BEN-GURION UNIVERSITY OF THE NEGEV
FACULTY OF ENGINEERING SCIENCES
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Research Proposal for the Ph.D. degree

Robust Speaker Recognition for Reverberant Speech

By: **Noam Shabtai**
Supervised by:

Dr. Boaz Rafaely
Dr. Yaniv Zigel

August 13, 2007

Abstract

This report presents a *research proposal* on the subject of robust speaker recognition for reverberant speech. Speaker recognition is the process of recognizing who is speaking on the basis of individual features that were extracted from a speech signal. Speaker recognition can be used in voice dialing, banking by telephone, telephone shopping, etc. Speaker recognition performance degrades due to *reverberation* that is often present in such applications, which has an effect on the feature extraction.

Feature extraction of the speech signal is performed by the Mel-Cepstrum approach which yields the MFCC feature vectors. Statistical distribution of the feature vectors is modeled by the GMM approach. Robustness stands for effective feature extraction in a reverberant environment, and reducing the degradation in performance of speaker recognition systems due to reverberation.

This proposal describes the approach and methods regarding room reverberation and studying its effect on the feature vectors. These methods include an experimental setup in which detailed future experiments are presented. In these experiments we will try to isolate various acoustic parameters of the room, and examine their effect on the feature vectors. The acoustic parameters under examination are reverberation time, absorption of the walls, distance between speaker and microphone, and room dimensions.

Initial simulations were used to measure the effect of reverberation on the feature vectors in two ways. The first, by comparing the MFCC vectors of clean and reverberant matching speech frames. The second, by calculating the total distance between the distribution of the MFCC vectors of the clean speech, to the distribution of the MFCC vectors of the reverberant speech. Initial results are presented. The conclusion for both ways was that the distortion made to the feature vectors increases with the increase of reverberation time.

Contents

Part I

Research Proposal

Chapter 1

Introduction and Motivation

Speaker recognition is the process of recognizing who is speaking on the basis of individual information included in a speech signal [1]. *Speaker recognition* is divided into two topics namely, *speaker verification* (SVR) and *speaker identification* (SID) [2] [1]. SVR is considered to be a simpler operation, which is used to verify whether the speaker is an hypothesized speaker or not. SID uses the SVR results to identify which entity among a list the speaker represents.

Speaker recognition systems extract feature vectors in order to parameterize speakers individual information from the speech signal. A popular feature extraction method for speaker recognition systems is the *Mel frequency cepstral coefficients* (MFCC) on its various modes [3]. Speakers are recognized according to a pattern matching of the statistical distribution of their feature vectors in the feature space, with known statistical distribution patterns of feature vectors that are related to other speakers. The statistical distribution of the feature vectors is modeled by statistical modeling methods. *Gaussian mixture models* (GMM) has become a dominant approach for statistical modeling of speech feature vectors [4].

Speaker recognition has a lot of use in telephone based applications [1]. However, often is the case that *reverberation* is present in such applications, due to the surrounding room environment [5]. The presence of reverberation adds distortion to the feature vectors, which cause a degradation in the performance of speaker recognition systems [6]. Reverberation is considered to be a form of *convolutive noise*, in a sense that a clean input speech signal is filtered with a transfer function of a room to form a reverberant output

[1] In [2] the terms ASV and ASI replace SVR and SID respectively, to indicate *automatic* recognition.

speech signal [7]. In the cases where noise is considered to be a background additive noise, *speech enhancement* methods which suppress the noise such as spectral subtraction [8, 9], Wiener filtering [10],and microphone arrays [11], had been previously developed. The problem becomes more complicated, however, in cases where reverberation is present due to its convolutive character.

Methods for feature extraction of speech signals with convolutive noise in general usually use feature normalization techniques such as the *cepstral mean subtraction* (CMS) [12, 1] and the variance normalization methods [13]. Pattern matching of statistical distribution of the feature vectors in the case of convolutive noise is performed by using post processing techniques such as the Znorm, Hnorm, Tnorm, etc. verification score normalizations [12, 1, 14]. All these methods whether employ normalization to the feature vectors or to the verification scoring, are being used in order to reduce the channel effect due to hand-set type [15], linear channel distortion [16], and microphone type [17]. Although these might also include convolutive noise effects, they are typically characterized by a *short-duration* impulse response. For that reason, they might not be as useful to reduce the effect of reverberation with long duration, which is often the case in room acoustics. In fact, the use of these methods to reduce the effect of reverberation in speaker recognition has not been presented yet is the literature.

Reducing the effect of convolutive noise that is characterized by a *long-duration* impulse response can be achieved by training the speaker recognition system with the feature vectors that were extracted from the reverberant speech data [7, 18, 19]. Another set of solutions may use *microphone arrays* [20]. Microphone arrays may be used to estimate the room impulse response and then perform *de-reverberation* by applying the inverse response. This can be done by the *matched-filter* approach [21], or by exploiting characteristics of the speech signal [22], etc. Microphone arrays may also be used to perform *speech enhancement* by *beamforming* in which the direction from which the direct sound is approaching is emphasized. However, traditional approaches to beamforming fail to compensate for the negative effect of reverberation, and therefore recent work is done to adapt beamforming methods for speech recognition applications [5]. Moreover, microphone arrays result in increased cost and processing requirements, and many telephony applications use not an array but a single microphone.

In spite of what have been stated above, there appears to be a lack of theoretical understanding of the effect of reverberation on the feature vectors and on their statistical

distribution in the feature space. The solutions which were mentioned above, suggest only refinements to the usual feature extraction and pattern matching routines. Hence, this research will be focused on modeling the effect of reverberation on the feature vectors, and on developing new ways of feature extraction for speaker recognition in reverberant environments.

The organization of this proposal is as follows. Research objective and expected innovations of the research are discussed in Chapter 2. Chapter 3 presents the approach and methods to be used in order to achieve that objective, including a detailed experimental setup. Chapter 4 summarizes in a research plan the steps that had been taken so far, and the expected steps to be taken.

Chapter 2

Research Objective and Expected Innovations

The objective of this research is to develop a speaker recognition method that is robust to reverberation. The research is expected to examine the effect of reverberation on the feature vectors and on their statistical distribution in the feature space. The study of this effect will be used in order to propose *new* or *modified* feature extraction technique, that will form feature vectors whose distortion due to reverberation is reduced. The new feature extraction technique is then to be implemented in a speaker recognition system, to form robust speaker recognition for reverberant speech.

The research is expected to yield several innovations:

1. A better understanding of the effect of reverberation on speech feature vectors.
2. New or modified feature extraction technique.
3. Experimental efficiency verification of the new feature extraction technique.

Chapter 3

Research Approach and Methods

This chapter discusses the proposed research approach and methods to reduce the effect of reverberation on speaker recognition. Section 3.1 presents the speaker recognition system that is used in this research. Section 3.2 presents the general approach for tackling the problem of deterioration in performance of speaker recognition systems due to the effect of reverberation on the feature vectors. The rest of the chapter introduces methods and experiments to be conducted in order to measure, analyze, and suppress this kind of effect.

3.1 The Speaker Recognition System

Figure 3.1 shows a block diagram of the speaker recognition system. In general, speaker recognition divides into two phases. The first, represented on top of Figure 3.1, is the *training* phase. During the training phase speaker models are generated. The second, represented on bottom of Figure 3.1, is the *testing* phase, during which speaker models that were generated in the training phase are being used in pattern matching for speaker verification. In this work we typically assume that the input speech signal in the testing phase is convoluted with an impulse response of a room which results in a reverberant speech signal [1].

A popular feature extraction method for speaker recognition systems is the *Mel frequency cepstral coefficients* (MFCC) on its various modes [3]. Therefore, a major part of the research is devoted to studying the effect of reverberation on the MFCC feature

[1] Speaker recognition systems usually train on clean (non-reverberant) speech data.

vectors of a speech signal.

After feature extraction, pattern recognition methods are applied in order to perform statistical modeling of the distribution of the feature vectors. In the training phase the modeling of that distribution is used for speaker model estimation and generation of speaker models. During the testing phase, a decision is made to the feature vectors of a speaker's speech signal, regarding the match of their statistical distribution pattern with each one of the previously generated speaker models. Since *Gaussian mixture models* (GMM) has become a dominant approach for statistical modeling of speech feature vectors distribution [4], this will be the statistical modeling method of choice in this research as well.

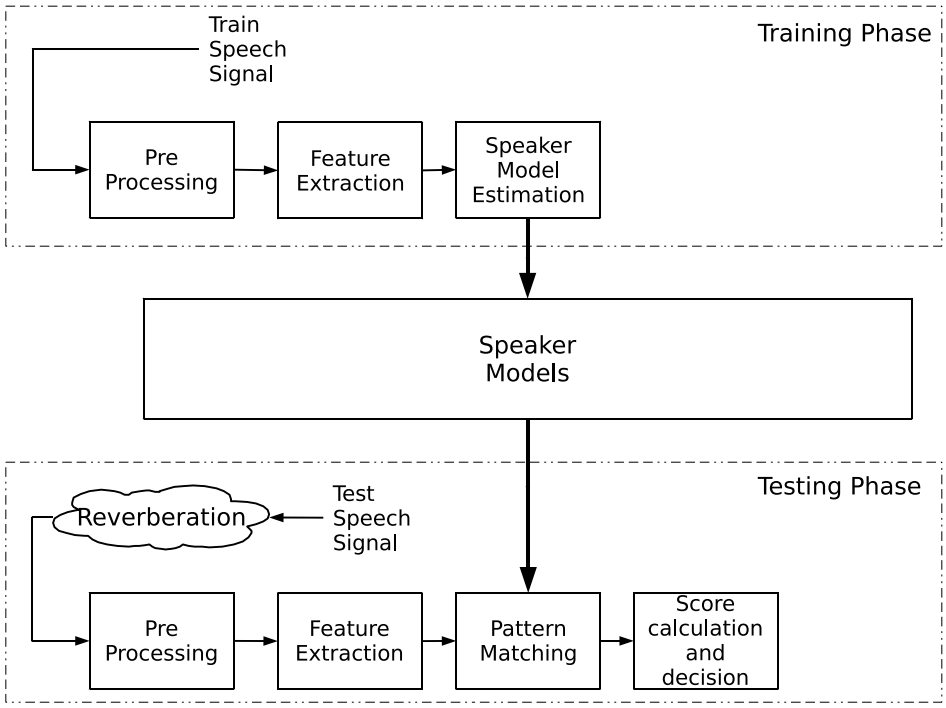


Figure 3.1: Block diagram of the speaker recognition system.

In this research speech is taken from the 1999 *national institute of standards and technology – speaker recognition evaluation* (NIST-SRE) data base [23]. Reverberation is added by convoluting the speech signal with either simulated or measured room impulse response. Speaker recognition algorithms are performed either by MATLAB or by the SID simulator that is supplied by *NICE-systems Ltd.* as part of a Magneton research project. When using MATLAB, feature extraction is performed with the *voicebox* toolbox [24], and statistical modeling is performed with the *netlab* toolbox [25]. Output analysis

is performed in MATLAB.

3.2 General Research Approach

This research combines theoretical analysis and simulations. Acoustic models are being used to simulate reverberation, and to study its effect on different aspects of the speech signal. This effect is measured at different stages of the speaker recognition system using spectrograms, feature vector distortion measures, and distribution distance measures. Reliable measures are required in order to monitor the effect of reverberation as a function of the reverberation parameters. Thus, part of the research is dedicated to developing such measure tools. Likewise, simulation tools of reverberation and speaker recognition are required, and need to be developed. Such an approach allows us to study and model the effect of the acoustic channel on the speech signal under different conditions. Hopefully, the results of this study will facilitate the development of feature extraction methods for robust speaker recognition.

3.3 Room Reverberation

Reverberation is modeled by the impulse response of a room [2]. Reverberant speech is modeled by the output of the convolution of that impulse response with the speech signal. The effect of reverberation on the speech signal is to be studied both in the time and the frequency domains. In the time domain, smearing of adjacent frames is expected to be seen if the impulse response is long. In the frequency domain, the effect might be seen by resonances in low frequencies, and by some spectral coloring in high frequencies.

There are three main ways to model enclosed sound field, namely,

1. The *modal model* (see Appendix B.1).

2. The *diffuse-field model* (see Appendix B.2).

3. The *image method* (see Appendix B.3).

[2]Here we assume *linear time invariant* (LTI) system, but that is not always the case. For example, a movement of a speaker inside the room causes the response to be time dependent.

In general, experiments are to be designed by using the different models and different room parameters. These experiments will be used in order to study the effect of reverberation on different aspects of the speech signal.

The modal model provides information about resonance frequencies. It is typically used at low frequencies. This model could be used to study the effect of resonances or standing waves on the speech features, by controlling the resonant frequencies, their density, and damping behavior.

The diffuse-field model represents a statistical model with damping, and ignores spatial and frequency detailed effects. The damping causes a decay which can be used to describe the 'tail' of the room impulse response in the time domain. In this 'tail', many sources exist and arranged in a dense manner, generating a type of noisy response with a low amplitude. This model will be used in order to give theoretical support on acoustic quantities such as reverberation time, modal density, and radius of reverberation (see Appendix B.2).

The image method is a general model which can be used at a wide frequency range. This method gives good description of the first reflections from the walls, which are the dominant ones. This model will be used in order to control these reflections by changing their amplitude and density and by that studying their effect on the speech features. The image method is currently being used to simulate impulse response of a shoe-box-type room. In order to have the response include the effect of the transducers, it is convoluted with an *infinite impulse response* (IIR) *Band Pass Filter* (BPF) [3].

In order to yield the simulated reverberant speech, the clean speech signal is convoluted with the impulse response of the room. Other than the simulated room impulse response, measured responses are taken, for better representation of realistic reverberation channels.

3.4 Reverberant Speech in the STFT Domain

The effect of reverberation on the STFT of a speech signal can be visualized by representing the *spectrogram* of a reverberant signal, and comparing it to the one of the clean signal. The expected result would be that in the reverberant spectrogram there will be an evidence for smearing of voiced phonemes onto later frames. That would be since rever-

[3] *High pass filter* (HPF) alone can also serve for that cause.

beration causes an averaging of adjacent time frames. Hence, the greater reverberation time, the wider smearing is expected to be.

The difference between the clean and reverberant STFT is believed to be diminished if using long frame windows. However, windows longer than 30 ms make the speech signal inside the frame to be non-stationary [26]. In that case, a feature vector might represent more than one speech feature. The effect of window length on performance of the speaker recognition system under reverberation will therefore be studied.

Theoretical modeling of the effect of reverberation in the STFT domain can be performed by applying the equations of STFT to the output of a linear filter, representing room impulse response. A preliminary model and conclusions are presented in Appendix A.2.

3.5 Feature Extraction

Feature extraction in this research yields the *Mel frequency cepstral coefficients* (MFCC) feature vectors (see Appendix C.2). The following configuration is typically used:

1. 20-30 ms STFT window length,
2. 20 triangular filters in the Mel scale filter-bank,
3. using only the 12 MFCC parameters from c_1 to c_{12},
4. compute dynamic parameters delta and/or delta-delta by time-derivation of the MFCC, and
5. perform *cepstral mean subtraction* (CMS).

This configuration is being used in a parallel research as a part of a Magneton project in cooperation with *NICE-systems Ltd.*

3.6 Statistical Modeling

The distribution of the MFCC feature vectors is modeled by the *Gaussian mixture model* (GMM). The number of mixtures in the speakers GMM will be highly dependent on the amount of enrollment speech, e.g. 64-256 mixtures [12], 1024 mixtures when using

Gaussian mixture model – unified background model (GMM-UBM) [4], or more when using subpopulations within the GMM-UBM [4]. GMM-UBM will be used in *adaptive Gaussian mixture model* (AGMM) [4]. However, the amount of mixtures might be reduced for computational efficiency and convenience.

The distribution distance measure is likely to be the symmetric *Kullback-Leibler* divergence (see Appendix D.4). The special case of one mixture in the GMM will be used for testing the behavior of the analytic solution of the divergence integral. In the general case where more than one mixture are used, the distance is to be computed numerically by the *earth movers distance Kullback Leibler divergence* (EMD-KL) [27] or by any other way that may be encountered during the research.

3.7 Experimental Setup

The impulse response of a room depends highly on its acoustic parameters. Following is a list of experiments in which we will try to isolate various acoustic parameters of the room, and examine their influence on the distribution of the speech feature vectors in the feature space. In this way, we may have a better understanding of the effect of reverberation on the performance of speaker recognition systems. The acoustic parameters under examination are reverberation time T_{60}, absorption of the walls, distance between speaker and microphone, and room dimensions.

3.7.1 Experiment 1

Experiment Aim

To investigate the effect of the room impulse response decay time, quantified using the reverberation time parameter T_{60}, on the MFCC feature vectors.

Experiment Method

1. Use a clean speech data-set of both male and female speakers from the 1999 NIST-SRE training database [23]. This data set is composed of free speech recordings of

[4] In [4], there is use of 1024 mixtures for male subpopulation GMM-UBM, and for female subpopulation GMM-UBM. The two models are then pooled together to create 2048 mixture GMM-UBM.

duration one-minute each. The total size of the data-set is initially set to several minutes, and gradually increased until a consistant behavior can be determined.

2. Use the image method to generate room impulse responses. Reverberation time T_{60} is set by the reflection coefficients of the walls. Room dimensions and source-microphone location are to remain constant [5]. The distance from the source to the microphone is chosen to be greater than the radius of reverberation.

   The Schroeder integral is used to compute T_{60} of each response. This is done by considering T_{20} as the time which takes the impulse response energy to decay from -5dB to -25dB. Then, applying $T_{60} = 3 \cdot T_{20}$.

3. Convolute the speech data with the room impulse response in order to generate reverberant speech with various reverberation times, in the range of $0 - 3$ sec.

4. Compute MFCC feature vectors for each speech recording according to Section 3.5.

Analysis of Results

We wish to investigate the changes in the behavior of the MFCC feature vectors due to change in reverberation time. This can be done by using distance measures to come out with a single number that describes the distance or change, or by looking in more details at the behavior of the MFCC distributions.

1. Calculate distance measures that use the MFCC first and second order moments, e.g.

   (a) Euclidean distance between overall mean vectors, and

   (b) Mahalanobis distance between overall mean vectors using the average matrix of the overall covariance matrices (see Appendix D.4).

2. Model the statistical distribution of the MFCC feature vectors using GMM and calculate distribution distance measures. The distribution distance measure is likely to be the symmetric *Kullback-Leibler* divergence. GMM and divergence calculation are made according to Section 3.6.

3. More detailed analysis of the changes in the GMM, e.g.

[5] Varying these parameters is the base of another experiment

(a) investigating change in mean and variance for a gradual change in T_{60}, and

(b) comparing the changes for all Gaussians in the mixture.

This analysis should be done with low number of Gaussians, for having the ability to relate between the Gaussians of the clean and the reverberant data.

4. Check if the change is the same for static MFCC, delta-MFCC, and delta-delta-MFCC coordinate groups in the MFCC feature vectors. This is done by using 12 dimension feature space for modeling each group.

5. Examine the effect for different frequency bands. This would be done by avoiding the DCT, CMS, variance normalization, and time-derivation operations in the generation of the MFCC. Compare this effect with the frequency response of the room.

Summary and Conclusions

Summarize the main results and try to conclude a pattern or a behavior in the effect of reverberation time on speech parameters. Try to explain why these results are expected (or not).

Experiment Hypothesis

For long reverberation times (much longer than the STFT window length), there will be time-smearing, and so overlap of features. Increase of reverberation time increases this time-smearing, and is likely to increase the distance between the distribution of the clean MFCC feature vectors and the distribution of the reverberant ones. This effect might also cause the means of the GMM model to come closer together. Furthermore, the GMM might need fewer mixtures. In the limit, having very long reverberation time, each speech segment will contain similar spectrum and similar set of features, and the MFCC distribution can be described by a single Gaussian.

3.7.2 Experiment 2

Experiment Aim

Here we wish to isolate the distance from the speaker to the microphone as an acoustic parameter. Sabin's T_{60} is to remain constant, and Schroeder's T_{60} is changed by changing

that distance.

Experiment Method

Repeat experiment 1 method on Section 3.7.1 except that $T60$ is controlled by changing the distance from the speaker to the microphone (see item 2).

Analysis of Results

In addition to the analysis in Section 3.7.1, create a plot of the reverberation time vs. the distance from the speaker to the microphone. Yield an analysis of the distortion to the features as a direct function of the distance from the speaker to the microphone.

Experiment Hypothesis

When the distance from the speaker to the microphone is smaller than the radius of reverberation, the signal seems to be affected more by the direct speech than by its refelections. In that case, the feature vectors should not be affected even if the room is extremely reverberant according to Sabin's T_{60}.

The opposite case may also be examined. When the distance from the speaker to the microphone increases beyond the radius of reverberation, the signal seems to be affected not by the direct speech but by its refelections. Furthermore, the noise filed is considered to be diffused, which leads to the assumption that the effect on the feature vectors should remain the same if the distance is greater than the radius of reverberation.

3.7.3 Experiment 3

Experiment Aim

Here we wish to isolate the dimensions of the room as an acoustic parameter. Reverberation time depends on the room dimensions, in a way that the time interval between the reflections is increased with the increase of the volume of the room.

Experiment Method

Repeat experiment 1 method on Section 3.7.1 except that $T60$ is controlled by changing the dimensions of the room (see item 2). The change in room dimensions is performed

uniformly to all the walls in the room by multiplying each dimension with a same factor. The mean of all three dimensions represents the room dimensions parameter. This parameter is to be given the symbol L.

Analysis of Results

In addition to the analysis in Section 3.7.1, create a plot of the reverberation time vs. the room dimensions parameter, L (see Section 3.7.3). Yield an analysis of the distortion to the features as a direct function of L.

Experiment Hypothesis

Increase in room dimensions results in less frequent reflections. Since with each reflection a part of the energy is absorbed, reverberation time increases because the loss of energy becomes less frequent. However, despite the fact that reverberation time increases, the speech signal seems not to receive any additional energy from its reflections. It seems only to receive this energy at later speech frames. This transfer of reflection energy from close to far timings might have an effect on the distortion to the feature vectors. It is hard though to predict the manner of the effect before this experiment is performed.

3.7.4 Additional Experiments

Additional experiments may be used in order to investigate the effect of acoustic parameters of the room. Table 3.1 summarizes some acoustic parameters apart from the ones that had been mentioned in the three preceding experiments (Sections 3.7.1 – 3.7.3) and associated models that can be used in order to examine their effects.

Early to late energy ratio is a parameter that can affect the speech signal in a way that more energy in the late part of the impulse response gives rise to reverberation. Frequency dependent decay is a parameter that is associated with the effect of absorption dependence on frequency.

Modification of the distance from the speaker to the microphone was discussed on the second experiment in Section 3.7.2. This modification can help control the positioning of the reflections within the time domain of the impulse response. However, unlike the case of the reflection density dependence on the room dimensions (see experiment 3 in Section 3.7.3), positioning of the reflections is not straight forward.

Last, there is the nature of the impulse response in the frequency domain that can serve as an acoustic parameter. This nature may be of either *low pass filter* (LPF) or *high pass filter* (HPF) etc. This parameter can be set in two ways. One way is to use the dependence of the absorption in the room of different materials on the frequency. The less realistic yet simpler way is to filter the response to the shape that we want to inspect.

Table 3.1: Acoustic parameters and associated models.

Parameter Type	**Acoustic Parameter**	**Associated Model**
Time-domain energy decay.	Early to late energy ratio.	Image methods for generation of impulse response.
	Frequency-dependent decay.	Improved frequency-dependant reflection coefficients for frequency-dependant decay.
		Schroeder integral for decay analysis.
		Diffuse-field model for theoretical support.
Time-domain reflections behavior.	Exact position of individual reflection in early response.	Image method, modification of speaker and microphone positions.
Frequency-domain response.	Nature of response (LP, HP, ...).	Image method with frequency dependant decay.
		Filter to shape overall response.

Chapter 4

Research Plan

Table 4.1 shows the planning of the research throughout the four years of studying towards the Ph.D. degree. The first year is already accomplished.

The review regarding room acoustics and reverberation, speech feature vectors, and speaker recognition was performed mainly by text-books and early articles, and therefore placed at the appendix of this proposal. The review regarding recent developments in speaker recognition in reverberant environments is found in Chapter 1. Simulations of reverberant speech features vectors are introduced in Appendix A, which also describe the basic conjectures that have been made regarding the effect of reverberation, and the ways to analyze it. The initial results regarding these simulations had been taken into consideration while designing a set of experimental investigations that describe the future research. These experiments are shown in Chapter 3.

The second year is dedicated to the understanding of the effect of reverberation on speech feature vectors and on their statistical distribution in the feature space. This would be performed according to the set of experimental investigations as presented in this proposal. For that purpose, experimental tools are required in order to measure the effect of reverberation. These tools contain the data set, the simulation tool, and the measuring tools. The data-set will consist of the 1999 NIST-SRE data-base, as mentioned in Chapter 3. The simulation tool is currently the image method, hopefully to be improved to use consideration of frequency dependent absorption. The measuring tools measure the distortion in the statistical distribution of the feature vectors due to reverberation, and come out with a single number that describes the distortion. Analysis of the results will be done in order to try to conclude a pattern or a behavior in the effect

of reverberation on speech feature vectors.

During the third year, a new or modified feature extraction technique is to be developed. The performance of this feature extraction technique is to be investigated in a speaker recognition system, such as the SID simulator (see Chapter 3), by feeding it with the new extracted feature vectors. Hopefully, the use of the new feature vectors will come up with higher recognition rate than using the usual feature vectors in the presence of reverberation.

During the fourth year, final refinements and performance analysis is to be conducted in order to come out with a final formulation of an algorithm for the new feature extraction technique. A Ph.D. thesis will be composed and submitted.

Table 4.1: Research plan.

Year	**Planning**
First	Literature review on room acoustics and reverberation, speech feature vectors, and speaker recognition.
	Literature review on speaker recognition in reverberant environments.
	Simulations of reverberant speech feature vectors, and basic conclusions regarding the distortion caused by reverberation.
Second	Designing a set of experimental investigations.
	Submition of a research proposal.
	Developing tools for simulations in acoustics, feature extraction, and speaker recognition.
	Conducting experiments in order to understand the effect of reverberation on the feature vectors and their statistical distribution.
	Developing measure tools for the effect of reverberation.
	Analysis of the results and deduction of conclusions regarding the effect.
Third	Developing new feature extraction technique.
	Experimental investigation and performance analysis in a speaker recognition system.
Forth	Refinements to the feature extraction technique.
	Final performance analysis.
	Writing and submition of a Ph.D. Thesis.

Part II

Appendix

Appendix A

Initial Results

During the first year of research, several simulations and mathematical evaluations had been carried out. Those assisted to maintain an initial impression of the effect of reverberation. This chapter summarizes the initial results and the derived conjectures regarding the effect of reverberation.

Section A.1 presents the first step in the research in which an impulse response of a room is simulated. Measuring quantities such as reverberation time and radius of reverberation are calculated. Refer to Appendix B.2 for definitions of these. Time-domain waveforms and energy decay curves of both simulated and measured room responces are presented.

Section A.2 discusses the initial conjectures regarding the effect of reverberation on the STFT domain. Spectrogram of simulated reverberant speech is compared to the one of clean speech. Phoneme analysis is performed on these spectrograms. Definition of phonemes is done according to [26] and [28]. A basic formulation of the effect of reverberation on the STFT domain is also presented. This formulation leads to the hypothesis that in cases where the reverberation time is longer than the frame window that is used for STFT of clean signals, longer windows should be used in the STFT of reverberant signals. The usage of long windows in STFT is discussed in Section A.3. Such a utilization of STFT seems to diminish the differences between clean and reverberant spectrograms.

Section A.4 discusses the effect of reverberation on feature vectors of a speech signal. The analysis is done frame-by-frame and distortion of the feature vectors was measured with the Euclidean distance. The distortion was seemed to increase monotonously with

the increase of reverberation time.

Section A.5 introduces the use of the dynamic features delta and delta-delta MFCC. The effect of reverberation is measured first on the overall MFCC features. Then, due to difference in magnitude between static and dynamic features, the distance is measured separately for static, delta, and delta-delta MFCC features.

In Section A.6, the distribution of the feature vectors is demonstrated in a cross-section of two dimensional feature sub-space of the overall feature space. The two dimensions are chosen according to the biggest variance criterion. Clean and reverberant feature vectors are displayed one vs. another in that 2-D feature sub-space.

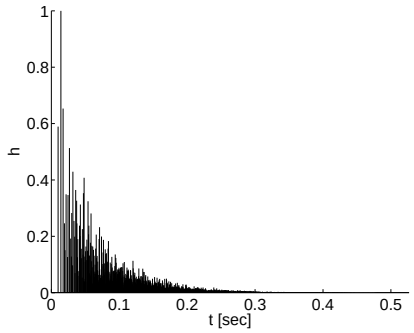
Reasons that are stated in Sections A.4 and A.5 leads to the conclusion that it would be better if the effect of reverberation is measured on the overall distribution of the feature vectors in the feature space. Hence, Section A.7 discusses the distortion that is made to the distribution of the clean feature vectors due to reverberation. The distance between the distribution of the clean vectors to the one of the reverberant vectors is calculated as a function of reverberation time. Same as in A.4, the distortion increased monotonously with the increase of reverberation time.

Section A.8 discussed the operation of *cepstral mean subtraction* (CMS) performed on the feature vectors, under the effect of reverberation. Section A.9 examine the change in that effect due to change in STFT window length.

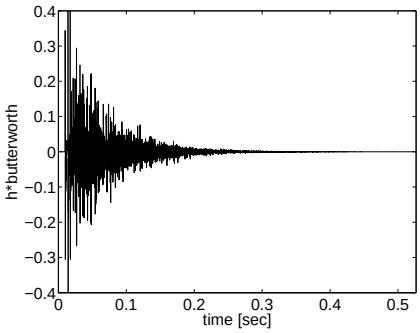
A.1 Simulating Reverberation

The impulse response of a simulated room is shown in Figure A.1. Room dimensions has been taken to be $5 \times 6 \times 7\text{m}^3$. Reflection coefficient from each wall is 0.8. Reverberation time T_{60} had been calculated to be 0.439 sec according to Sabin equation. Radius of reverberation r_d had been calculated as 1.238m. The speaker and microphone are located at $\boldsymbol{x}_s = (1, 1, 1)$ [m] and $\boldsymbol{x}_m = (3, 3, 3)$ [m], respectively. The distance between the speaker and the microphone is $\sqrt{2^2 + 2^2 + 2^2} = 3.464$m, which means that most energy is received from reverberation, rather than from the direct sound.

Figure A.1(a) shows the impulse response $h(t)$ as calculated by the image method. The same room response that includes the transducers effect $h_f(t)$ is shown in Figure A.1(b). This effect was simulated by passing the impulse response through a 4'th order Butterworth BPF with cutoff frequencies at 120 and 3600Hz.



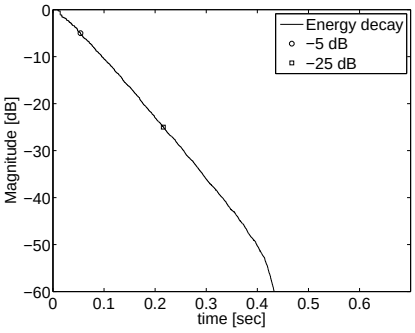
(a) $h(t)$, formed straight forward using the image method.



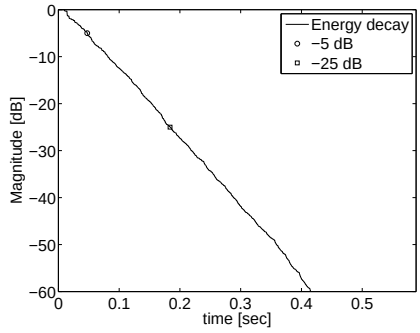
(b) $h_f(t)$, includes the transducers effect.

Figure A.1: Time domain simulated room impulse response.

The decay curve of the response is shown in Figure A.2. Figures A.2(a) and A.2(b) refer to the decay of the impulse responses $h(t)$ and $h_f(t)$, respectively. Reverberation time T_{60} according to Schroeder integral was calculated as 0.486 sec for $h(t)$, and 0.409 sec for $h_f(t)$.



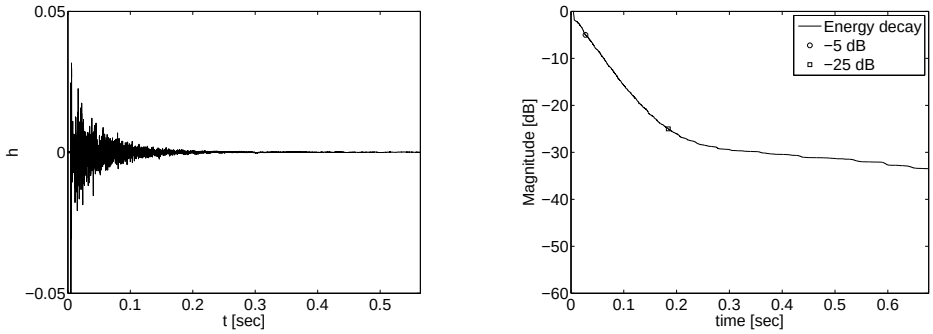
(a) of $h(t)$ from Fig. A.1(a),



(b) of $h_f(t)$ from Fig. A.1(b),

Figure A.2: Decay curve

Figure A.3 shows a measured impulse response of a real office. The distance from the speaker to the microphone was 1.55m. Reverberation time T_{60} had been calculated by the Schroeder integral as 0.470 sec. It can be seen from Figure A.3(a) that its nature in the time domain is more similar to the simulated response that includes the transducers effect, that is $h_f(t)$ from Figure A.1(b), than to the delta functions structure imposed by $h(t)$ in Figure A.1(a).



(a) Time response. (b) Decay curve.

Figure A.3: Measured impulse response of a real room.

A.2 Reverberation in the STFT Domain

Figure A.4 shows the effect of reverberation on the spectrogram of a speech signal. The clean signal on top of Figure A.4 is of the English word 'seven' spoken by a male speaker, taken from the TIDIGIT database [29]. The two peaks from 0.1 to 0.2 sec are caused by the breathing and preparation of the speaker to produce speech. The high-pass structure from 0.3 to 0.4 sec is of the *unvoiced* fricative consonant phoneme `'s'`. The four formants from 0.4 to 0.55 sec are of the *voiced* front vowel phoneme `'e'` [1]. Then there is a short drop from three formants to one formant, which is due to the *voiced* fricative consonant phoneme `'v'`. After that short drop there is a structure of four formants until 0.625 sec. This is due to the return of the phoneme `'e'`. From 0.625 sec until the end of the utterance at 0.8 sec there are three formants due to the *voiced* nasal consonant phoneme `'n'`.

The clean signal was reverberated by a convolution with the impulse response from Figure A.1(b). The effect can be seen in the manner that both voiced and unvoiced phonemes are smearing into later phonemes. The smearing can be seen in Figure A.4. The following are the most obvious evidences.

1. The two peaks between 0.1 to 0.2 sec which are unrelated to speech but to the breathing of the speaker, are smeared toward the speech itself.

2. At 0.4 sec, the high frequency noise from the phoneme `'s'` is smeared onto the four formants of the phoneme `'e'`. It seems even to last with low amplitude until the end of the utterance.

[1] The highest two formants may seem like one formant due to their close frequencies.

3. The phoneme ‘v’ that used to be at 0.55 sec in the clean spectrogram is vanished from the reverberant spectrogram due to the smearing of the preceding ‘e’ phoneme formants.

4. The second‘e’ phoneme at 0.6 sec is smearing to the later ‘n’ phoneme from 0.625 sec until the end of the utterance.

5. The whole spectrographic representation of the reverberant word seems to be longer than the one of the clean word.

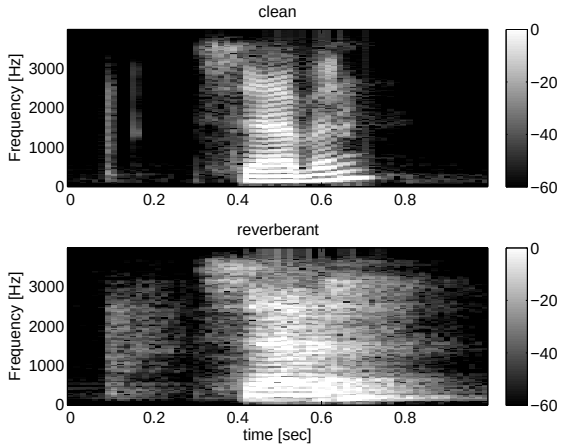


Figure A.4: Spectrograms of clean (on top) and reverberant (on bottum) signals.

A formulation of the effect of reverberation in the STFT domain had been performed by applying the equations of STFT on mathematical terms of convoluted speech signal with a room transfer function. Let $s(t)$ be a speech excitation signal in a room. That room has reverberation characteristics that can be modeled by an impulse response $h(t)$. This system is illustrated in Figure A.5. $g(t)$ is the window function of the STFT. $Y(t, \omega)$ is the STFT result, which is a function of the time t in which the window is placed, and the frequency ω.

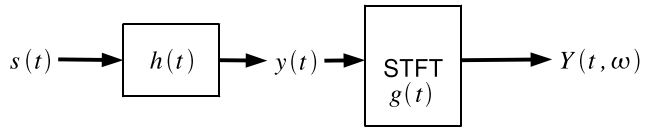


Figure A.5: A system model for STFT of a reverberant signal.

Then, $Y(t,\omega)$ can be expressed as:

$$
\begin{aligned}
Y(t,\omega) &= \int_{-\infty}^{\infty} y(\tau')\, g(t-\tau')\, e^{-j\omega\tau'} \mathrm{d}\tau' && \text{(A.1)} \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} s(\tau)\, h(\tau'-\tau)\, \mathrm{d}\tau \right] g(t-\tau')\, e^{-j\omega\tau'} \mathrm{d}\tau' \\
&= \int_{-\infty}^{\infty} s(\tau) \left[\int_{-\infty}^{\infty} h(\tau'-\tau)\, g(t-\tau')\, e^{-j\omega\tau'} \mathrm{d}\tau' \right] \mathrm{d}\tau \Big|_{\tau'-\tau\to u} \\
&= \int_{-\infty}^{\infty} s(\tau) \left[\int_{-\infty}^{\infty} h(u)\, g(t-u-\tau)\, e^{-j\omega u} \mathrm{d}u \right] e^{-j\omega\tau} \mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} s(\tau) \left[H(t-\tau,\omega) \right] e^{-j\omega\tau} \mathrm{d}\tau && \text{(A.2)}
\end{aligned}
$$

where $H(t,\omega)$ is the STFT of $h(t)$. In other words, Equation (A.1) which corresponds to Figure A.5, can be replaced by (A.2) that describes a STFT operation with the STFT of $h(t)$ as the window function, as shown in Figure A.6.

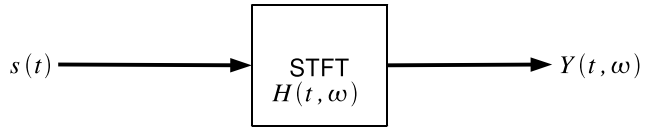
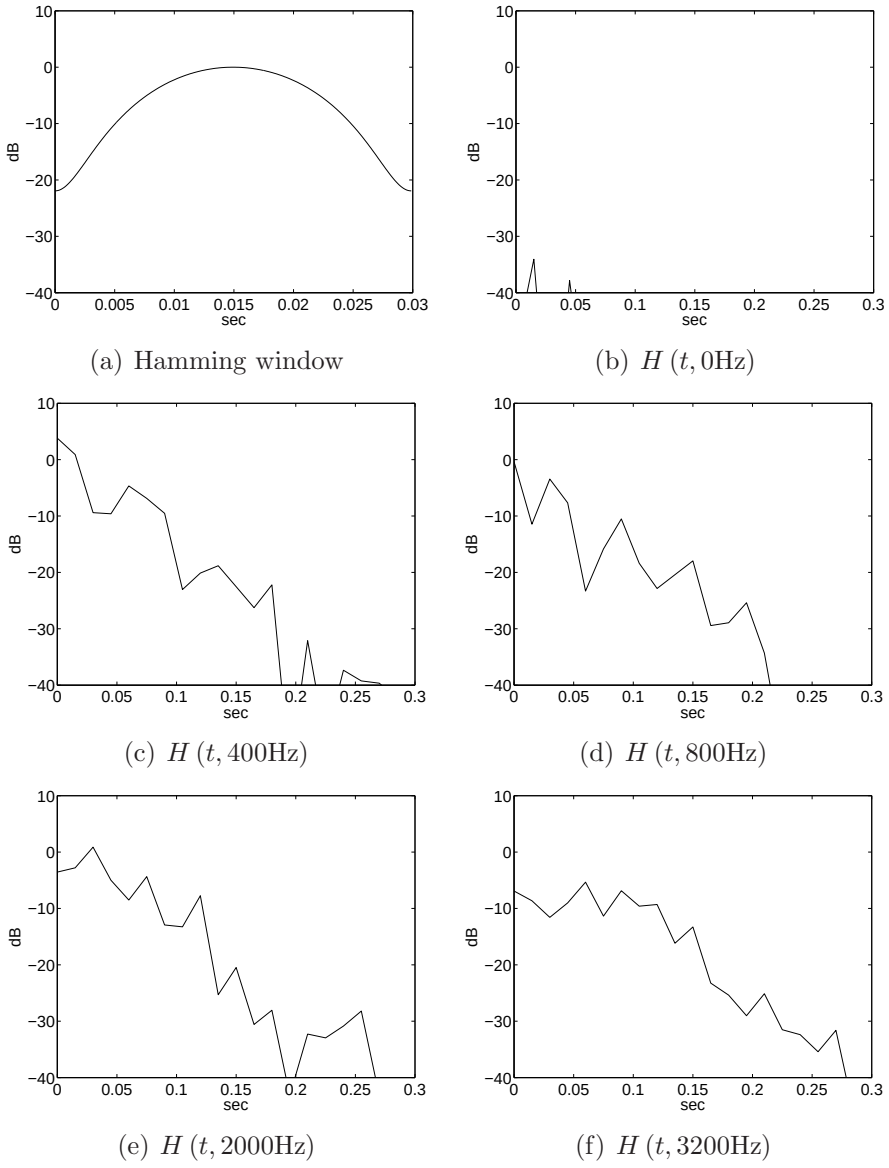


Figure A.6: STFT of a clean speech signal, using the STFT of $h(t)$ as the window function.

According to Equation (A.2) and Figure A.6, the STFT of the impulse response is the window that is used to extract the STFT of the reverberant signal. Since the impulse response is of long time duration, its STFT is also likely to be of long time duration. This result justifies the usage of longer frame windows, to be discussed on Section A.3.

In order to examine the assumption that $H(t,\omega)$ is of long duration, it had been simulated by calculating the STFT of the simulated impulse response $h_f(t)$ from Figure A.1(b) in Section A.1. $H(t,\omega)$ had been compared to the frame window $g(t)$. Figure A.7 shows the comparison between Hamming window type $g(t)$, and $H(t,\omega)$ in the frequencies of 0Hz (DC), 400Hz, 800Hz, 2000Hz, and 3200Hz. Except to DC, $H(t,\omega)$ seems to be longer than $g(t)$ at any frequency. The reason is that the reverberation time is longer than the duration of the Hamming window. Hence, the assumption was found to be correct.



(a) Hamming window

(b) $H\left(t, 0\text{Hz}\right)$

(c) $H\left(t, 400\text{Hz}\right)$

(d) $H\left(t, 800\text{Hz}\right)$

(e) $H\left(t, 2000\text{Hz}\right)$

(f) $H\left(t, 3200\text{Hz}\right)$

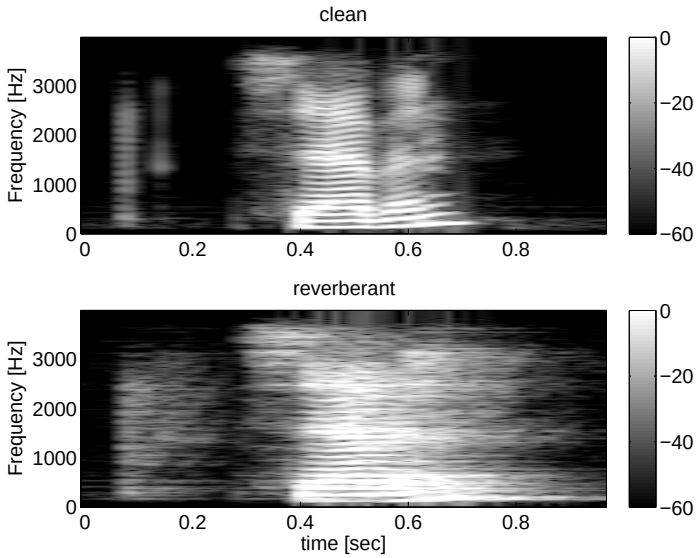
Figure A.7: Comparison between Hamming window $g\left(t\right)$ and STFT of the room impulse response $H\left(t, \omega\right)$ at different frequencies.

A.3 Long Windows

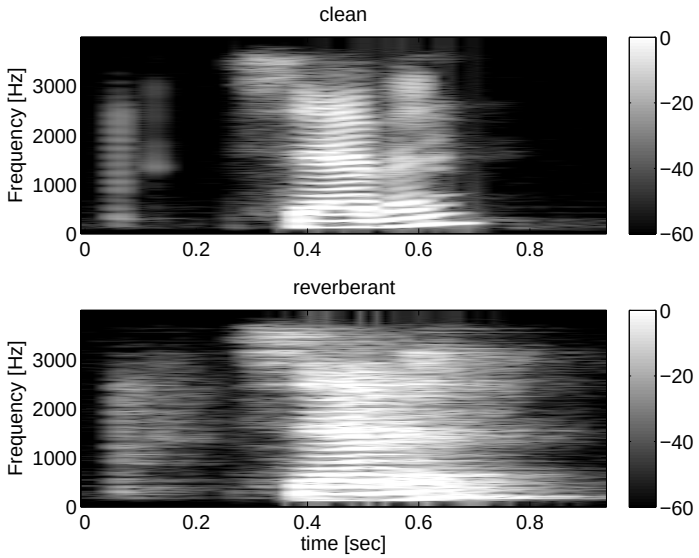
Spectrograms of STFT using long window duration are shown in Figure A.8. Figure A.8(a) shows clean vs. reverberant spectrogram of a speech signal with 60 ms duration window. According to the phoneme labeling in Section A.2 the voiced vowel phoneme `'e'` appears twice, first at 0.4 sec, and then again at 0.625 sec. This phoneme appears in a structure of four formants. In Figure A.8(a) however, the formant at the lowest frequency among the four of the first `'e'` phoneme, appear to smear onto the lowest

frequency formant of the second ‘e’ phoneme. In the case of 90 ms window, shown in Figure A.8(b), smearing in the clean spectrogram occurs also in the higher frequency formants.

This behavior is similar to the formant smearing in the reverberant spectrogram. Recalling the case of 30 ms window on Section A.2 in Figure A.4 , the smearing was evident only in the reverberant spectrogram. Hence, there seems to be less difference between the clean and reverberant spectrograms when using long windows.



(a) 60 ms window.



(b) 90 ms window.

Figure A.8: Spectrograms of clean and reverberant signals with long duration windows.

A.4 Measuring the Effect of reverberation on the Feature Vectors Frame by Frame

As an initial experiment that can give some information about the effect of reverberation on the MFCC feature vectors, the distance between frames of a clean speech signal and corresponding frames of the reverberant speech was measured with dependence on T_{60}. Intuition leads to the assumption that MFCC will change a more with increase of reverberation time. In order to test that hypothesis, a half minute long speech signal of a male speaker from the 1999 NIST-SRE database [23] was reverberated with various room impulse response functions of different reverberation time.

Reverberation time parameter was T_{60}. Room dimensions were $5 \times 6 \times 7\text{m}^3$, and remained constant. The reflection coefficient of the walls, R, had been given values from 0 to 0.97, to form increase in T_{60} from 0 to 3 sec. 12 coordinate MFCC feature vectors where calculated for 50% overlapping 30 ms time frames using Hamming window, and 20 triangular filters in the filterbank. For every impulse response, the mean of the Euclidean distance from MFCC vector of clean frames to the one of the time matching reverberant frames was calculated. Averaging throughout the progress in frame time was performed in order to yield the mean MFCC distance. Hence, the distance between a clean and reverberant feature vector given T_{60} had been assumed to be a first order *ergodic* discrete-time random signal.

The result is shown in Figure A.9. Each stem represents a result of a room. Indeed the distance is a monotonous increasing function of the reverberation time. This result strengthens the hypothesis.

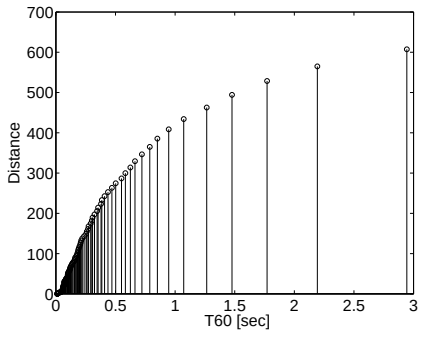


Figure A.9: Distortion in MFCC vs. reverberation time.

Also seen in Figure A.9 is that in the region of $[0, 0.4 \text{ sec}]$, the stems are organized

in a very dense manner relative to the region of [1 sec, 3 sec]. It implies that T_{60} is an increasing but not linear function of R. Furthermore, it is *strictly convex* [2] as can be seen in Figure A.10. The measurement of the distance used only the Schroeder integral form of T_{60}. Nevertheless, Figure A.10 shows that both Schroeder integral and Sabin equation derived T_{60} are strictly convex.

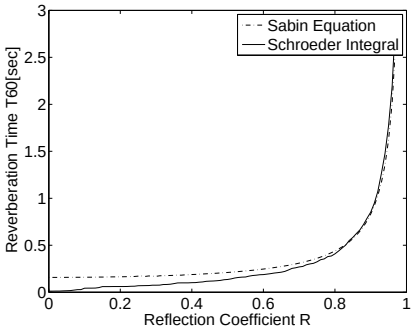


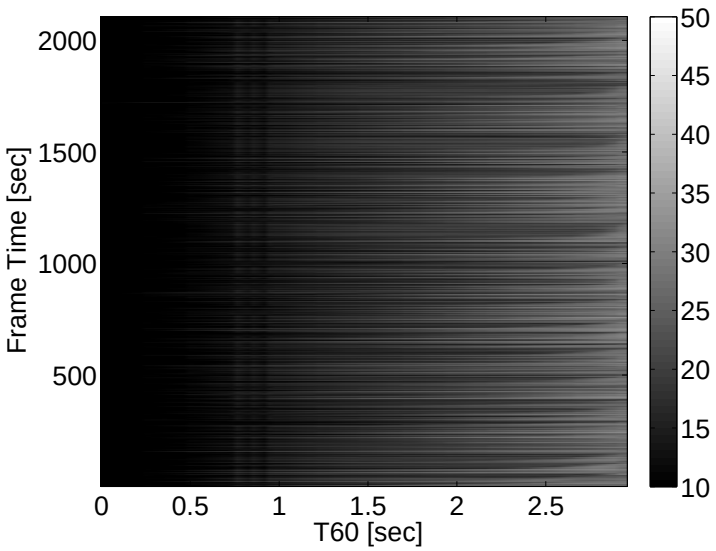
Figure A.10: T_{60} is a strictly convex function of R.

Separation in both frame time index and feature coordinate index has been made. The results are shown in Figure A.11. Figure A.11(a) shows the distance calculated for different time frames. Here, results were calculated in the same manner of Figure A.9, except that averaging over time frames was not performed. It can be seen from Figure A.11(a) that the distance is a monotonous increasing function of reverberation time for every time frame, except to the range of [0.7 sec, 1 sec]. This result supports the hypothesis with a slight restriction due to the a-monotonous behavior in that range.

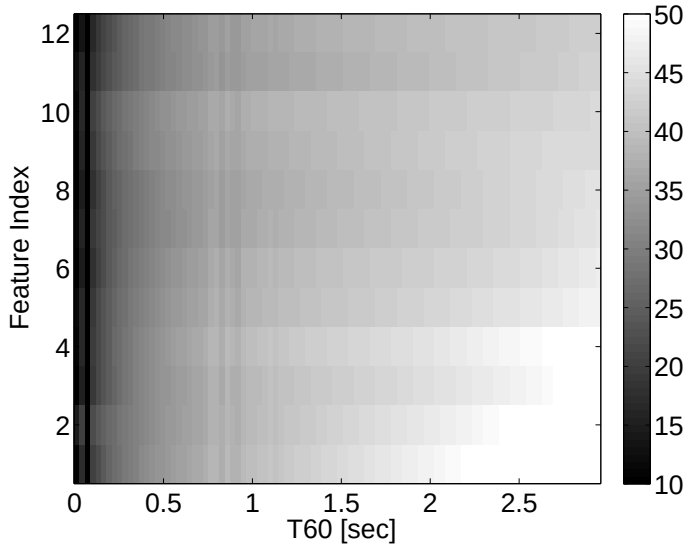
Figure A.11(b) shows the distance calculated for different feature indices. Feature index vectors were created by frame time index coordinates. 12 vectors where calculated for every impulse response, each represents a feature index from c_1 to c_{12} throughout the whole half minute speech signal. The Euclidean distance from clean vector to a matching feature index reverberant one was calculated for every room impulse response. It can be observed from Figure A.11(b) that $c_1 - c_4$ are increasing considerably with comparison to the last feature indices.

However, Figure A.11(b) does not provide information about the effect in different frequency bands, due to the *discrete cosine transform* (DCT) operation that is performed as the final stage of MFCC extraction. For having frequency analysis of the effect,

[2] $f(x)$ is considered *convex* if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ for any $t \in [0,1]$. If the sign $\leq$ can be switched in $<$ then $f(x)$ is *strictly convex*.



(a) Distance for different time frames.



(b) Distance for different feature indices.

Figure A.11: Distance as a function of frame time index or feature index, dB scale.

one must bypass the DCT, as presented in Figure A.12. In Figure A.12, 20 feature index vectors are calculated, each represents a filter in the Mel-scale filterbank. Here, no particular behavior seems to be evident for different frequency bands.

To conclude, measuring the distance framewise shown that indeed the distance increases with the increase of reverberation time. However, the effect of continuous phoneme smearing might not be taken into account by this method, due to the hard separation into frames. For that reason, a distance measure tool that is representative for the over-

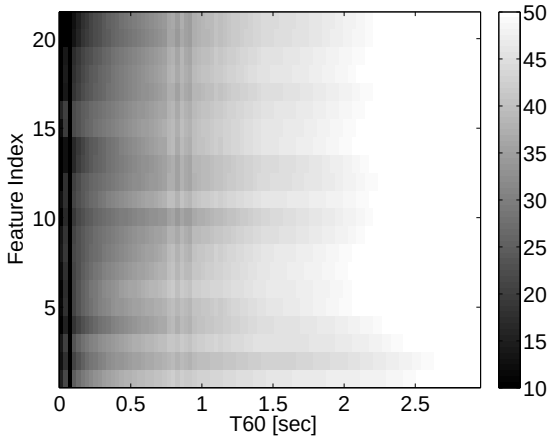


Figure A.12: Distance as a function of the Mel frequency band, dB scale.

all effect of reverberation on the speech features should be developed. A preliminary examination of it is presented in Section A.7.

A.5 Dynamic Features

The effect of reverberation on the dynamic features delta-MFCC and delta-delta-MFCC (see Appendix C.2) was examined, and compared to the effect on the static features. In order to do so, the same speech signal from Section A.4 was taken. Reverberation time T_{60} parameter was controlled in the same manner of Section A.4, with same room dimensions. Along with the static MFCC coordinates $c_1 - c_{12}$, delta-MFCC and delta-delta-MFCC were calculated as coordinates $c_{13} - c_{24}$ and $c_{25} - c_{36}$ respectively. The mean Euclidean distance from clean and reverberant feature vectors was calculated as a function of T_{60}, in a similar way to the one in Section A.4. The results are shown in Figure A.13, which shows a monotonous increase in distance.

The contribution of the dynamic coefficients delta and delta-delta was tested. Figure A.14 shows a distance calculation for every coordinate among $c_1 - c_{36}$. Delta and delta-delta contribution to the distance seems to be negligible. The reason is that the magnitude of the dynamic values is low, compared with the static ones, i.e. $c_{25} << c_{13} << c_1$.

Due to the big difference in magnitude, the MFCC, delta-MFCC, and delta-delta-MFCC were separated into groups of 12 coordinates. Each coordinate group was regarded as a separate feature vector. Figure A.15 shows the mean Euclidean distance as a function of T_{60} for each one of the three coordinate groups. The difference in behavior of the groups

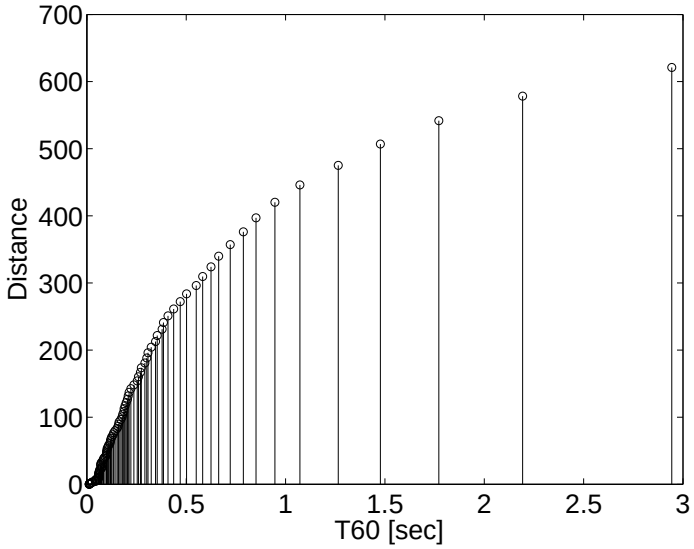


Figure A.13: Distance of MFCC with delta and delta-delta coefficients.

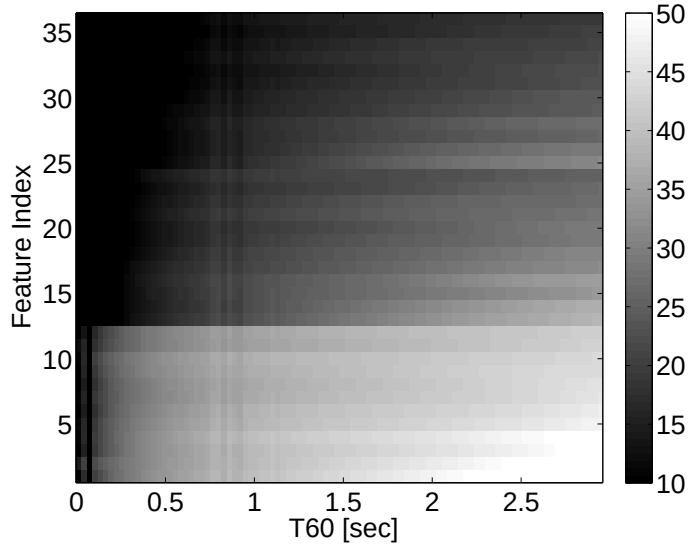
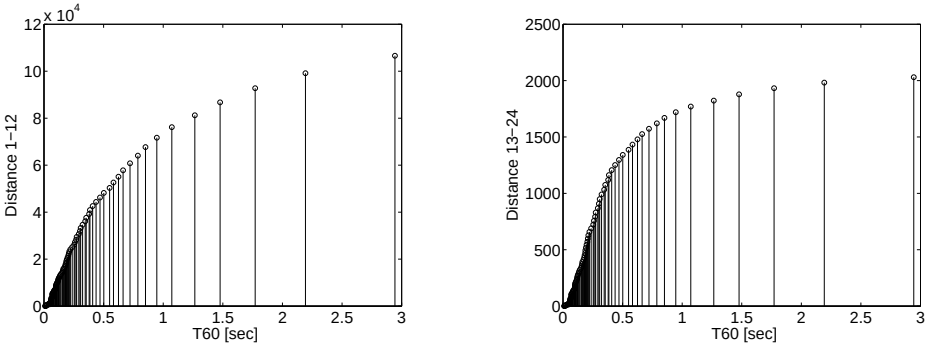
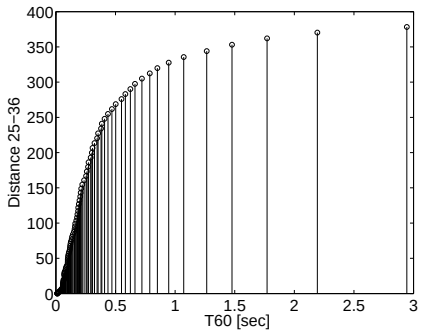


Figure A.14: Distance as a function of feature index including dynamic coefficients, dB scale.

can be seen in Figure A.15, in the region of $[0, 0.5$ sec$]$. The static features distance seems to increase slower than the dynamic features distance. Furthermore, the delta-delta-MFCC features distance seems to increase faster than the delta-MFCC features distance. All in all, it can be assumed that for low reverberation times (i.e. $T_{60} < 0.5\,\text{sec}$), the more dynamic the feature is, the more changed it would be due to reverberation.



(a) ...static MFCC features, (b) delta-MFCC features, and



(c) delta-delta-MFCC features.

Figure A.15: Distance of feature vectors, separated into groups of ...

A.6 2-D Display of Statistical Distribution

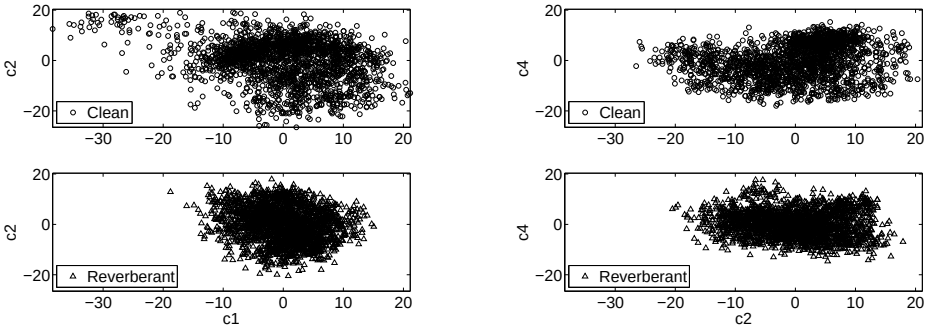
In order to display the distribution of the feature vectors in a 2-D plane, two indices of the feature vector coordinates should be chosen. The indices are chosen according to the magnitude of their variance. This operation is referred in the literature as *principle component analysis* (PCA) [30].

The distribution of the clean and reverberant feature vectors of the the same speech signal from Section A.4 is shown in Figure A.16. Room dimensions are $5 \times 6 \times 7\text{m}^3$, and reflection coefficient from each wall is 0.95. Reverberation time is 1.77 sec according to Schroeder integral. Circles represent the clean speech signal feature vectors, and triangles represent the reverberant feature vectors.

c_1 and c_2 had been found to be the largest variance coordinates of the clean speech feature vectors. Reverberation had seemed to change that however, as the largest variance coordinates of the reverberant speech feature vectors where c_2 and c_4. The conclusion is that the choice of coordinates of biggest variance is variant under reverberation, and hence the total feature space should be regarded when measuring the effect of reverberation on

the feature vector.

Figure A.16(a) shows the distribution of coordinate c_1 as a function of the coordinate c_2. Figure A.16(b) shows the distribution of coordinate c_2 as a function of the coordinate c_4. In both Figures, the clean feature vectors seem to be distributed with larger variance than the reverberant ones. The reverberant feature vectors seem to 'collapse' to the center. This phenomenon strengthens the hypothesis that reverberation causes an averaging throughout the whole feature space.



(a) PCA according to clean feature vectors.

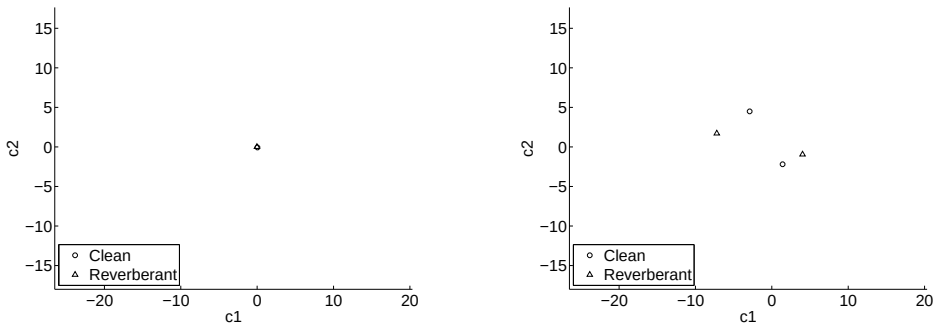
(b) PCA according to reverberant feature vectors.

Figure A.16: MFCC feature vectors distribution of clean vs. reverberant signals.

The effect of the reverberation on the mean vector is visualized in 2-D in Figure A.17(a). It seems like there is almost no effect on the mean vector of the distribution. However, the distribution can be modeled with GMM of more than one Gaussian. In that case, the effect of reverberation on the mean vectors can be visualized in a general manner, but it is hard to make conclusions regarding each and every mean vector. The reason is that the origin of the reverberant mean feature vector in the clean set cannot be determined. Figures A.17(b) – A.17(d) illustrate that point. This is yet another reason to examine the effect of reverberation on the distribution in the overall feature space.

A.7 Measuring the Effect of reverberation on the Distribution of the Feature Vectors

For reasons that had been mentioned in Sections A.4 and A.5 it was clear that the effect of reverberation on the overall distribution of the feature vectors should be examined. The same speech signal from Section A.4 was taken. Reverberation time T_{60} parameter



(a) 1 Gaussian

(b) 2 Gaussians

(c) 3 Gaussians

(d) 4 Gaussians

Figure A.17: 2-D visualization of the effect of reverberation on the mean vectors.

was controlled in the same manner of Section A.4, with same room dimensions. Only static features were issued, and calculated in the same manner as in Section A.4. The feature vectors were assumed to have a normal distribution with a full covariance matrix. The speech was reverberated using the image method and the distance between the distribution of the clean feature vectors and the distribution of the reverberant feature vectors was calculated as a function of the reverberation time T_{60}. The distance was calculated using the symmetric Kullback-Leibler divergence. (see Appendix D.4). The results are shown in Figure A.18 which include only reverberation times that are less than 1 sec.

As Expected, the divergence is an increasing monotonous function of the reverberation time. That result has significant meaning because it measures the effect on the cross-frame time distribution of the feature vectors from all frames, as oppose to the frame-by-frame comparison that was made in Section A.4.

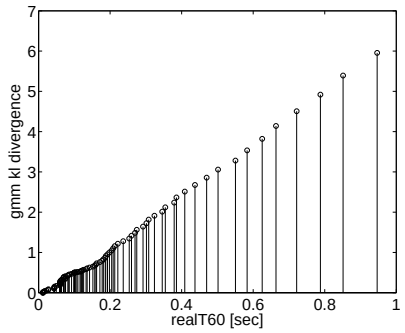


Figure A.18: Kullback-Leibler divergence from clean GMM to reverberant GMM of one center and full covariance matrix.

A.8 Cepstral Mean Subtraction

The same experiment in Section A.7 was performed with the use of *cepstral mean subtraction* (CMS). The results are shown in Figure A.19.

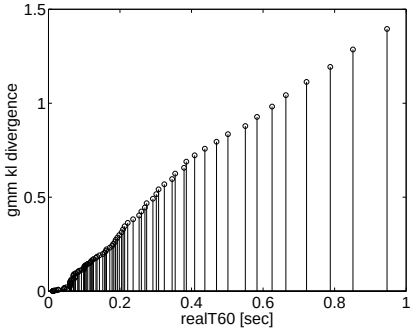


Figure A.19: Kullback-Leibler divergence from clean GMM to reverberant GMM of MFCC feature vectors, where CMS was performed.

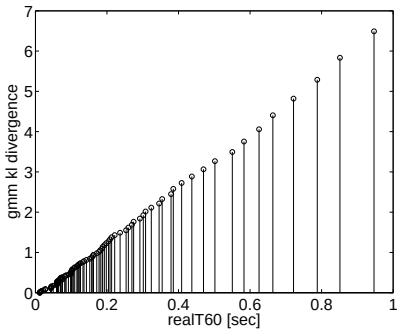
In general, the usage of CMS decreased the values of the divergence, approximately by a factor of 4. However, There is an evidence of sharp changes in the curvature of the graph due to the CMS, unlike in Figure A.18 without the CMS. For reverberation time T_{60} in the region of $[0, 0.1 \text{ sec}]$, there is a sharp increase of the divergence which then turns plateau until $T_{60} = 0.2$ sec.

A.9 Mixing CMS with Long STFT frames

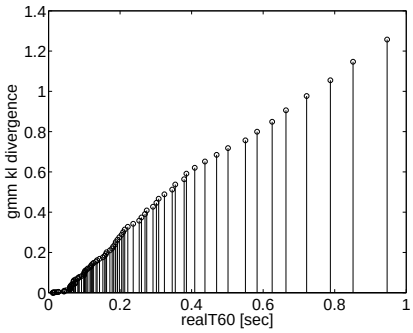
Figures A.20 – A.22 show the comparison of the KL divergence from clean to reverberant signal with and without CMS as a function of reverberation time, with the duration of

the STFT window as a parameter. Figures A.18 and A.19 from Sections A.7 and A.8 had already shown the KL divergence with and without CMS when using 30 ms STFT window length. Figure A.20 repeats Figures A.18 and A.19 but with 60 ms STFT window length. Figures A.21 and A.22 show the cases of 90 and 120 ms STFT window length respectively.

Figures A.20 to A.22 show that the divergence gets smaller in general by using CMS, when the STFT window increases. This result fits the assumption that longer STFT windows approximate the reverberation to simulate a channel, in which the effectiveness of CMS had been proven. Also, it seems like with the increase of STFT window length, the changes in divergence curvature are relaxed for the no-CMS configuration.

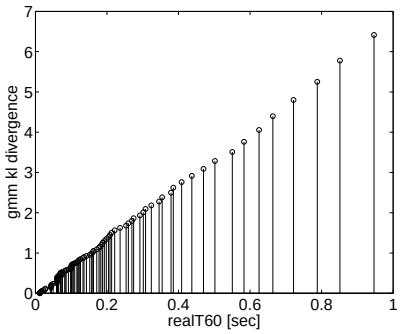


(a) Without CMS.

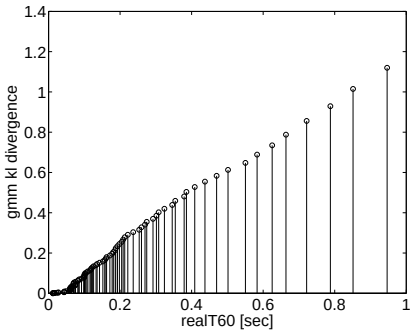


(b) With CMS.

Figure A.20: KL distance, 60 ms STFT window.

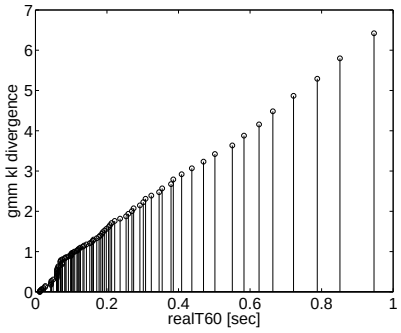


(a) Without CMS.

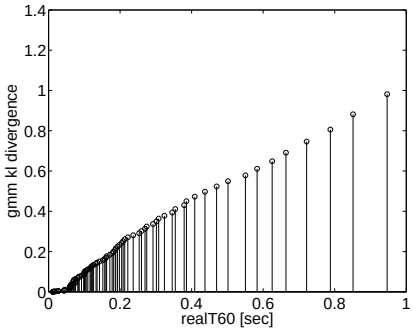


(b) With CMS.

Figure A.21: KL distance, 90 ms STFT window,



(a) Without CMS.



(b) With CMS.

Figure A.22: KL distance, 120 ms STFT window.

Appendix B

Room Acoustics and Reverberation

This chapter models reverberation in a small room. Section B.1, *modal model* of a room, relies on references [31, 32]. In general, it is effective for low frequencies, where separated resonances exist. It can be used to calculate the frequency response of a room, and following by that, its impulse response. In the cases where the frequencies are high, or alternatively, the modes are dense, this model is no longer efficient.

For high frequencies, more efficient analysis can be made according to the *diffuse-field model* of a room, which is introduced in Section B.2 and relies on references [31, 32, 33]. The diffuse-field model supplies an energetic analysis of the sound in the room. It can be used to calculate power, to model the decay, and to measure the reverberation time. It cannot however be used to calculate the impulse response of a room.

Section B.3 describes the *image method*, developed by *Allen and Berkley* [34]. The image method is an algorithm to simulate reverberation in a small room in a fast and efficient way. This method is efficient for high frequencies as well as for low frequencies. This method also considers the absorption of the sound that is caused by the walls, and not necessarily assumes rigid walls.

B.1 Modal Model of a Room

The modal model of a room can be used to calculate a frequency response of a room at low frequencies. The impulse response of the room can be reached by operating the inverse Fourier transform on the frequency response.

Section B.1.1 discusses the trivial case where all the walls in the room are rigid. This

case results in developing base functions whose linear combination comprise the frequency response.

Section B.1.2 discusses the case where the walls of the room are no longer considered rigid, and a sound source exists. In that case the frequency response includes damping considerations of the energy of the direct sound from the sources. A point-to-source frequency response can be calculated according to this model, where 'point' refers to the spatial location in which the response is calculated, and 'source' refers to the spatial location of the sound source.

Section B.1.3 discusses the issue of mode density within a frequency bandwidth. A term that describe this density is introduced, and used to decide whether the acoustic field can or cannot be efficiently modeled by the modal model.

B.1.1 Rigid Walls

The acoustic pressure field p is represented by:

$$p(\boldsymbol{x}, t) = p(\boldsymbol{x})\, e^{j\omega t} \tag{B.1}$$

where $\boldsymbol{x} = (x_1, x_2, x_3)$ and t represent displacement and time respectively, and ω represents the angular frequency of the sound wave.

$p(\boldsymbol{x}, t)$ in (B.1) is a solution of the *wave equation*:

$$\nabla^2 p - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{B.2}$$

where c is the speed of sound, and therefore for an harmonic field without sound sources must maintain the *Helmholtz equation*:

$$\nabla^2 p + k^2 p(\boldsymbol{x}) = 0 \tag{B.3}$$

$$k = \frac{2\pi}{\lambda} = \frac{\omega}{c} \tag{B.4}$$

where k is the *wave number*, and λ is the *wave length*.

Equation (B.3) is solved by the method of variable separation to the following form:

$$p(\boldsymbol{x}) = \sum_{n=(0,0,0)}^{(\infty,\infty,\infty)} a_n \psi_n(\boldsymbol{x}) \tag{B.5}$$

$$\psi_n(\boldsymbol{x}) = \sqrt{\epsilon_{n_1}\epsilon_{n_2}\epsilon_{n_3}} \cos\frac{n_1\pi}{L_1}x_1 \cos\frac{n_2\pi}{L_2}x_2 \cos\frac{n_3\pi}{L_3}x_3 \tag{B.6}$$

Every permutation $n = (n_1, n_2, n_3)$ is referred to as a *mode*. Thus, a general solution is a linear combination of all possible modes. The base functions $\psi_n(\boldsymbol{x})$ are known as the *modal shape* functions. The coefficients a_n are the *mode amplitudes*.

The discrete frequencies in which the modes exist are called *mode frequencies*. They are also known as the *resonance frequencies* of the room. The resonance frequencies are related to the discrete *mode wave numbers* in the following form:

$$k_n^2 = \left(\frac{\omega_n}{c}\right)^2 = k_1^2 + k_2^2 + k_3^2 = \left(\frac{n_1\pi}{L_1}\right)^2 + \left(\frac{n_2\pi}{L_2}\right)^2 + \left(\frac{n_3\pi}{L_3}\right)^2 \tag{B.7}$$

B.1.2 Sources and Non-Rigid Walls

In case where the walls are not rigid, the impulse response should include damping. An exhaustive development of the modal model with sources and damping can be found in [32]. Suppose we have a point source of strength q_s that is located at $\boldsymbol{y}_s$. Then a_n of Equation (B.5) is a function of the frequency ω and equals to

$$a_n(\omega) = \frac{\omega\rho_0 c^2}{V\left(2\xi_n\omega\omega_n + j\left(\omega_n - \omega\right)^2\right)} \cdot q_s\psi_n(\boldsymbol{y}_s) \tag{B.8}$$

where ξ_n is the *damping coefficient* of the mode n. ω_n is the resonance frequency of the mode n and calculated according to Equation (B.7). V is the volume of the room and equals to

$$V = L_1 \cdot L_2 \cdot L_3 \tag{B.9}$$

ρ_0 is the air density in equilibrium, or, the *ambient air density*.

B.1.3 Modal Density

Every mode describes the dependence of the sound pressure on frequency resonance that is related to that mode, on its spatial location within the room. The *resonance frequencies*

can be described as discrete points in a 3D space where:

$$f_n = (f_1, f_2, f_3) = \left(\frac{n_1 c}{2L_1}, \frac{n_2 c}{2L_2}, \frac{n_3 c}{2L_3} \right) \tag{B.10}$$

The number of modes within a sphere of radius f_0 can be calculated by referring all f_n that maintain

$$\sqrt{f_1^2 + f_2^2 + f_3^2} < f_0 \tag{B.11}$$

Since f_n is positive, the only point that is relevant among all those which created by the permutations of $n = (\pm n1, \pm n2, \pm n3)$, is the one that is related to $n = (n1, n2, n3)$. Hence, only Eighth of the sphere's volume is relevant.

N_{f_0} is defined as the number of modes that maintain (B.11) which are in a frequency box of volume $\frac{c}{2L_1}\frac{c}{2L_2}\frac{c}{2L_3} = \frac{c^3}{8V}$, and which are set by the permutations of $n = (\pm n1, \pm n2, \pm n3)$. In order to calculate N_{f_0}, we divide eighth of a sphere's volume of radius f_0 by the volume of the frequency box to have:

$$N_{f_0} \approx \frac{4}{3}\pi V \left(\frac{f_0}{c} \right)^3 \tag{B.12}$$

We define the *modal density* as the change of the number of modes as the function of the frequency:

$$\frac{\mathrm{d}N}{\mathrm{d}f} = 4\pi V \frac{f^2}{c^3} \tag{B.13}$$

In the cases where the number of modes within a bandwidth of -3dB is over 3, then the sound pressure field is considered complicated, or *diffused*. In these cases the modal model is no longer efficient due to the high order of the impulse response, caused by the large amount of modes. The frequency which defines the transition from the modal model to the diffused model is known as the *Schroeder frequency* [35]

B.2 Diffuse-Field Model of a Room

For high frequencies, an efficient analysis can be made by the statistical *diffuse-field model* of a room, as opposed to the modal model which is efficient only in low frequencies. The diffuse-field model supplies an energetic analysis of the sound in the room. It can be used to calculate power, to model the decay, and to measure the reverberation time. It cannot however be used to calculate the impulse response of a room.

B.2.1 Diffuse-Field

For diffuse sound field we assume that the acoustic field consists of an infinite number of plane waves which travel in all directions. This assumption is a reasonable due to the large amount of modes. We also assume uniform amplitude for all waves. This assumption is less reasonable however, because reflected waves are usually of lower amplitudes than the direct waves, and therefore the amplitude is not uniform. Yet, we make that assumption for the convenience of the solution, and the fact that the result is reasonable. A third assumption that we make is that all waves have random phase. This assumption is reasonable due to the random travel paths in a room.

B.2.2 Energy Transfer

The pace of energy transfer, or *average power*, to a boundary surface element of area ΔS in a diffuse-field is given in the following equation [33]:

$$\frac{\Delta E}{\Delta t} = \frac{\epsilon \cdot \Delta S \cdot c}{4} \tag{B.14}$$

where ϵ is the *energy spatial density* in the space of the room. In the limit where Δt approaching zero, we get the *power for area unit*, or *intensity*:

$$I = \frac{1}{\Delta S}\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\epsilon \cdot c}{4} \tag{B.15}$$

B.2.3 Absorption

We assume that the boundary surfaces have area S all together. Let us define the *absorption surface*, A, as the boundary surface that swallows energy. A is defined as the weighted average of the total absorption on the boundary surfaces. That is,

$$A = \sum_i A_i = \sum_i S_i a_i = S\hat{a} \tag{B.16}$$

where i is the index of a surface, the summation goes over all the boundary surfaces, and S_i is an i'Th boundary surface with an absorption of a_i. $\hat{a} = \frac{A}{S}$ is the mean absorption.

According to Equations (B.15) and (B.16), if a room has an absorption surface A,

then the pace in which the energy absorbed would be

$$A \cdot \frac{\epsilon \cdot c}{4} \tag{B.17}$$

Assuming a source in the room that emits energy with power Π, we can deduce the equation that describe the dynamics of energy in the room:

$$A \cdot \frac{\epsilon \cdot c}{4} + V\frac{\mathrm{d}\epsilon}{\mathrm{d}t} = \Pi \tag{B.18}$$

Equation (B.18) is a first order differential equation. Therefore, its solution is exponential. Assuming that the source starts its emission at time $t = 0$, the solution for ϵ is of the form:

$$\epsilon(t) = \frac{4\pi}{Ac}\left[1 - e^{-\frac{t}{\tau}}\right] u(t) \tag{B.19a}$$

$$\tau = \frac{4V}{Ac} \tag{B.19b}$$

that is, the time constant τ is dependent on the room volume V, and on the absorption surface A. In the steady state the solution would be:

$$\lim_{t\to\infty} \epsilon(t) = \frac{4\Pi}{Ac} \tag{B.20}$$

which is independent on the volume of the room.

B.2.4 Reverberation Time

The solution of the diffuse-field can be used to calculate the *reverberation time*, which is one of the most important parameters in room acoustics. Let as assume that the source in the room is active for a long duration, and then terminated at time $t = 0$. In that case we can write the equation of energy as:

$$\epsilon = \epsilon_0 e^{-\frac{t}{\tau}} \qquad , t \geq 0 \tag{B.21a}$$

$$\tau = \frac{4V}{Ac} \tag{B.21b}$$

The time which takes to the *sound pressure level* (SPL) in the room to diminish by 60dB, is defined as *reverberation time*, and denoted by T_{60}.

Sabine Equation

Since ϵ is proportional to the square of the pressure, we can use Equation (B.21) to find T_{60} by letting ΔSPL to be equal to 60 dB where

$$\Delta\text{SPL} \quad = \quad 10\log_{10}\frac{\epsilon_1}{\epsilon_2} \quad = \quad \frac{t_2 - t_1}{\tau}\cdot 10\log_{10} e \quad = \quad \frac{\Delta t}{\tau}\cdot 4.34 \tag{B.22}$$

Substituting $\Delta\text{SPL} = 60$ in (B.22) and using Equation (B.21b) yields $\Delta t = T_{60}$:

$$T_{60} = 0.161\frac{V}{A} \tag{B.23}$$

Equation (B.23) is known as the *Sabine Equation* [36].

Measurement of Reverberation Time from the Impulse Response

If one measures T_{60} in order to compare it with the theoretical *Sabine Equation* on (B.23), he has to, by definition, play noise in the room for a long duration of time, terminate it, and messure the decay. An alternative way that had been suggested by *Schroeder* in 1965 [37] is to messure reverberation time from the *source-to-microphone* impulse response.

Assume a source $x(t)$ is producing white noise $n(t)$ of variance σ^2 until time $t = 0$, and is then turned-off. We assume that the source-to-microphone impulse response $h(t)$ had been measured. The noise that is measured in the microphone $s(t)$ would be:

$$s(t) = \int_{-\infty}^{0} n(\tau)\, h(t-\tau)\, \mathrm{d}\tau \tag{B.24}$$

Representing the energy of the noise by the expectation of the square of the noise $E[s^2(t)]$ yields [37]:

$$E\left[s^2(t)\right] = \sigma^2 \int_{t}^{\infty} h^2(\tau)\, \mathrm{d}\tau \tag{B.25}$$

Equation (B.25) allows us to integrate $h(t)$ instead of measuring the noise in the room. T_{60} is found according to (B.25) as the one that maintains (B.26):

$$1000 \cdot \int_{T_{60}}^{\infty} h^2(\tau)\, \mathrm{d}\tau = \int_{0}^{\infty} h^2(\tau)\, \mathrm{d}\tau \tag{B.26}$$

B.2.5 Radius of Reverberation

Assuming a source in the room, *radius of reverberation* r_d is defined as the distance from the source at which energy received from the direct field is equal to the energy received from the reverberant field. r_d is given by [33],

$$r_d = \frac{1}{4}\sqrt{\frac{A}{\pi}} \tag{B.27}$$

For distance that is smaller than r_d most of the energy is received from the direct field and vice verse.

B.3 The Image Method

The *image method* [34] assumes a rectangular room enclosure. This model calculates the room impulse response using a time-domain image expansion method. It is aimed to be simple, fast, and easy to use algorithm.

Assumming an angle independent reflection coefficient β, the point-to-source room impulse response according to the image method is

$$\begin{aligned} p\left(t, \boldsymbol{X}, \boldsymbol{X}'\right) &= \sum_{\boldsymbol{p}=(0,0,0)}^{(1,1,1)} \sum_{\boldsymbol{r}=(-\infty,-\infty,-\infty)}^{(\infty,\infty,\infty)} \beta_{x_1}^{|n-q|} \beta_{x_2}^{|n|} \beta_{y_1}^{|l-j|} \beta_{y_2}^{|l|} \beta_{z_1}^{|m-k|} \beta_{z_2}^{|m|} \\ &\quad \times \frac{\delta\left(t - \frac{|\boldsymbol{R}_p + \boldsymbol{R}_r|}{c}\right)}{4\pi\left|\boldsymbol{R}_p + \boldsymbol{R}_r\right|} \end{aligned} \tag{B.28}$$

where $\boldsymbol{X} = (x, y, z)$ and $\boldsymbol{X}' = (x', y', z')$ are the speaker and microphone location, respectively. x_1, x_2, y_1, y_2, z_1, z_2, are all wall indices. The triplet vector $\boldsymbol{p} = (q, j, k)$ defines the vector $\boldsymbol{R}_p$ as:

$$\boldsymbol{R}_p = (x - x' + 2qx', y - y' + 2qy', z - z' + 2qz') \tag{B.29}$$

and the triplet vector $\boldsymbol{r} = (n, l, m)$ defines the vector $\boldsymbol{R}_r$ as:

$$\boldsymbol{R}_r = 2\left(nL_x, lL_y, mL_z\right) \tag{B.30}$$

where L_x, L_y and L_z are the room's dimension in the x, y and z directions respectively.

Appendix C

Speech Features

This chapter summarizes digital speech signal processing and feature extraction techniques. An important tool in digital speech signal processing is the *short-time Fourier transform* (STFT) that is introduced in Section C.1. The feature extraction method that is presented here is the *Mel frequency cepstral coefficient* (MFCC). MFCC is fairly discussed in Section C.2.

C.1 Short Time Fourier Analysis

Speech signals can be modeled as non stationary random processes. In order to perform frequency analysis of it, it is typically splitted into frames where it is considered stationary. It has been shown [26] that in 10 – 30 msec frames the speech signal has random parameters that change slowly with time, due to anatomical reasons that are related to the change rate of the vocal tract. Such an analysis is called *short-time Fourier analysis* and is performed by the *short-time Fourier transform* (STFT) [26, 38]. If we let $x[n]$ be a discrete time signal with sample time T_s (i.e. $t = nT_s$), then its STFT, $X_n(\theta)$ at time n would be [26, 38]

$$X_n(\theta) = \sum_{m=-\infty}^{\infty} w[n-m]\, x[m]\, e^{-j\theta m} \tag{C.1}$$

where $\frac{\theta}{T_s}$ is the *angular frequency* ω, and $w[n]$ is a finite duration window sequence. In order to prevent distortions in the frequency domain due to noncontinuous character of the step function in the time domain, $w[n]$ is usually chosen to be a *Hamming* or a

Hanning window of the forms [39]

$$w_{Hamming}[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad \text{(C.2)}$$

$$w_{Hanning}[n] = 0.5\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \quad \text{(C.3)}$$

respectively, where N is the window duration and $0 \leq n \leq N-1$.

Let $X_n(\theta)$ be the STFT of $x(n)$. If we take the dB values of $X_n(\theta)$ and associate every value with a specific color, we get a two dimensional display of the energy of the speech signal, depending on time and frequency. Such a representation is known to be the spectrographic view or the *spectrogram* of the speech signal $x(n)$. Spectrograms can serve for visualizing the STFT, and therefore can be a monitoring tool for distortions in the STFT domain. In this research, the spectrogram helps to visualize the effect of reverberation on the STFT domain. Spectrograms can also serve as a phoneme labeling tool. One can see a list of phonemes and their matching spectrograms in [26, 28]. Hence, effects on the STFT can be issued to a certain phoneme or a phoneme type (e.g. voiced or unvoiced).

C.2 Cepstral Analysis

C.2.1 Cepstrum

The idea of cepstrum was first introduced by Bogert et all in 1963 [40]. Let $s(t)$ represent an output frame of a linear system $h(t)$ where $u(t)$ is an excitation signal.

$$s(t) = h(t) * u(t) \quad \text{(C.4)}$$

Applying Fourier transform on (C.4) yields:

$$S(\omega) = H(\omega) \cdot U(\omega) \quad \text{(C.5)}$$

Then, the definition of the *cepstrum* would be:

$$\begin{aligned} c(t) &= \mathcal{F}^{-1}\{\log S(\omega)\} && \text{(C.6)} \\ &= \mathcal{F}^{-1}\{\log H(\omega)\} + \mathcal{F}^{-1}\{\log U(\omega)\} && \text{(C.7)} \\ &= c_h(t) + c_u(t) && \text{(C.8)} \end{aligned}$$

In (C.8) $c_h(t)$ is attributed to the system, while $c_u(t)$ is attributed to the excitation signal. By an operation that is called *liftering*, which is referred to filtering in the new time domain, one can isolate the characteristics of the transfer function, or of the excitation signal. Isolating the characteristics of the transfer function can be thought as a deconvolution operation. If the linear system is a room, then this operation can be a method for *de-reverberation* [41, 42, 43]

C.2.2 Mel Scale and MFCC

Mel Scale

The *Mel scale* was proposed by Stevens et all in 1937 [44]. It offers a transformation from the real frequency to the *perceived frequency*, which is known as *pitch* [1]. The transformation is made in the following form:

$$f_{Mel} = \log\left(1 + \frac{f_{Hz}}{700}\right) \cdot \frac{1000}{\log\left(1 + \frac{1000}{700}\right)} \qquad \text{(C.9)}$$

where f_{Mel} is the pitch perceived when the frequency is f_{Hz}.

MFCC

The *Mel scale frequency cepstral coefficients* (MFCC) are calculated in a similar way to the cepstrum, except that the frequency domain is warped into Mel frequency domain by applying (C.9). Then, energy is calculated in the Mel frequency domain over equally spaced overlapping triangle shaped filter banks.

An example for common configuration of the algorithm of MFCC can be [45]:

1. Sampling frequency is set to 8 kHz.

2. Time frame duration is set to 20 msec. This results in 160 samples per time frame.

[1] not to be confused with *fundamental frequency*.

3. Framing in the time domain is done with the use of 50% overlapping Hamming windows.

4. Log spectral energy is calculated in the Mel-scale frequency domain with the use of 50% overlapping triangle windows. Number of windows is no less than 20. This results in a series of $N \geq 20$ coefficients.

5. The *discrete cosine transform* (DCT) of the series of length N from step 4 is calculated. $c_{13} \dots c_N$ are eliminated. This results in a series $c_0 \dots c_{12}$.

6. c_0 is eliminated. The MFCC vector is of 12 length.

C.2.3 Dynamic Features

The cepstrum represents the local spectral properties of a given frame of speech. However, it does not characterize the temporal or transitional information in a sequence of speech frames. In automatic speech or speaker recognition systems improved performance had been achieved by introducing cepstral derivatives into the feature space. Cepstral derivatives are believed to capture the transitional information in the speech [1].

An estimation of the first derivative of the cepstrum can be achieved by Δc_n in the following equation [12]:

$$\Delta c_n = \frac{\sum_{k=-l}^{l} k \cdot c_{n+k}}{\sum_{k=-l}^{l} |k|} \tag{C.10}$$

Δc_n is referred to as the *delta cepstrum*. The second order cepstrum derivative can also be estimated and used as a feature. It is referred to as the *delta-delta cepstrum* and noted by $\Delta\Delta c_n$ using the following equation [12]:

$$\Delta\Delta c_n = \frac{\sum_{k=-l}^{l} k^2 \cdot c_{n+k}}{\sum_{k=-l}^{l} k^2} \tag{C.11}$$

If c_n represents a MFCC feature vector, then Δc_n and $\Delta\Delta c_n$ are the *delta*-MFCC and *delta-delta*-MFCC respectively.

C.2.4 Cepstral Mean Subtraction

Suppose we extract feature vectors for T frames where t represents a discrete time index of a frame so that $t = 0 \dots T-1$. Let $\boldsymbol{c}_t$ represent the t-th frame's MFCC feature vector,

$$\boldsymbol{c}_t = \begin{bmatrix} c_{t_1} \\ \vdots \\ c_{t_{12}} \end{bmatrix} \tag{C.12}$$

Once a collection of MFCC feature vectors throughout all frames is calculated, it can have a mean vector $\boldsymbol{\mu}$ by considering the random process $\boldsymbol{c}_t$ to be ergodic and using averaging in time, so that

$$\boldsymbol{\mu} = \frac{1}{T} \sum_{t=0}^{t=T-1} \boldsymbol{c}_t \tag{C.13}$$

Cepstral mean subtraction (CMS) is the operation of subtracting the mean vector from each MFCC feature vector, i.e.

$$\tilde{\boldsymbol{c}}_t = \boldsymbol{c}_t - \boldsymbol{\mu} \qquad t = 0 \dots T-1 \tag{C.14}$$

CMS is often used in speaker verification systems for the removal of slowly varying convolutive noise due to communication channel [12]. Hence, CMS is effective when the communication channel is approximately time invariant. The motivation for using CMS in reverberant environments lays in the consideration of the room as a channel. However, the point to source impulse response of the room is varying with time due to the change in location of the speaker relative to the microphone, and due to variations in a speaker's fundamental frequency [6]. Hence in such applications, the mean feature vector is calculated over a limited time duration during which the speaker is assumed to stay at a constant location, and his fundamental frequency is constant.

Appendix D

Speaker Recognition and Statistical Modeling

Speaker recognition is the process of recognizing who is speaking on the basis of individual information included in a speech signal [1]. *Speaker recognition* is divided into two topics namely, *speaker verification* (SVR) and *speaker identification* (SID) [2] [1]. SVR is considered to be a simpler operation, which is used to verify whether the speaker is an hypothesized speaker or not. SID uses the SVR results to identify which entity among a list the speaker represents.

Speakers are recognized according to a pattern matching of the statistical distribution of their feature vectors in the feature space, with known statistical distribution patterns of feature vectors that are related to other speakers. The statistical distribution of the feature vectors is modeled by statistical modeling methods. *Gaussian mixture models* (GMM) has become a dominant approach for statistical modeling of speech feature vectors [4].

D.1 Speaker Verification

In *speaker verification* (SVR) the task is to verify whether the tested speaker is an hypothesized speaker or not. Let

$$X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T] \tag{D.1}$$

[1] In [2] the terms ASV and ASI replace SVR and SID respectively, to indicate *automatic* recognition.

be a segment of speech feature vectors $\boldsymbol{x}_t$ of discrete time $t \in \{1, 2, \ldots, T\}$. Let H_0 represent the event that the tested speaker is the hypothesized speaker, and let H_1 represent the opposite event. A mixed random variable λ maps the events H_0 and H_1 to the models λ_{hyp} and $\lambda_{\overline{hyp}}$ respectively. The models may consist of vectors and matrices that describe the distribution of a single feature vector. For example, if the model sets a Gaussian distribution, then λ consists of a mean vector and a covariance matrix.

The decision is then made according to

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p\left(X|\lambda_{\overline{hyp}}\right) \begin{cases} \geq \theta & \text{accept hypothesized speaker} \\ < \theta & \text{reject hypothesized speaker} \end{cases} \tag{D.2}$$

where $p(X|\lambda)$ is the class conditional distribution density function, and θ is a threshold. $\Lambda(X)$ is referred to as the *log-likelihood ratio* function.

D.2 Speaker Identification

In *speaker identification* the task is to decide which entity among a list of hypothesized speakers, the speaker represents. Assuming N speaker indices $n = 1 \ldots N$, let there be N models $\lambda_1 \ldots \lambda_N$ mapped by the random variable λ in the same manner like in Section D.1 so that

$$\lambda_n = \text{The model of speaker } n. \tag{D.3}$$

Let as define a model $\lambda_{\overline{hyp}}$ for no speaker. The conditional probability density of X given $\lambda_{\overline{hyp}}$ is estimated as a function of all the conditional probability densities of X given any hypothesis speaker. That is,

$$p\left(X|\lambda_{\overline{hyp}}\right) = f\left(p(X|\lambda_1), p(X|\lambda_2), \ldots p(X|\lambda_N)\right) \tag{D.4}$$

where f is some function. $\lambda_{\overline{hyp}}$ is also referred as the *background model* and denoted as λ_{bkg}.

The identification procedure is then carried out by the following steps:

1. For every speaker n, calculate

$$\Lambda(X|\lambda_n) = \log p(X|\lambda_n) - \log p(X|\lambda_{bkg}) \tag{D.5}$$

2. Find n which gives the maximum of $\Lambda(X|\lambda_n)$.

3. Apply

$$\Lambda(X|\lambda_n) \begin{cases} \geq \theta & \text{accept speaker n.} \\ < \theta & \text{reject all speakers.} \end{cases} \tag{D.6}$$

D.3 Gaussian Mixture Models (GMM)

The conditional density probability functions of the feature vectors given hypothesis speaker should be estimated. The dominant approach of modeling those probabilities is the *Gaussian mixture models* (GMM).

According to the GMM approach, if $\boldsymbol{x}$ is a feature vector, and λ is a class that represents the hypothesized speaker, then

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{x}) \tag{D.7}$$

Equation (D.7) determines that the conditional density probability of $\boldsymbol{x}$ given λ is a weighted sum of probability densities $p_i(\boldsymbol{x})$ of weight ω_i each. The constraint on ω_i in (D.7) is

$$\sum_{i=1}^{M} \omega_i = 1 \tag{D.8}$$

Every $p_i(\boldsymbol{x})$ in (D.7) is a conditional Gaussian probability density function given sub-class i

$$p_i(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_i)} \tag{D.9}$$

where D is the dimension of $\boldsymbol{x}$, $\boldsymbol{\mu}_i$ is the *mean vector* of sub-class i, and $\boldsymbol{\Sigma}_i$ is its *covariance matrix*. According to (D.9), the model λ in (D.7) can be denoted as [4]:

$$\lambda = \{\omega_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1\ldots M} \tag{D.10}$$

ω_i, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\mu}_i$ are iteratively set by the *expectation maximization* (EM) algorithm [46]. $\boldsymbol{\Sigma}_i$ can be chosen to be either *diagonal* or a *full* matrix. The interpretation of a diagonal covariance matrix is that the feature vector coordinates are independent of one another. The computation of the probability density is much simpler in this case. The advantage

of the full covariance matrix, however, is the greater generalization of the model.

D.4 Distance between Distributions

In speaker recognition applications, every speaker is recognized with a distribution model. Therefore, there needs to be a measure of dissimilarity between distribution models. The concepts in this section are mostly relied on [47] and [48].

Let λ_1 and λ_2 be two distribution models. For J being a dissimilarity between λ_1 and λ_2, it has to fulfill the following conditions:

1. $J = 0$ if the conditional probability density functions are identical, i.e. $p(\boldsymbol{x}|\lambda_1) = p(\boldsymbol{x}|\lambda_2) \quad \forall \boldsymbol{x}$.
2. $J \geq 0$.
3. J attains its maximum when the classes are disjoint, i.e. $\exists \boldsymbol{x}$ such that $p(\boldsymbol{x}|\lambda_1) = 0$ and $p(\boldsymbol{x}|\lambda_2) \neq 0$.

Measures that satisfy these conditions are the *Kolmogorov* variational distance, the *Chernoff* and *Bhattacharyya* dissimilarity measures, which are all given in [47].

In this research most used is the divergence approach, and the *Kullback-Leibler divergence* in specific, which is defined as:

$$D_{KL}(\lambda_1, \lambda_2) = \int p(\boldsymbol{x}|\lambda_1) \log\left(\frac{p(\boldsymbol{x}|\lambda_1)}{p(\boldsymbol{x}|\lambda_2)}\right) \mathrm{d}\boldsymbol{x} \tag{D.11}$$

The *symmetric* Kullback Leibler divergence, is defined as:

$$D_{KL} = \int p(\boldsymbol{x}|\lambda_1) \log\left(\frac{p(\boldsymbol{x}|\lambda_1)}{p(\boldsymbol{x}|\lambda_2)}\right) \mathrm{d}\boldsymbol{x} + \int p(\boldsymbol{x}|\lambda_2) \log\left(\frac{p(\boldsymbol{x}|\lambda_2)}{p(\boldsymbol{x}|\lambda_1)}\right) \mathrm{d}\boldsymbol{x} \tag{D.12}$$

In cases where the distribution is modeled as one Gaussian, an analytic solution for (D.12) is:

$$D_{KL} = \mathrm{Tr}\left\{\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right\} + \mathrm{Tr}\left\{\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right\} - 2S + \mathrm{Tr}\left\{\left(\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1}\right)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\right\} \tag{D.13}$$

where S is the dimension of the feature vectors. $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ represent *mean* and *covariance* matrix of λ_i respectively.

Bibliography

[1] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: a feature-based approach," *IEEE Signal Process. Mag.*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[2] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[3] T. Ganchev, N. Fakotakis, and G. Kokkinakisand, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. SPECOM*, 2005, pp. 191–194.

[4] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[5] N. Mesgarani, M. Slaney, and S. A. Shamma, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 6, pp. 2109–2121, Nov. 2006.

[6] Y. Pan and A. Waibel, "The effects of room acoustics on MFCC speech parameter," in *Proc. ICSLP*, 2000, pp. 129–132.

[7] P. J. Castellano, S. Sridharan, and D. Cole, "Speaker recognition in reverberant enclosures," in *Proc. ICASSP*, 1996, pp. 117–120.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[9] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, 1979, pp. 208–211.

[10] H. E. Ewalt and M. T. Johnson, "Combining multisource wiener filtering with parallel beamformers to reduce noise from interfering talkers," in *Proc. ICSP*, 2004, pp. 455–458.

[11] J. L. Flanagan, "Use of acoustic filtering to control the beamwidth of steered microphone arrays," *J. Acoust. Soc. Am.*, vol. 78, no. 2, pp. 423–428, Aug. 1985.

[12] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, S. Meignier, T. Merlin, J. O. Garcia, I. M. Chagnolleau, D. P. Delacretaz, and D. A. Reynolds, "A tutorial on

text-independent speaker verification," *EURASIP J. Appl. Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[13] C. P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Speech Audio Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[14] Y. Zigel and M. Wasserblat, "How to deal with multiple targets in speaker identification systems?" in *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, 2006, pp. 1–7.

[15] S. P. Kishore and B. Yegnanarayana, "Speaker verification: minimizing the channel effects using autoassociative neural network models," in *Proc. ICASSP*, 2000, pp. 1101–1104.

[16] R. A. Bates and M. Ostendorf, "Reducing the effects of linear channel distortion on continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 594–597, Sep. 1999.

[17] I. R. Titze and W. S. Winholts, "Effect of microphone type and placement on voice perturbation measurements," *J. Speech and Hearing Research*, vol. 36, no. 6, pp. 1177–1190, Dec. 1993.

[18] J. S. Gammal and R. A. Goubran, "Speaker recognition in reverberant environment," *J. Can. Acoust.*, vol. 32, no. 3, pp. 134–135, Sep. 2004.

[19] ——, "Combating reverberation in speaker verification," in *Proc. IMTC*, 2005, pp. 687–690.

[20] J. G. Rodriguet, J. O. Garcia, C. Martin, and L. Hentrindez, "Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays," in *Proc. ICSLP*, 1996, pp. 1333–1336.

[21] J. L. Flanagan, A. C. Surendran, and E. E. Jan, "Spatially selective sound capture for speech and audio processing," *EURASIP J. Speech Commun.*, vol. 13, no. 1–2, pp. 207–222, Oct. 1993.

[22] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum kurtosis subband adaptive filtering," in *Proc. ICASSP*, 2001, pp. 3701–3704.

[23] NIST, "The 1999 speaker recognition evaluation," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 1–18, Jan./Apr./Jul. 2000.

[24] M. Brooks, "Voicebox: Speech processing toolbox for matlab," Imperial College, London, Tech. Rep., 2003, available at `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[25] I. T. Nabney, *Netlab: Algorithms for Pattern Recognition.* New York: Springer, 2002, available at `http://www.ncrg.aston.ac.uk/netlab/`.

[26] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, N.J.: Prentice-Hall, 1978.

[27] Y. Rubner, C. Tomasi, and L. Guibas, "The earth movers distance as a metric for image retrieval," *Int. J. Comp. Vision*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[28] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Chichester, England: Prentice-Hall, 1993.

[29] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, 1984, pp. 328–331.

[30] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Singapore: Academic Press, 1990.

[31] H. Kuttruff, *Room Acoustics*. New York: Spon Press, 2000.

[32] P. A. Nelson and S. J. Elliott, *Active Control of Sound*, 5th ed. London: Academic Press, 2000.

[33] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*. New York: John Wiley, 2000.

[34] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[35] M. R. Schroeder and K. H. Kuttruff, "On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima," *J. Acoust. Soc. Am.*, vol. 34, no. 1, pp. 76–80, Jan. 1962.

[36] W. C. Sabine, *Collected Papers on Acoustics*. Los Altos: Peninsula Publishing, 1993, (Originally 1921).

[37] M. R. Schroeder, "New method of measuring the reverberation time," *J. Acoust. Soc. Am.*, vol. 37, no. 3, pp. 409–412, Mar. 1965.

[38] J. B. Allen and L. R. Rabiner, "A unified approach to short-time fourier analysis and synthesis," *Proc. IEEE*, vol. 65, no. 11, Nov. 1977.

[39] B. Porat, *A Course in Digital Signal Processing*. New York: John Wiley, 1997.

[40] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency aanalysisof time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in *Proc. Symposium on Time Series Analysis*, 1963, pp. 209–243.

[41] S. Subramaniam, A. P. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 392–396, Sep. 1996.

[42] R. A. Kennedy and B. D. Radlovic, "Iterative cepstrum-based approach for speech dereverberation," in *Proc. ISSPA*, 1999, pp. 55–58.

[43] M. Tohyama, R. H. Lyon, and T. Koike, "Source waveform recovery in a reverberant space by cepstrum dereverberation," in *Proc. ICASSP*, 1993, pp. 157–160.

[44] S. S. Stevens, J. E. Volkman, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *J. Acoust. Soc. Am.*, vol. 8, no. 3, pp. 185–190, Aug. 1937.

[45] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[46] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[47] A. R. Webb, *Statistical Pattern Recognition*, 2nd ed. West Sussex, England: John Wiley, 2002, ch. Appendix A.

[48] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Englewood Cliffs, N.J.: Prentice-Hall, 1982.