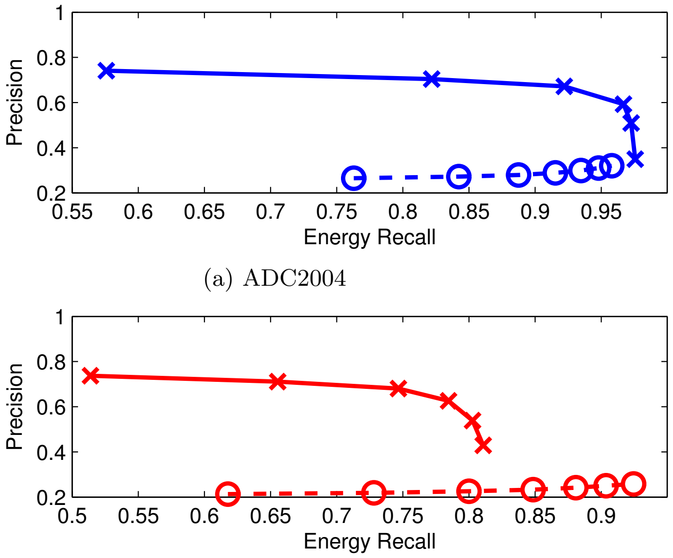


Figure 7.6: The f-measures of the proposed (solid) and the MPD (dashed) methods. The blue lines (marker: x) correspond to the ADC2004 dataset, and the red (marker: o) to the Mirex05 dataset. The configurations correspond to different r_{min} and A values, as described in the previous experiment.



(a) ADC2004

(b) Mirex05

Figure 7.7: The precision/energy recall curves of the proposed (solid) and MPD (dashed) methods in creating sets of TFR points related to the melodic content. The points correspond to different r_{min} and A values for the MPD method, as described in earlier experiment.

In Figure 7.7, the energy recall of the proposed and MPD methods is depicted, as a function of the point precision of each method. We observe that both methods are successful in selecting the TFR points that bare the most significant amount of energy of the harmonic components. Furthermore, it is shown that the MPD method is able to produce higher energy recall measures, especially in the case of Mirex05. Nevertheless, the corresponding precision values are too low to yield any useful application.

7.3.3 Evaluation of discrete pitch contours

To compare the proposed system with the state-of-the-art, an experiment was designed, mapping the retrieved TFR points into a new bi-dimensional grid, as in described in

Section 4.2. The resulting contours are post-processed with the algorithm described in Section 7.2.4, so that missing harmonic parts are repaired. The contours are evaluated with the following measures, as proposed in [Salamon et al., 2011]:

Contour precision Pr: The total number of melody contour points retrieved, divided by the total number of contour points retrieved by the algorithm.

Contour recall Re: The total number of melody contour points retrieved, divided by the total number of contour points in the ground truth.

Contour f-measure F : The geometric mean of the contour precision and contour recall measures.

The same evaluation measures are applied for the evaluation of the contour extraction process of the MELODIA vamp plug-in. The comparative results are presented in Table 7.1. As shown there, in both test datasets the proposed contour extraction method results in a higher f-measure, as compared to the MELODIA contour extraction method. In the ADC2004 dataset the proposed method improves both the precision and recall metrics, compared to the MELODIA method. Although this is not the same for the Mirex05 dataset, the proposed method is producing more balanced precision/recall pairs of values, and therefore results in higher f-measures for both datasets. This is an interesting finding as it means that the selection process that leads to the pitch contours, *i.e.,* the DRS and the multi-pitch extraction algorithm, is more successful than the literature method in retrieving points that are actually related to the melody line. Note that for the dataset Mirex05, the f-measure reported for the proposed method is higher than the one reported for MELODIA, although the corresponding precision/recall values do not follow the same trend. This is because the reported f-measure is an average itself, and not the f-measure of the average precision/recall values. In the partial results the precision/recall values of the proposed system are consistently more balanced than the ones of MELODIA. This fact, *i.e.,* the more balanced precision/recall values produced by the proposed system compared to the MELODIA in the partial result, explains the apparent inconsistency of Table 7.1. For the Mirex05 dataset, we observe that although MELODIA results in higher or equal precision/recall values, the f-measure is higher for the proposed system. This is observed since the average f-measure or each partial results is reported (and not the f-measure of the average precision and recall).

For completeness, the precision/recall curves of different configurations of the proposed system are depicted in Figure 7.8. From this representation, the effect of the post-processing step applied after the contour extraction becomes more evident: opposite to the previous experimental results, there is no trade-off between the precision and recall metrics and both are improving with the use of higher values for r_{min}. This is explained

Dataset	Method	Pr	Re	F
ADC2004	MEL.	0.58	0.7	0.63
	Pr.	0.75	0.8	0.76
Mirex05	MEL.	0.48	0.77	0.59
	Pr.	0.48	0.73	0.63

Table 7.1: Comparative results for two different datasets. The output of the pitch contour extraction step of the MELODIA vamp plugin (MEL.) is compared to the corresponding output of the proposed system (Pr.). For Mirex05, the f-measure reported for the proposed method is higher than the one reported for MELODIA, although the corresponding precision/recall values do not follow the same trend. This is because F is the average of the f-measures for each test excerpt, and not the f-measure of the average precision/recall values.

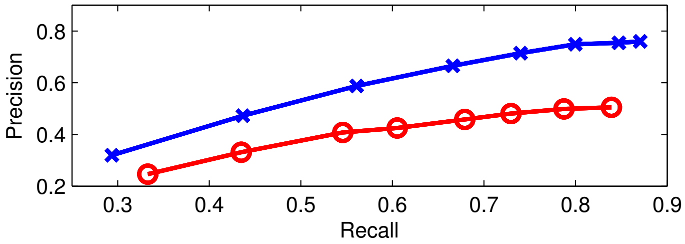


Figure 7.8: The precision/recall curves for the two datasets (ADC2004: blue-x, Mirex05: red-o), after the contour extraction and the post-processing steps. The points correspond to different r_{min} values, in the range from 0.1 to 0.8.

by the fact that, in the best cases, the post-processing improves the precision of the system by removing points that do not belong to any of the detected harmonic groups. At the same time, the post-processing improves the recall with the addition of the contours (*i.e.,* missing roots of harmonic groups) that have been missed in the earlier steps.

7.3.4 Glass ceiling analysis

Assuming a perfect pitch selection process, the melody can be identified correctly as long as there is a contour following this predominant component. This idea has been adopted in the literature for performing "glass ceiling" analysis. In Table 7.2 the results of this type of analysis are presented as reported in the literature, for the works of Salamon[Salamon and Urbano, 2012] and Dressler [Dressler, 2011]. Similarly here, we use it for analysing the proposed method. The pitch is considered to be correctly identified if there is one contour that lies within half semitone from the ground truth value. Based on this, we calculate the overall accuracy, considering that the voicing information is known. We observe that although the proposed system provides a satisfactory "glass-ceiling" result for one of the used datasets (Mirex05) and improves the reported state-of-the-art results, the same is not the case for the other dataset. It is important to indicate that, generally,

Dataset	Salamon	Dressler	Proposed	No pp
ADC2004	90%	93%	89.8%	68.6%
Mirex05	71%	77%	80.6%	63.6%
Average	80.5%	85%	85.2%	66.1%

Table 7.2: "Glass-ceiling" analysis assuming a perfect pitch selection step. The last column ("No pp") corresponds to the results of the proposed method, before the application of the post-processing step.

in the ADC2004 dataset the melody stands out from the background more clearly than in the Mirex05 dataset. This seems to affect the literature methods more than the proposed method which does not demand the presence of a very salient component in order to correctly identify it as a part of the melody line. On average for the two datasets, the proposed method yields the same result as the best of the two state-of-the-art methods. Finally, the last column of Table 7.2 reports the "glass-ceiling" results of the proposed system, when the post-processing is not applied. The improving effect of the harmonic group correction method becomes immediately evident.

Although the above results are satisfactory, it is worth noting here that the "perfect pitch selection process", assumed for this type of analysis, is benefited from a set of MPC that maximizes the recall measure, since the more relevant contours are retrieved, the higher the glass-ceiling will be. This however comes in contrast with the proposed method and parameter optimization, which is performed taking into account the behaviour in terms of both precision and recall.

7.4 Conclusions

In this chapter, we presented a method that detects the spectral regions of polyphonic music signals where melodic components are active, and groups these components in harmonic sets of MPC. The use of the RS in the core of the system provides a set of finely tuned contours that ensure the minimization of errors related to the limitations of the STFT. The MPC extraction algorithm is based on a dominance weighting of the TFR data, which is shown to successfully emphasize the points that belong to the most important harmonic components of the input signal. Finally, a post-processing step that detects and corrects harmonic groups is applied on the set of MPC. This step adds harmonics that were missed in earlier stages, and discards contours that do not respect the expected harmonic structure of the signal.

A main goal of this work was the design of a melody extraction system which incorporates blocks of processing that are shown to produce satisfactory intermediate results. Therefore, the experiments were designed so that each decision is proved to be advantageous to the final system. Indeed, the MPC extraction method has been shown superior to the only alternative method that is available in the literature as a means of selecting the harmonic points of a RS, namely the MPD method. Furthermore, the MPCs have been compared against the contours produced by the state-of-the-art system MELODIA and the results showed that the proposed method produces melodic contours of higher quality. Finally, the "glass-ceiling" analysis proved the benefits added by the proposed post-processing step, and also, the competitiveness of the proposed method to two state-of-the-art methods in the area.

Chapter 8

Conclusions and future work

The RS can be a powerful tool in the quest for effective voice based solutions in high-tech applications. As any time-frequency representation of non-stationary acoustic signals, the RS provides a description of the temporal evolution of the spectral components of the signal, while it improves the localization of these components. In this work, we studied various aspects of this representation, from its theoretical characteristics to the particularities of its implementation. Then, we proceeded to investigate the most critical aspect, *i.e.,* whether this representation can benefit the field of acoustic signal processing, specifically for speech and singing voice analysis. In the following we summarize, and discuss, the most relevant contributions of this thesis towards answering this research question.

8.1 Main contributions

Literature Review In Chapter 2 we compiled and presented an overview of the most relevant theoretical topics. This review highlighted several of the challenges faced in automatic analysis of human voice, and pointed out how these challenges are addressed from the most commonly employed feature extraction techniques in speech processing applications. Furthermore, we discussed the most interesting aspects that differentiate singing voice from speech. The effect of various acoustic conditions, such as the addition of reverberation or background music on a voice signal was also overviewed. A particular emphasis was given on the topic of reverberation, as its effects and characteristics play an important role in several of the topics addressed in this work.

In Chapter 3 we focused on the RS, and presented a comprehensive overview of the related literature. We emphasized the relation of this representation with the IF and the GD of the analysed signal, and discussed some of its limitations, as for instance, the inherent noise and the need for a re-quantization step. This study was then extended with the proposal of two different representations, based on the RS: the reassigned cepstrum and the DRS. Both representations were motivated and then discussed in the context of the different acoustic conditions that we focus on, *i.e.,* reverberation and musical background. The attention of the remaining chapters was shifted from theoretical topics to practically demonstrating the experimental value of the proposed representations. This was achieved by studying the RS, in the context of three distinct modules, each very important for a number of different acoustic signal processing application. These modules were (i) speech feature extraction, (ii) objective quality measures, and (iii) pitch contour extraction. Each of these modules was evaluated from the scope of final applications.

Speech feature extraction Feature extraction is a fundamental step in the majority of systems that are concerned with the automatic analysis and understanding of acoustic signals, and serves a variety of different purposes. The main goal of feature extraction is to summarize information, and keep all the information that is relevant for the subsequent processing while discarding anything that is redundant in some sense. In speech recognition one can argue that the only important information relates to the phonetic structure. However, when used as a front-end for complex statistical frameworks, feature extraction also serves the purpose of decorrelating the matrices used as observations, thus enabling their use with, *e.g.,* an HMM classifier. MFCC and PLP features have been used extensively in numerous systems, and have been the most popular choices for the front-end of phone level speech segmentation, and speech recognition.

As described in Chapter 4, this work contributed in the area of speech feature extraction with a set of acoustic features called TFRCC, which is a reassigned version of the MFCC features. Although TFRCCs are a rather direct concept, their implementation manifests certain aspects to address, as for example the problematic re-quantization step. In our proposed method, we combine this re-quantization step with the application of the Mel scale filter-bank in a single step which exploits bi-dimensional windowing. The TFRCC features have been experimentally verified in the contexts of phone-level speech segmentation and speech recognition.

Concerning phone-level speech segmentation, the TFRCCs were shown to behave in a similar manner as the MFCC features, while improving the accuracy of boundary detection under strict evaluation tolerances. We believe that this is directly related to the particularly sharp representation of the RS, which offers an excellent temporal resolution, while keeping a very good spectral resolution as well. The experimental activities showed however that the application of the IDCT at the last stage of the TFRCC feature extraction did not yield the expected improvements, compared to the use of the filter-bank output as a feature set. This suggests a future direction, where alternative methods to decorrelate the filter-bank output as obtained from the RS can be investigated. As far as ASR and DSR experiments is concerned, the presented experimental activities regarded different corpora, acoustic models and front-end configurations. TFRCC features showed consistent improvements upon the MFCC features, both in close-talk and reverberant conditions.

Objective quality measures Objective speech quality measures have been exploited for many years in the fields of speech processing as a means of evaluating diverse distortions introduced in the signal, as for instance distortions due to a communication channel, or an enhancement algorithm. Although subjective listening test is the most accurate method to evaluate the quality of speech signals, the very demanding process and strict rules that need to be followed shifts the interest to the alternative of objective measures. Numerous measures have been studied in the literature, and extensive activities investigated their correlation with the outcome of listening tests, for evaluating different tasks, showing still not satisfactory results.

Inspired by the power of the RS to represent speech, which was supported by the success in exploiting it for feature extraction, we investigated how the RS can be further used for objective speech quality measures. In Chapter 5 we proposed the use of reassigned versions of traditional objective speech quality measures, and one inspired by the continuous nature of RS data, called RPWD. We discussed on the properties of these measures, and investigated their dependency on the amount of speech used to evaluate them. From an experimental standpoint, we were particularly interested to address evaluation of distortion due to reverberation. In the literature, reverberation is normally “measured” either based on parameters such as DRR and T_{60}, or with the effect it has on the output of a particular system, *e.g.,* how much it can lower recognition rate. Instead, we investigated how objective distance measures can be exploited to quantify reverberation effects. We studied the ability of traditional, and reassigned measures to evaluate distortions due to reverberation. Our findings were numerous, as we found that objective measures, and in particular the CD and RCD are very good in characterizing reverberation in a way consistent with DRR and T_{60}. Concerning the RPWD measure it was found to produce disappointingly noisy results. Nevertheless, we believe that it is a promising measure which deserves further investigation as it is defined on the raw RS data and it is free from errors due to the re-quantization and further processing steps.

In Chapter 6 we extended the use of objective speech quality measures in the field of DSR and, in particular, for multi-microphone processing based on CS. We proposed a novel method to perform CS in an informed way, and extended this to a blind solution. The contributions of this work were also extended to the experimental findings. In evaluating the proposed method, we also demonstrated the relation between the oracle CS and reverberation, as measured through the DRR. In addition, we demonstrated the relation between CD/RCD and WER, and the intuitive nature of the proposed CS. To our knowledge, this work constituted the first systematic attempt to analyse CS and relate it to the characteristics of the reverberant environment within which the experiments are performed. Finally, the proposed CS methods were evaluated in terms of recognition performance and it was found that the RCD has the potential to lead to improvements over the CD, especially when used jointly with TFRCC features.

Pitch contour extraction The last module addressed in this thesis was the extraction of pitch contours, mainly focused on singing voice. A set of finely tuned contours, that describe the temporal evolution of the predominant harmonic components, which we called MPC, can be exploited mainly for the task of melody extraction from polyphonic music, although it can be also used in the core of a PDA. Apparently, the time-frequency representation exploited in systems concerned with melody extraction, and in case pitch detection, is of critical importance to their success, as the final results are largely based on this representation. In the literature there have been multiple attempts to improve the resolution of the spectrogram, for instance through the MR-FFT, as a means of improving the accuracy of melody extraction.

In Chapter 7 we proposed to use the RS, and perform melody extraction exploiting the infinite resolution it offers. The improved representation of the harmonic structure, offered by the RS and further enhanced by the DRS, was utilized for the design of a MPC extraction algorithm, *i.e.,* the first step towards pitch extraction from speech and melody extraction from polyphonic music signals. The proposed algorithm operated directly on the raw dominance weighted RS data, a fact that minimized fine pitch errors due to re-quantization and smoothing.

Finally, experimental activities demonstrated the ability of the proposed method to detect the predominant harmonic components in polyphonic music signals, and produce melodic contours that describe these components. The evaluation was performed, first, by directly evaluating the sets of selected points with a comparison to the points obtained with the application of a MPD criterion. The proposed algorithm showed overall a better f-measure than this baseline, proving its ability to detect more accurately the TFR points that relate to the predominant melody. Following this, we evaluated the discrete MPC and found them to compare favourably to those obtained from two state-of-the-art systems. Finally, a “glass-ceiling” analysis showed the potential to outperform these systems.

8.2 Future work

The different results, presented in this thesis, report the positive effects that the RS can have on various tasks related to the analysis and understanding of speech and singing voice signals. Nevertheless, there is still room for improvement within most of the proposed solutions.

First, further investigation is needed regarding the MPD criterion, as a means either to de-speckle the RS, or to emphasize its impulsive or harmonic components. The use of the second order derivatives of the phase, as proposed by the MPD criteria, can lead to an improved visualization of the RS. However, it is not yet clear how exactly this method can be exploited for an improved representation of speech, through the TFRCC features and the DRS representation.

Another direction that we believe needs to be further explored is the behaviour of the RS under noisy acoustic conditions. Throughout the experimental activities reported in this thesis, we had various indications that the RS is particularly strong in representing signals degraded by sources of distortions other than reverberation and music. Initial experimenting showed that in noisy data TFRCC features were more robust than MFCC in recognition experiments and that RCD was better than CD in evaluating reverberant and noisy speech utterances in the context of CS. This research direction demands further attention.

Finally, concerning melody extraction, we are interested in the design of a novel method to select the MPC subset that relates to the melody component, when more than one harmonic groups are active simultaneously. As mentioned earlier, this step should be different than what is normally found in the literature within the melody line tracking component and, possibly, has to focus on the discrimination between contours that result from different sound sources. This aspect can be studied as a point of connection to the extensive work that exists in the area of audio source separation and separation-based approaches to melody line extraction.

The qualitative and quantitative evaluations that were performed and reported in this thesis, support the use of the RS as a valid approach for the time-frequency representation of acoustic signals. In general, we identified two main sources of discrepancies as fas as the RS is concerned and these are the presence of random like noise and the need for a re-quantization step. We experimented with default and novel solutions for both of these two gaps, and showed that it is possible to improve the current state-of-the-art. Nevertheless, we believe that there is still room for improvement and that an optimal framework to work with the RS, managing the inherent noise and the raw data defined in the continuous time-frequency domain, is yet to be described.

Bibliography

Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. ITU-T, ITU-T Rec P. 835, 2003.

J. B Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

J. B. Allen and L. R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.

V. M. Alvarado. *Talker Localization and Optimal Placement of Microphones for a Linear Microphone Array Using Stochastic Region Contraction.* PhD thesis, BROWN UNIVERSITY, 1990.

ANSI. American national psychoacoustical terminology - s3.20. New York.

V. Arora and L. Behera. On-line melody extraction from polyphonic audio using harmonic cluster tracking. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21 (3):520–530, 2013.

F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, 1995.

F. Auger, O. Lemoine, P. Goncalvés, and P. Flandrin. The time-frequency toolbox. `http://tftb.nongnu.org/`. [Online; accessed 2017].

F. Auger, P. Flandrin, Y. T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.T. Wu. Time-frequency reassignment and synchrosqueezing: An overview. *IEEE Signal Processing Magazine*, 30(6):32–41, 2013.

R. G. Baraniuk, P. Flandrin, A. J. Janssen, and O. J. Michel. Measuring time-frequency information content using the rényientropies. *IEEE Transactions on Information Theory*, 47(4):1391–1409, 2001.

J. Barker, R. Marxer, E. Vincent, and S. Watanabe. The Third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines. In *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

J. R. Bellegarda and C. Monz. State of the art in statistical methods for language and speech processing. *Computer Speech & Language*, 35:163–184, 2016.

J. Benesty, S. Makino, and J. Chen. *Speech Enhancement*. Springer, Berlin, Germany, 2005.

C. M. Bishop. Pattern recognition. *Machine Learning*, 128:1–58, 2006.

B. Boashash. Estimating and interpreting the instantaneous frequency of a signal. i. fundamentals. *Proceedings of the IEEE*, 80(4):520–538, 1992.

B. Boashash and V. Sucic. Resolution measure criteria for the objective assessment of the performance of quadratic time-frequency distributions. *IEEE Transactions on Signal Processing*, 51(5):1253–1263, 2003.

P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences 17*, pages 97–110, 1993.

B. P. Bogert, M. J. Healy, and J. W.Bra Tukey. The frequency analysis of time series for echoes: cepstrum, pseudo-auto covariance, cross-cepstrum, and shaft cracking. In *Proceedings of the Symposium on Time Series Analysis*, pages 209–243, 1963.

M. Brandstein and S. Griebel. Explicit speech modeling for microphone array applications. In M. Brandstein and D. Ward, editors, *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 7, pages 133–153. Springer, 2001.

F. Brugnara, D. Falavigna, and M. Omologo. Automatic segmentation and labeling of speech based on Hidden Markov Models. *Speech Communication*, 12(4):357–370, 1993.

A. Buzo, A. Gray, R. Gray, and J. Markel. Speech coding based upon vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):562–574, 1980.

D. Cabrera, J. Xun, and M. Guski. Calculating reverberation time from impulse responses: a comparison of software implementations. *Acoustics Australia*, pages 1–10, 2016.

P. Cancela and S. Magallanes. Tracking melody in polyphonic audio. mirex 2008. In *In MIREX Audio Melody Extraction Contest Abstracts*, 2008.

J. R. Carson and T. C. Fry. Variable frequency electric circuit theory with application to the theory of frequency-modulation. *Bell System Technical Journal*, 16(4):513–540, 1937.

M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

J. Cohen, T. Kamm, and A. G. Andreou. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. *The Journal of the Acoustical Society of America*, 97(5):3246–3247, 1995.

L. Cohen. Time-frequency distributions-a review. *Proceedings of the IEEE*, 77(7):941–981, 1989.

L. Cohen. What is a multicomponent signal? In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 5, pages 113–116. IEEE, 1992.

L. Cohen. *Time-frequency analysis*, volume 778. Prentice hall, 1995.

P. Cosi, D. Falavigna, and M. Omologo. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In *EuroSpeech*, 1991.

L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos. The DIRHA simulated corpus. *Proc. of International Conference on Language Resources and Evaluation*, 5, may 2014.

K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

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, 1980. ISSN 0096-3518.

A. De Cheveigne. Pitch perception models. In *Pitch*, pages 169–233. Springer, 2005.

A. De Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

J. de Ville et al. Théorie et applications de la notion de signal analytique. *Cables et transmission*, 2(1):61–74, 1948.

L. Deng and D. O'Shaughnessy. *Speech processing: a dynamic and optimization-oriented approach*. CRC Press, 2003.

D. Deutsch. Speaking in tones. *Scientific American Mind*, 21(3):36–43, 2010.

D. D. Dirks, D. E. Morgan, and J. R. Dubno. A procedure for quantifying the effects of noise on speech recognition. *Journal of Speech and Hearing Disorders*, 47(2):114–123, 1982.

M. Dolson. The phase vocoder: A tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

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.

K. Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution fft. In *9th International Conference on Digital Audio Effects*, DAFx, pages 247–252, 2006.

K. Dressler. Audio melody extraction for mirex 2009. In *MIREX 2009 - Music Information Retrieval Evaluation eXchange, MIREX Melody Extraction*, 2009.

K. Dressler. Pitch estimation by the pair-wise evaluation of spectral peaks. In *42nd International Conference of Audio Engineering Society*, AES, 2011.

J. Droppo and A. Acero. Environmental robustness. In *Springer Handbook of Speech Processing*, pages 653–680. Springer, 2008.

J. R. Dubno, D. D. Dirks, and D. E. Morgan. Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, 76(1): 87–96, 1984.

J. L. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP, pages 105–108, 2009.

A. Dziewonski, S. Bloch, and M. Landisman. A technique for the analysis of transient seismic signals. *Bulletin of the seismological Society of America*, 59(1):427–444, 1969.

J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor. Estimation of room acoustic parameters: The ACE challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1681–1693, 2016.

Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

G. Evermann and P. C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Speech Transcription Workshop*, volume 27, 2000.

G. Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108-th Audio Engineering Society Convention*, AES, 2000.

J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speeck Recognition and Understanding*, ASRU, pages 347–354, 1997.

K. Fitz. Time-frequency analysis - matlab toolbox. `http://www.cerlsoundgroup.org/Kelly/timefrequency.html/`, 2007. [Online; accessed 2017].

K. Fitz and S. A. Fulop. A unified theory of time-frequency reassignment. *arXiv preprint arXiv:0903.3080*, 2009.

K. Fitz and L. Haken. On the use of time-frequency reassignment in additive sound modeling. *Journal of the Audio Engineering Society*, 50(11):879–893, 2002.

K. Fitz, L. Haken, and P. Christensen. Transient preservation under transformation in an additive sound model. In *In Proc. International Computer Music Conference*, 2000.

J. L. Flanagan, C. H. Coker, L. R Rabiner, R. W. Schafer, and N. Umeda. Synthetic voices for computers. *IEEE spectrum*, 7(10):22–45, 1970.

J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko. Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78(5):1508–1518, 1985.

J. L. Flanagan, J. B. Allen, and M. A. Hasegawa-Johnson. *Speech Analysis Synthesis and Perception.* Springer-Verlag, third edition, 2008.

P. Flandrin. *Time-frequency/time-scale analysis*, volume 10. Academic press, 1998.

P. Flandrin, F. Auger, E. Chassande-Mottin, et al. Time-frequency reassignment from principles to algorithms. *Applications in time-frequency signal processing*, 5:179–203, 2002.

P. Flandrin, M. Amin, S. McLaughlin, and B. Torrésani. Time–frequency analysis and applications. *IEEE Signal Processing Magazine*, 30(2013):19–150, 2013.

G. B. Folland and A. Sitaram. The uncertainty principle: a mathematical survey. *Journal of Fourier analysis and applications*, 3(3):207–238, 1997.

J. W. Forgie and C. D. Forgie. Results obtained from a vowel recognition computer program. *The Journal of the Acoustical Society of America*, 31(11):1480–1489, 1959.

D Friedman. Instantaneous-frequency distribution vs. time: An interpretation of the phase structure of speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10, pages 1121–1124. IEEE, 1985.

D. B. Fry. Theoretical aspects of mechanical speech recognition. *Journal of the British Institution of Radio Engineers*, 19(4):211–218, 1959.

S. A. Fulop. *Speech Spectrum Analysis.* Springer Berlin Heidelberg, 2011.

S. A. Fulop and S. F. Disner. The reassigned spectrogram as a tool for voice identification. In *International Congress of Phonetic Sciences*, pages 1853–1856, 2007.

S. A. Fulop and K. Fitz. A spectrogram for the twenty-first century. *Journal of the Acoustical Society of America*, 2(3):26–33, 2006.

S. A. Fulop and K. Fitz. Separation of components from impulses in reassigned spectrograms. *Journal of the Acoustical Society of America*, 121(3):1510–1518, 2007.

S. A. Fulop and Y. Kim. Speaker identification made easy with pruned reassigned spectrograms. In *Proceedings of Meetings on Acoustics*, volume 19. Acoustical Society of America, 2013.

S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272, 1981.

S. Furui. *Advances in Speech Signal Processing.* Electrical and Computer Engineering. Taylor & Francis, 1991.

D. Gabor. Theory of communication. *IEE*, 93:429–457, 1946.

T.J. Gardner and M.O. Magnasco. Instantaneous frequency decomposition: An application to spectrally sparse sounds with fast frequency modulations. *The Journal of the Acoustical Society of America*, 117(5):2896–2903, 2005.

J. Garofolo, D. Graff, D. Paul, and D. Pallett. Continous speech recognition (CSR-I) Wall Street Journal (WSJ0) News Complete. *LDC93S6A. DVD. Linguistic Data Consortium, Philadelphia*, 1993a.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N*, 93:27403, 1993b.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, 1993c.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.

Y. Gong. Speech recognition in noisy environments: A survey. *Speech communication*, 16 (3):261–291, 1995.

M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43 (4):311 – 329, 2004.

M. Goto. PreFEst: A predominant-f0 estimation method for polyphonic musical audio signals. In *2nd Music Information Retrieval Evaluation eXchange*, MIREX, 2005.

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, 2010.

A. Gray and J. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, Oct 1976.

C. Guerrero. *Information Fusion Approaches for Distant Speech Recognition in a Multi-microphone Setting*. PhD thesis, University of Trento, 2016.

C. Guerrero, G. Tryfou, and M. Omologo. Channel selection for distant speech recognition - exploiting cepstral distance. In *Annual Conference of the International Speech Communication Association*, INTERSPEECH, 2016.

P. Guillemain and R. Kronland-Martinet. Characterization of acoustic signals through continuous linear time-frequency representations. *Proceedings of the IEEE*, 84(4):561–585, 1996.

S. W. Hainsworth. *Techniques for the automated analysis of musical audio*. PhD thesis, University of Cambridge, 2003.

S. W. Hainsworth and P. J. Wolfe. Time-frequency reassignment for music analysis. In *International Computer Music Conference*, 2001.

S. W. Hainsworth, M. D. Macleod, and P. J. Wolfe. Analysis of reassigned spectrograms for musical transcription. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 23–26. IEEE, 2001.

S. Handel. Timbre perception and auditory object identification. *Hearing*, 2:425–461, 1995.

J. H. Hansen and B. L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *International Conference on Spoken Language Processing*, volume 7 of *ICSLP*, pages 2819–2822, 1998.

M. Harper. The automatic speech recognition in reverberant environments (ASpIRE) Challenge. In *Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2015.

W. M. Hartmann. Pitch, periodicity, and auditory organization. *The Journal of the Acoustical Society of America*, 100(6):3491–3502, 1996.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87:1738, 1990.

H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE transactions on speech and audio processing*, 2(4):578–589, 1994.

Wolfgang Hess. *Pitch determination of speech signals: algorithms and devices.* Springer-Verlang, 1983.

F. B. Hildebrand. *Advanced calculus for engineers.* Prentice-Hall, 1949.

I. Himawan, P. Motlicek, S. Sridharan, D. Dean, and D. Tjondronegoro. Channel selection in the short-time modulation domain for distant speech recognition. In *Proceedings of Interspeech*, number EPFL-CONF-209075, 2015.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara. Revising Perceptual Linear Prediction (PLP). In *12th Annual Conference of the International Speech Communication Association*, pages 2997–3000, 2005.

J. P. Hosom. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4):352–368, 2009.

C. L. Hsu and J. S. R. 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, 2010.

Y. Hu and P. C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, 49(7):588–601, 2007.

Y. Hu and P. C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):229–238, 2008.

X. Huang, A. Acero, H. W. Hon, et al. *Spoken language processing: A guide to theory, algorithm, and system development*, volume 15. Prentice Hall PTR New Jersey, 2001.

Y. Huang, J. Benesty, and J. Chen. *Time delay estimation and source localization*, pages 1043–1064. Springer, 2008.

Y. A. Huang and J. Benesty. *Audio signal processing for next-generation multimedia communication systems*. Springer Science & Business Media, 2007.

F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Electronic Communication Japan*, 53-A:36–43, 1970.

M. Ito and M. Yano. Sinusoidal modeling for nonstationary voiced speech based on a local vector transform. *The Journal of the Acoustical Society of America*, 121(3):1717–1727, 2007.

F. Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

F. Jelinek, L. Bahl, and R. Mercer. Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, 21(3): 250–256, 1975.

H. Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

S. Jo, S. Joo, and C. D. Yoo. Melody pitch estimation based on range estimation and candidate extraction using harmonic structure model. In *Annual Conference International Speech Communication Association*, INTERSPEECH, pages 2902–2905. ISCA, 2010.

T. Jo and M. Koyasu. Measurement of reverberation time based on the direct-reverberant sound energy ratio in steady state. In *INTER-NOISE and NOISE-CON Congress and*

Conference Proceedings, number 2, pages 579–582. Institute of Noise Control Engineering, 1975.

M. Joos. Acoustic phonetics. *Language*, 24(2):5–136, 1948.

B. H. Juang, S. Levinson, and M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of markov chains (corresp.). *IEEE Transactions on Information Theory*, 32(2):307–309, 1986.

H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In *EuroSpeech*, volume 99, pages 2781–2784, 1999.

B. Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, 1986.

F. Keiler and S. Marchand. Survey on extraction of sinusoids in stationary sounds. In *Digital Audio Effects (DAFx) Conference*, pages 51–58, 2002.

P. Kenny. A small footprint i-vector extractor. In *Odyssey*, volume 2012, pages 1–6, 2012.

M. Khadkevich. *Music signal processing for automatic extraction of harmonic and rhythmic information.* PhD thesis, University of Trento, 2011.

M Khadkevich and M. Omologo. Reassigned spectrum-based feature extraction for gmm-based automatic chord recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2013(1):1–12, 2013.

Y. J. Kim and A. Conkie. Automatic segmentation combining an hmm-based approach and spectral boundary correction. In *INTERSPEECH*, 2002.

K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, WASPAA, pages 1–4, 2013.

A. Klapuri and M. Davy. *Signal processing methods for music transcription.* Springer, 2006.

A. P. Klapuri, A. J. Eronen, and J. T. Astola. Analysis of the meter of acoustic musical signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):342–355, 2006.

K. Kodera, C. De Villedary, and R. Gendrin. A new method for the numerical analysis of non-stationary signals. *Physics of the Earth and Planetary Interiors*, 12(2):142–150, 1976.

K. Kodera, R. Gendrin, and C. de Villedary. Analysis of time-varying signals with small BT values. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):64–76, 1978.

W. Koenig, H. K. Dunn, and L. Y. Lacy. The sound spectrograph. *The Journal of the Acoustical Society of America*, 18(1):19–49, 1946.

K. Kumatani, J. McDonough, J. F. Lehman, and B. Raj. Channel selection based on multichannel cross-correlation coefficients for distant speech recognition. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays*, HSCMA, pages 1–6, 2011.

H. Kuttruff. *Acoustics: An Introduction*. CRC Press, 2007.

H. Kuttruff. *Room acoustics*. CRC Press, fifth edition, 2009.

J. E. Lane. Pitch detection using a tunable iir filter. *Computer Music Journal*, 14(3): 46–59, 1990.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

C J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

H. Leung and V. Zue. A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, volume 9, pages 73–76. IEEE, 1984.

J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12):1586–1604, 1979.

R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman. The precedence effect. *The Journal of the Acoustical Society of America*, 106(4):1633–1654, 1999.

H. Y. Lo and H. M. Wang. Phonetic boundary refinement using support vector machine. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–933. IEEE, 2007.

P. C. Loizou. *Speech enhancement: theory and practice*. CRC press, 2013.

J. Makhoul. Spectral analysis of speech by linear prediction. *IEEE Transactions on Audio and Electroacoustics*, 21(3):140–148, 1973.

L. Malfait, J. Berger, and M. Kastner. P. 563&8212; The ITU-T Standard for Single-Ended Speech Quality Assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, 2006.

L. L. Mangu. *Finding consensus in speech recognition*. PhD thesis, 2000. AAI9964156.

J. D. Markel and A. Jr Gray. *Linear prediction of speech*, volume 12. Springer Science & Business Media, 2013.

M. Marolt. On finding melodic lines in audio recordings. In *Proceedings of the 7th International Conference on Digital Audio Effects*, DAFx, pages 199–204, 2004.

T. B. Martin, A. L. Nelson, and H. J. Zadell. Speech recognition by feature-abstraction techniques. Technical report, DTIC Document, 1964.

M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer. Hidden Markov Model training with contaminated speech material for distant-talking speech recognition. *Computer Speech & Language*, 16(2):205–223, 2002.

R. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976.

G. F. Meyer, F. Plante, and F. Berthommier. Segregation of concurrent speech with the reassigned spectrum. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 2, pages 1203–1206. IEEE, 1997.

B. Milner. A comparison of front-end configurations for robust speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–797, 2002.

J. A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, pages 32–38, 1977.

C. Myers, L. Rabiner, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(6):623–635, 1980.

T. Nakatani and T. Irino. Robust and accurate fundamental frequency estimation based on dominant harmonic components. *The Journal of the Acoustical Society of America*, 116(6):3690–3700, 2004.

P. A. Naylor and N. D. Gaubitch. *Speech dereverberation.* Springer Science & Business Media, 2010.

D. J. Nelson. Cross-spectral methods for processing speech. *The Journal of the Acoustical Society of America*, 110(5):2575–2592, 2001.

D. J. Nelson. Instantaneous higher order phase derivatives. *Digital Signal Processing*, 12 (2–3):416 – 428, 2002.

A. M. Noll. Short-time spectrum and "cepstrum" techniques for vocal-pitch detection. *The Journal of the Acoustical Society of America*, 36(2):296–302, 1964.

Y. Obuchi. Multiple-microphone robust speech recognition using decoder-based channel selection. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, ITRW, 2004.

Y. Obuchi. Noise robust speech recognition using delta-cepstrum normalization and channel selection. *Electronics and Communications in Japan (Part II: Electronics)*, 89(7): 9–20, 2006.

J. E. Odegard, R. G. Baraniuk, and K. L. Oehler. Instantaneous frequency estimation using the reassignment method. In *SEG Technical Program Expanded Abstracts 1997*, pages 1941–1944. Society of Exploration Geophysicists, 1997.

H. F. Olson and H. Belar. Phonetic typewriter. *The Journal of the Acoustical Society of America*, 28(6):1072–1081, 1956.

Alan Oppenheim and Ronald Schafer. *Discrete-time signal processing.* Prentice Hall Inc., 1989.

R. F. Orlikoff and J. C. Kahane. Structure and function of the larynx. *Principles of Experimental Phonetics. Mosby, St. Louis, Baltimore/etc*, pages 112–181, 1996.

R. S. Orr. Dimensionality of signal sets. In *San Diego, CA*, pages 435–446. International Society for Optics and Photonics, 1991.

R. L. Ott and M. T. Longnecker. *An introduction to statistical methods and data analysis.* Cengage Learning, 2008.

C. H. Page. Instantaneous power spectra. *Journal of Applied Physics*, 23(1):103–106, 1952.

R. P. Paiva, T. Mendes, and A. Cardoso. Melody detection in polyphonic musical signals: Exploiting perceptual rules, note salience, and melodic smoothness. *Computer Music Journal*, 30(4):80–98, 2006.

M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra. A survey of convolutive blind source separation methods. *Multichannel Speech Processing Handbook*, 2007.

G. Peeters and X. Rodet. Non-stationary analysis/synthesis using spectrum peak shape distortion, phase and reassignement. In *Proceedings of the International Congress on Signal Processing Applications Technology-ICSPAT*, 1999.

P. M. Peterson. Simulating the response of multiple microphones to a single acoustic source in a reverberant room. *The Journal of the Acoustical Society of America*, 80(5): 1527–1529, 1986.

M. Piszczalski. A computational model of music transcription. 1986.

C. J. Plack and R. P. Carlyon. Loudness perception and intensity coding. *Hearing*, pages 123–160, 1995.

F. Plante and W. A. Ainsworth. Formant tracking using reassigned spectrum. In *Fourth European Conference on Speech Communication and Technology*, 1995.

F. Plante, G. Meyer, and W. Ainsworth. Improvement of speech spectrogram accuracy by the method of reassignment. *IEEE Transactions on Speech and Audio Processing*, 6 (3):282–287, 1998.

G. E. Poliner and D. P. W. Ellis. A classification approach to melody trascription. In *6th International Society for Music Information Retrieval Conference*, ISMIR, pages 161–166, 2005.

G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and Beesuan O. Melody transcription from music audio: Approaches and evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(4):1247–1256, 2007.

R. K. Potter et al. Visible speech. 1947.

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P.and others Schwarz. The Kaldi speech recognition toolkit. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, ASRU, 2011.

S. Quackenbush, T. Barnwell, and M. Clements. *Objective measures of speech quality.* Englewood Cliffs, NJ, Prentice-Hall. Signal Processing Series, 1988.

L. Rabiner and B. J. Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-015157-2.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, Englewood Cliffs, NJ, 1978.

L. R. Rabiner and R. W. Schafer. *Theory and application of Digital Speech Processing.* PEARSON, 2011.

D. V. Rabinkin, R. J. Renomeron, J. C. French, and J. L. Flanagan. Estimation of wavefront arrival delay using the cross-power spectrumphase technique. In *132nd Meeting of the Acoustical Society of America*, volume 100, page 2697. Citeseer, 1996.

R. Rajan and H. A. Murthy. Group delay based melody monopitch extraction from music. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 186–190. IEEE, 2013.

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, 2010.

R. Rasch and R. Plomp. The perception of musical tones. *The psychology of music*, 2: 89–112, 1999.

R. Ratnam, D. L. Jones, B. C. Wheeler, William D. O'Brien Jr., C. R. Lansing, and A. S. Feng. Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892, 2003.

M. Ravanelli and M. Omologo. On the selection of the impulse responses for distant-speech recognition based on contaminated speech training. In *INTERSPEECH*, pages 1028–1032, 2014.

M. Ravanelli and M. Omologo. Contaminated speech training methods for robust dnn-hmm distant speech recognition. In *16th Annual Conference of the International Speech Communication Association*, volume 1 of *INTERSPEECH*, 2015.

M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo. Impulse response estimation for robust speech recognition in a reverberant environment. In *20th European Signal Processing Conference*, EUSIPCO, pages 1668–1672, 2012.

M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo. The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding*, ASRU, pages 275–282, 2015.

ITUT Recommendation. Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. *ITU-T Recommendation*, page 862, 2001.

D. R. Reddy. Approach to computer speech recognition by direct analysis of the speech wave. *The Journal of the Acoustical Society of America*, 40(5):1273–1273, 1966.

A. Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, pages 749–752, 2001.

C. Roads. *The computer music tutorial.* MIT press, 1996.

D. Robinson. A new determination of the equal-loudness contours. *Audio, IRE Transactions on*, AU-6(1):6–13, January 1958.

S. Rossignol, X. Rodet, J. Soumagne, J. L. Colette, and P. Depalle. Feature extraction and temporal segmentation of acoustic signals. In *ICMC: International Computer Music Conference*, pages 1–1, 1998.

M. Ryynänen. *Automatic Transcription of Pitch Content in Music and Selected Applications.* PhD thesis, Tampere University of Technology, 2008.

M Ryynänen and A. Klapuri. Transcription of the singing melody in polyphonic music. In *7th International Society for Music Information Retrieval Conference*, ISMIR, 2006.

J. Salamon. *Melody Extraction from Polyphonic Music Signals.* PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain, 2013.

J. Salamon and E. Gómez. A chroma-based salience function for melody and bass line estimation from music audio signals. In *6th Sound and Music Computing Conference*, pages 331–336. Citeseer, 2009.

J. Salamon and E. Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

J. Salamon and J. Urbano. Current challenges in the evaluation of predominant melody extraction algorithms. In *ISMIR*, volume 12, pages 289–294, 2012.

J. Salamon, E. Gómez, and J Bonada. Sinusoid extraction and salience function design for predominant melody estimation. In *14th International Conference on Digital Audio Effects*, DAFx, pages 73–80, 2011.

T. H. Sang and W. J. Williams. Renyi information and signal-dependent optimal kernel design. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 2, pages 997–1000. IEEE, 1995.

M. Schedl, E. Gómez, J. Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014.

M. R. Schröder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412, 1965.

M. R. Schröder. Diffuse sound reflection by maximum- length sequences. *The Journal of the Acoustical Society of America*, 57(1):149–150, 1975.

B. Secrest and G. Doddington. An integrated pitch tracking algorithm for speech systems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 8, pages 1352–1355, 1983.

C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura. Speech recognition based on space diversity using distributed multi-microphone. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, pages 1747–1750, 2000.

K. N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, 111(4):1872–1891, 2002.

S. Stevens and E. Volkmann, J .and Newman. The mel scale equates the magnitude of perceived differences in pitch at different frequencies. *J. Acoust. Soc. Am*, 8(3):185–190, 1937.

J. Sundberg et al. *The acoustics of the singing voice.* Scientific American, 1977.

C. Sutton, E. Vincent, M. D. Plumbley, and J. P. Bello. Transcription of vocal melodies using voice characteristics and algorithm fusion. In *Music Information Retrieval Evaluation eXchange*, 2006.

J. Suzuki and K. Nakata. Recognition of japanese vowels-preliminary to the recognition of speech. *Journal of the Radio Research Laboratory*, 37(8):193–212, 1961.

H. Tachibana, T. Ono, N. Ono, and S. Sagayama. Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source. In *Acoustics speech and signal processing (icassp), 2010 ieee international conference on*, pages 425–428, 2010.

Y. Tachioka, S. Watanabe, J. Le Roux, and J. R. Hershey. Discriminative methods for noise robust speech recognition: A chime challenge benchmark. In *Proceedings of the CHiME 2013 International Workshop on Machine Listening in Multisource Environments*, 2013.

P. Taylor. *Text-to-speech synthesis.* Cambridge university press, 2009.

E. Terhardt. Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, 55(5):1061–1069, 1974.

E. Terhardt, G. Stoll, and M. Seewann. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *Journal of the Acoustical Society of America*, 71 (3):679–688, 1982.

D. T. Toledano. Neural network boundary refining for automatic speech segmentation. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 6, pages 3438–3441. IEEE, 2000.

J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere. A study of complexity and quality of speech waveform coders. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP, pages 586–590, 1978.

B. Van der Pol. The fundamental principles of frequency modulation. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93 (23):153–158, 1946.

B. D. Van Veen and K. M. Buckley. Beamforming: A versatile approach to spatial filtering. *IEEE ASSP Magazine*, 5(2):4–24, 1988.

K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey. Sequence-discriminative training of deep neural networks. In *INTERSPEECH*, pages 2345–2349, 2013.

T. K. Vintsyuk. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, 4(1):52–57, 1968.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3):328–339, 1989.

S. Wang, A. Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on selected areas in communications*, 10(5):819–829, 1992.

D. B. Ward, R. A. Kennedy, and R. C. Williamson. Theory and design of broadband sensor arrays with frequency invariant far-field beam patterns. *The Journal of the Acoustical Society of America*, 97(2):1023–1034, 1995.

L. Welling, S. Kanthak, and H. Ney. Improved methods for vocal tract normalization. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 761–764. IEEE, 1999.

M. B. Wesenick and A. Kipp. Estimating the quality of phonetic transcriptions and segmentations of speech signals. In *4th International Conference on Spoken Language*, pages 129–132, 1996.

F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9:288–298, 2001.

E. Wigner. On the quantum correction for thermodynamic equilibrium. *Physical review*, 40(5):749, 1932.

W. J. Williams, M. L. Brown, and A. O Hero III. Uncertainty, information, and time-frequency distributions. In *San Diego,'91, San Diego, CA*, pages 144–156. International Society for Optics and Photonics, 1991.

M. Wolf. *Channel selection and reverberation-robust automatic speech recognition.* PhD thesis, Universitat Politècnica de Catalunya, 2013.

M. Wolf and C. Nadeu. Towards microphone selection based on room impulse response energy-related measures. In *Workshop on Speech and Language Technologies for Iberian Languages*, pages 61–64, 2009.

M. Wolf and C. Nadeu. On the potential of channel selection for recognition of reverberated speech with multiple microphones. In *INSTERSPEECH*, pages 80–83, 2010.

M. Wolf and C. Nadeu. Channel selection measures for multi-microphone speech recognition. *Speech Communication*, 57:170–180, 2014.

M. Wölfel. Channel selection by class separability measures for automatic transcriptions on distant microphones. In *Annual Conference of the International Speech Communication Association*, INTERSPEECH, pages 582–585, 2007.

M. Wölfel and J. McDonough. *Distant Speech Recognition.* Wiley, 2009.

T. C. Yeh, M. J. Wu, J.R. Jang, W. L. Chang, and Liao I. B. A hybrid approach to singing pitch extraction based on trend estimation and hidden markov models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP, pages 457–460, 2012.

T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *Signal Processing Magazine, IEEE*, 29 (6):114–126, 2012.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, D. Odell, J.and Ollason, D. Povey, et al. The HTK book.

D. Yu and L. Deng. *Automatic speech recognition: A deep learning approach.* Springer, 2014.

J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang. Automatic phonetic segmentation using boundary models. In *14th Annual Conference of the International Speech Communication Association*, pages 2306–2310, 2013.

P. Zahorik. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5):2110–2117, 2002.

E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2):248–248, 1961.