UPC

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

PhD Thesis

Channel Selection and Reverberation-Robust Automatic Speech Recognition

Author: Martin Wolf
Advisor: Prof. Climent Nadeu

TALP Research Center

Department of Signal Theory and Communication

Universitat Politècnica de Catalunya

Barcelona, Spain

October 2013

Túto prácu venujem mojej rodine.

Abstract

If speech is acquired by a close-talking microphone in a controlled and noise-free environment, current state-of-the-art recognition systems often show an acceptable error rate. The use of close-talking microphones, however, may be too restrictive in many applications. Alternatively, distant-talking microphones, often placed several meters far from the speaker, may be used. Such setup is less intrusive, since the speaker does not have to wear any microphone, but the Automatic Speech Recognition (ASR) performance is strongly affected by noise and reverberation. The thesis is focused on ASR applications in a room environment, where reverberation is the dominant source of distortion, and considers both single- and multi-microphone setups.

If speech is recorded in parallel by several microphones arbitrarily located in the room, the degree of distortion may vary from one channel to another. The difference among the signal quality of each recording may be even more evident if those microphones have different characteristics: some are hanging on the walls, others standing on the table, or others build in the personal communication devices of the people present in the room. In a scenario like that, the ASR system may benefit strongly if the signal with the highest quality is used for recognition. To find such signal, what is commonly referred as Channel Selection (CS), several techniques have been proposed, which are discussed in detail in this thesis.

In fact, CS aims to rank the signals according to their quality from the ASR perspective. To create such ranking, a measure that either estimates the intrinsic quality of a given signal, or how well it fits the acoustic models of the recognition system is needed. In this thesis we provide an overview of the CS measures presented in the literature so far, and compare them experimentally. Several new techniques are introduced, that surpass the former techniques in terms of recognition accuracy and/or computational efficiency. A combination of different CS measures is also proposed to further increase the recognition accuracy, or to reduce the computational load without any significant performance loss. Besides, we show that CS may be used together with other robust ASR techniques, like matched condition training or mean and variance normalization, and that the recognition improvements from both approaches are cumulative up to some extent. An online real-time version of the channel selection method based on the variance of the speech sub-band envelopes, which was developed in this thesis, was designed and implemented in a smart room environment. When evaluated in experiments with real distant-talking microphone recordings and with moving speakers, a significant recognition performance improvement was observed.

Another contribution of this thesis, that does not require multiple microphones, was developed in cooperation with the colleagues from the chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg, Erlangen, Germany. It deals with the problem of feature extraction within REMOS (REverberation MOdeling for Speech recognition), which is a generic framework for robust distant-talking speech recognition. In this framework, the use of conventional methods to obtain decorrelated feature vector coefficients, like the discrete cosine transform, is constrained by the inner optimization problem of REMOS, which may become unsolvable in a reasonable time. A new feature extraction method based on frequency filtering was proposed to avoid this problem.

Resum

Els sistemes actuals de reconeixement de la parla mostren sovint una taxa d'error acceptable si la veu és registrada amb micròfons pròxims a la boca del parlant, en un entorn controlat i lliure de soroll. No obstant, l'ús d'aquests micròfons pot ser massa restrictiu en moltes aplicacions. Alternativament, es poden utilitzar micròfons distants, els quals sovint són ubicats a diversos metres del parlant. Aquesta configuració és menys intrusiva, ja que el parlant no ha de portar a sobre cap micròfon, però el rendiment del reconeixement automàtic de la parla (ASR, de l'anglès Automatic Speech Recognition) en aquest cas es veu fortament afectat pel soroll i la reverberació. Aquesta tesi s'enfoca a aplicacions ASR en un ambient de sala, on la reverberació és la causa predominant de distorsió i es considera tant el cas d'un sol micròfon com el de múltiples micròfons.

Si la parla és gravada en paral·lel per diversos micròfons distribuïts arbitràriament a la sala, el grau de distorsió pot variar d'un canal a l'altre. Les diferències en qualitat entre els senyals enregistrats poden ser més accentuades si els micròfons tenen diferents característiques i col·locacions: uns a les parets, altres sobre la taula, o bé altres integrats en els aparells de comunicació de les persones presents a la sala. En un escenari com aquest, el sistema ASR es pot beneficiar enormement de l'utilització del senyal de més qualitat per al reconeixement. Per a trobar aquest senyal s'han proposat diverses tècniques, anomenades CS (de l'anglès Channel Selection), les quals es discuteixen detalladament en aquesta tesi.

De fet, la selecció de canal busca ordenar els senyals conforme a la seva qualitat des de la perspectiva ASR. Per crear tal rànquing es necessita una mesura que estimi la qualitat intrínseca d'un senyal, o bé una que valori com de bé aquest s'ajusta als models acústics del sistema de reconeixement. En aquesta tesi proporcionem un resum de les mesures CS fins ara presentades en la literatura, comparant-les experimentalment. A més, es presenten diverses noves tècniques que superen les anteriors en termes d'exactitud de reconeixement i / o eficiència computacional. També es proposa una combinació de diferents mesures CS amb l'objectiu d'incrementar l'exactitud del reconeixement, o per reduir la càrrega computacional sense cap pèrdua significativa de rendiment. A més mostrem que la CS pot ser utilitzada juntament amb altres tècniques robustes d'ASR, com ara matched condition training o la normalització de la variança i la mitjana, i que les millores de reconeixement de les dues aproximacions són fins a cert punt acumulatives. Una versió online en temps real del mètode de selecció de canal basat en la variança de les envolvents sub-banda de la parla, desenvolupada en aquesta tesi, va ser dissenyada i implementada en una sala intel·ligent. A l'hora d'avaluar experimentalment gravacions reals de micròfons no pròxims a la boca amb parlants en moviment, es va observar una millora significativa en el rendiment del reconeixement.

L'altra contribució d'aquesta tesi, que no requereix múltiples micròfons, va ser desenvolupada en col·laboració amb els col·legues del departament de Comunicacions Multimedia i Processament de Senyals de la Universitat de Erlangen-Nuremberg, Erlangen, Alemanya. Tracta sobre el problema d'extracció de característiques a REMOS (de l'anglès REverberation MOdeling for Speech recognition). REMOS és un marc conceptual genèric per al reconeixement robust de la parla amb micròfons llunyans. L'ús dels mètodes convencionals per obtenir els elements decorrelats del vector de característiques, com ara la transformada cosinus discreta, està limitat pel problema d'optimització inherent a REMOS. Aquest faria que, utilitzant les eines convencionals, es tornés un problema irresoluble en un temps raonable. Per resoldre aquest problema hem desenvolupat un nou mètode d'extracció de característiques basat en filtrat frecuencial.

Resumen

Los actuales sistemas de reconocimiento del habla muestran a menudo una tasa de error aceptable si la voz es registrada por micrófonos próximos a la boca del hablante, en un entorno controlado y libre de ruido. Sin embargo, el uso de estos micrófonos puede ser demasiado restrictivo en muchas aplicaciones. Alternativamente, se pueden emplear micrófonos distantes, los cuales a menudo se ubican a varios metros del hablante. Esta configuración es menos intrusiva ya que el hablante no tiene que llevar encima ningún micrófono, pero el rendimiento del reconocimiento automático del habla (ASR, del inglés Automatic Speech Recognition) en dicho caso se ve fuertemente afectado por el ruido y la reverberación. Esta tesis se enfoca a aplicaciones ASR en el entorno de una sala, donde la reverberación es la causa predominante de distorsión y se considera tanto el caso de un solo micrófono como el de múltiples micrófonos.

Si el habla es grabada en paralelo por varios micrófonos distribuidos arbitrariamente en la sala, el grado de distorsión puede variar de un canal a otro. Las diferencias de calidad entre las señales grabadas pueden ser más acentuadas si dichos micrófonos muestran diferentes características y colocaciones: unos en las paredes, otros sobre la mesa, u otros integrados en los dispositivos de comunicación de las personas presentes en la sala. En dicho escenario el sistema ASR se puede beneficiar enormemente de la utilización de la señal con mayor calidad para el reconocimiento. Para hallar dicha señal se han propuesto diversas técnicas, denominadas CS (del inglés Channel Selection), las cuales se discuten detalladament en esta tesis.

De hecho, la selección de canal busca ranquear las señales conforme a su calidad desde la perspectiva ASR. Para crear tal ranquin se necesita una medida que tanto estime la calidad intrínseca de una señal, como lo bien que ésta se ajusta a los modelos acústicos del sistema de reconocimiento. En esta tesis proporcionamos un resumen de las medidas

CS hasta ahora presentadas en la literatura, comparándolas experimentalmente. Diversas nuevas técnicas son presentadas que superan las técnicas iniciales en cuanto a exactitud de reconocimiento y/o eficiencia computacional. También se propone una combinación de diferentes medidas CS para incrementar la exactitud de reconocimiento, o para reducir la carga computacional sin ninguna pérdida significativa de rendimiento. Además mostramos que la CS puede ser empleada junto con otras técnicas robustas de ASR, tales como matched condition training o la normalización de la varianza y la media, y que las mejoras de reconocimiento de ambas aproximaciones son hasta cierto punto acumulativas. Una versión online en tiempo real del método de selección de canal basado en la varianza del speech sub-band envelopes, que fue desarrolladas en esta tesis, fue diseñada e implementada en una sala inteligente. Reportamos una mejora significativa en el rendimiento del reconocimiento al evaluar experimentalmente grabaciones reales de micrófonos no próximos a la boca con hablantes en movimiento.

La otra contribución de esta tesis, que no requiere múltiples micrófonos, fue desarrollada en colaboración con los colegas del departamento de Comunicaciones Multimedia y Procesamiento de Señales de la Universidad de Erlangen-Nuremberg, Erlangen, Alemania. Trata sobre el problema de extracción de características en REMOS (del inglés REverberation MOdeling for Speech recognition). REMOS es un marco conceptual genérico para el reconocimiento robusto del habla con micrófonos lejanos. El uso de los métodos convencionales para obtener los elementos decorrelados del vector de características, como la transformada coseno discreta, está limitado por el problema de optimización inherente a REMOS, lo que haría que, utilizando las herramientas convencionales, se volviese un problema irresoluble en un tiempo razonable. Para resolver este problema hemos desarrollado un nuevo método de extracción de características basado en filtrado frecuencial.

Acknowledgements

Foremost I want to thank Climent Nadeu, who provided me the guidance during this quest for knowledge. I will never forget his kindness, patience, encouragement and support that strongly motivated me to finish this work, despite many difficulties along the way. I feel very grateful and fortunate to be able to spent this time under his tutorship.

This thesis would not be written without the support of my family, my parents Dana and Pavel, and sister Jana, who have always been here for me. Thank you from the bottom of my heart for all your love. I also thank Sandrine for giving me the strength to take the final steps and to write this thesis.

I am especially grateful to Gregor Rozinaj, a friend and professor from my former university in the city of Bratislava, who stood at the beginning of this adventure.

During this time I had an opportunity to meet and to collaborate with amazing people. My gratitude goes to Walter Kellerman from the University of Erlangen-Nuremberg for giving me the opportunity to spent 4 months in his research group, and also to Roland Maas and Armin Sehr for a fruitful collaboration during that time. Thanks to the fellow doctorates, friends and colleagues from the department, for creating simulating and enjoyable research environment, especially Zelky, Adolfo, David, Carlos H., Henrik, Taras, Andrey, Carlos S., and Jordi. I also want to thank Diego Lendoiro for his help with the implementation in the smart room and Marc René Schädler from Carl von Ossietzky University of Oldenburg for many inspiring discussions.

I finally want to express gratitude to all my friends here in Spain, with whom I have spent unforgettable times climbing in the mountains. Thanks to you, Catalunya feels like another home to me.

Martin Wolf

June 2013

Acronyms

ASR	Automatic Speech Recognition
CHIL	Computers in the Human Interaction Loop
CS	Channel Selection
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DRR	Direct-to-Reverberant Ratio
EV	Envelope Variance
FBE	Filter-Bank Energies
FF	Frequency Filtering
FIR	Finite Impulse Response
GMM	Gaussian Mixture Model
HFA	Harmonicity-based Feature Analysis
HMM	Hidden Markov Model
HTK	Hidden Markov Model ToolKit
ICSI	International Computer Science Institute
IO	Input Output
LDA	Linear Discriminant Analysis
LP	Linear Prediction
LVCSR	Large Vocabulary Continuous Speech Recognition

MAP Maximum A Posteriori

MCCC Multichannel Cross-Correlation Coefficient

MFCC Mel-Frequency Cepstral Coefficients

MLLR Maximum-Likelihood Linear Regression

MRD Meeting Recorder Digit

MVN Mean and Variance Normalization

NIST National Institute of Standards and Technology

NMF Non-negative Matrix Factorization

PCA Principal Component Analysis

PCC Pan Correct Classification

PCCR Pan Correct Classification within a Range

PMAE Pan Mean Average Error

PMC Parallel Model Combination

RASTA RelAtive SpecTrAl processing

REMOS REverberation MOdeling for Speech recognition

RIR Room Impulse Response

RVM Reverberation Model

SMD Speaker to Microphone Distance

SNR Signal to Noise Ratio

STFT Short-Time Fourier Transform

UPC Universitat Politècnica de Catalunya

VTLN Vocal Tract Length Normalization

VTS Vector Taylor Series

WER Word Error Rate

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Computers and computer based devices have become an important part of our world. There are many tasks in our private and professional life that we can not imagine without them any more, and many others may be possible in the future. While in the past we were adapting to the machines and users had to spend significant time and effort to learn how to operate them, today the trend is to make this interaction more intuitive. Although we have become very skilled in using a keyboard, mouse or touch screen, there are many situations where it would be more convenient to use our voice.

Speech recognition research has been active for more than a half century. Despite the effort, there are not many widespread applications that people consider really useful and helpful in their day-to-day life. The main reason why speech recognition has not succeeded as it could be expected, although speech is the most natural way of communication for humans, is the complexity of the problem. Computer commands are simple, clear and unambiguous, and we give them through different specialized input devices. With speech, however, it is not so simple. Each person not only speaks differently, but even the same word from the same speaker may sound different the second time he or she says it. The speech signal contains a lot of information that is not related to the message, and to further complicate the problem, it may get distorted by other sounds coming from the environment. To recognize what was said and transcribe it to text is the aim of Automatic Speech Recognition (ASR).

Human-computer interaction is not the only possible ASR application. There are dictation and meeting transcription, audio captioning and indexing, voice-based search, or automatic translation of spoken language. Integration of many technologies from different fields is required to provide those services, but the correct speech transcription is the key component for all of them. It is therefore important to develop robust and reliable ASR systems. The fact that in our day-to-day lives we do not use as many ASR applications as may be possible suggests we have not achieved this yet.

State-of-the-art recognition systems often deliver satisfactory results in a noise-free and controlled environment where the recordings are made with a close-talking microphone. This kind of microphone is held or attached close to the mouth of the speaker, so the desired speech signal is much stronger than the interfering noises. The problem is, that such close-talking recording is not often possible or practical.

An alternative is to record the speech signal with a distant-talking microphone, which may be placed independently of the position and orientation of the speaker, often several meters far away. When the speech signal leaves the speaker's mouth it undergoes several transformations before it reaches the far microphone. First of all, the sound wave is attenuated as it propagates through the air. If the microphone is placed behind the speaker, attenuation by the head of the speaker also needs to be taken into consideration. Further distortion may be caused by different background noises, speech from overlapping speaker, or reverberation. Consequently, recognition accuracy (i.e. number of correctly recognized words) of the ASR system decreases.

In this thesis we focus on a speech recognition in a room using distant-talking microphones. A typical distant speech recognition scenario is shown in Figure 1.1. Additive (background) noise is any additive sound other than that of interest. It can be stationary, such as that caused by an air conditioning or a computer fan, or non-stationary, like door slams, keyboard clicks, interfering speaker or music. Speech signal and additive noises generated in the environment come from different sources and are usually uncorrelated. Reverberation, on the other hand, is highly correlated with speech. It is a phenomenon that occurs in enclosed spaces when the microphone does not pick up only the direct acoustic wave, but also the waves reflected by the walls and objects.

Many techniques increasing robustness of the ASR system to the distortion caused by additive interferences and reverberation using a single microphone have been developed over the years [1]. But, as illustrated in Figure 1.1, we do not have to limit ourselves to a single microphone setup. Actually, distant speech recognition can largely benefit from the use of multiple microphones, either by applying some of many multi-channel based noise reduction techniques [2–5] or in case of microphone arrays, by using beamforming [6, 7].

We may also encounter situations, where multi-microphone processing is practically inevitable if we want to achieve a good result. Imagine a scenario where several participants in a meeting have a laptop or some personal communication device with integrated microphone in front of them. Our objective is to transcribe that meeting

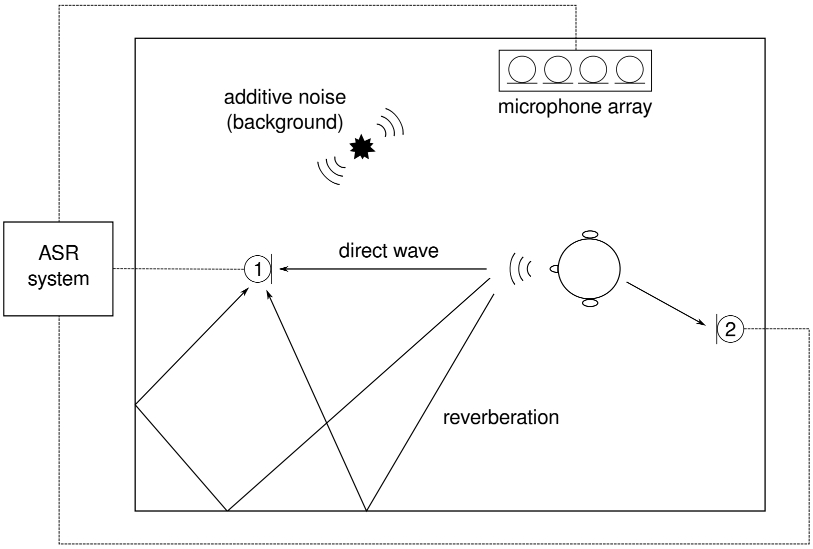


Figure 1.1: A typical distant speech recognition scenario.

using ASR. Assume, we are able to collect signals from all the microphones that are recording the sound. If we used only one microphone to recognize speech from all participants, the result would probably not be very good. On the other hand, if we take the advantage of having several microphones in the room, some possibly located closer to an active speaker, the recognition performance could be improved.

Of course, the described scenario presents a lot of challenges. Although a simple selection of the best microphone may provide significant increase in recognition accuracy, some additional methods will be needed to match the excellent speech recognition capabilities of humans. However, if there are several recordings of the same utterance available, each of them of different quality, it would be unreasonable not to use the best one for further processing. The methods that do this selection automatically and optimally from the point of view of the ASR system are investigated in this thesis. Follows a more detailed list of contributions.

1.1 Contributions of this Thesis

A less complex scenario than the aforementioned case was assumed during the development of this thesis. We focused mainly on the reverberation distortion, which is only one kind of distortion that may be encountered in reality. Nevertheless, reverberant environments present one of the biggest challenges for ASR. Contributions were made for both single and multi-microphone setups.

For setups with multiple microphones we discuss the Channel Selection (CS) problem. A detailed overview of the CS methods reported in the literature is provided. Almost all methods were implemented and compared experimentally on the same task, what has not been done so far. Several novel approaches were developed, including a very simple technique that matches or outperforms other CS methods in reverberant environments both, in terms of recognition performance of the ASR system, and computational complexity. We also show that different CS methods may be combined to further improve the recognition rate, or to reduce the computational cost without significant performance lost.

To develop this work a multi-microphone Room Impulse Response (RIR) database was recorded at the Universitat Politècnica de Catalunya (UPC) in the smart room, and was made publicly available. It may be used not only for CS experiments, but also to test other reverberation-related speech recognition problems. For the same room, a real-time CS client using one of the developed techniques was implemented. This client may be easily integrated with other speech recognition technologies.

Another topic discussed in this thesis is feature extraction within the REverberation MOdeling for Speech recognition (REMOS) framework [8], which is a generic framework for robust distant-talking speech recognition usually applied in single-microphone scenarios. This topic was developed in cooperation with the colleagues from the chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg, Erlangen, Germany, where REMOS was designed. In REMOS the use of conventional methods to obtain decorrelated feature vector coefficients, like the Discrete Cosine Transform (DCT), is not possible because the optimization problem, central to REMOS, becomes unsolvable in a reasonable time. A new feature extraction method based on Frequency Filtering (FF) [9] that fits well into the framework and partly decorrelates the coefficients was proposed. When tested in different reverberation scenarios, consistent recognition error reduction was observed compared to the previous implementation, which was using the highly correlated logarithmic mel-spectral features.

1.2 Thesis Structure

This thesis is organized as follows.

Chapter 2 provides an overview of the ASR concepts relevant to this work. We describe the basic building blocks of the Hidden Markov Model (HMM)-based ASR system. Then we demonstrate the effects of reverberation on the recognition performance and overview state-of-the-art approaches for reverberation-robust speech recognition.

Channel selection is introduced in Chapter 3. We classify and review the CS measures reported in the literature so far, as well as describe new methods developed in this thesis. In Chapter 4 the techniques are tested in different reverberant scenarios, using either signals convolved with measured RIRs, or real distant-talking microphone recordings. In the same chapter the combination of different CS measures is proposed, to further increase the recognition accuracy, or to reduce the computational load without any significant performance loss.

In Chapter 5 the integration of the new feature extraction method into REMOS is discussed. We briefly describe the REMOS framework to the extent necessary to understand the optimization-related problem, present the novel solution, and report the experimental results.

Finally, Chapter 6 summarizes the mayor contributions of this thesis and highlights some possible directions for future work.

The multi-microphone RIR database and the implementation of the real-time CS client in the UPC smart room are described at the end in Appendix A and B.

Chapter 2

A Review of Reverberation-Robust ASR

This thesis was developed mainly in the context of the UPC smart room, where distant-talking microphones are used to record the speech signal. Although both, additive noise and reverberation, are present in such environment, the latter is usually more dominant and challenging to cope with. The techniques described in this work are therefore primarily designed to deal with the reverberation, and the following discussion is for the most part focused on this problem[1].

The objective in this work is to improve the performance of ASR. We limit our discussion to HMM-based systems and only briefly describe the most important components for understanding the following chapters. Reader interested in more comprehensive introduction may refer to [10–13]. The chapter starts with a formulation of the speech recognition problem. Then we describe two feature extraction methods which were used in our work, and outline the very basics of HMM-based acoustic modeling. Follows a discussion about the reverberation. We explain this phenomenon and show its influence on the ASR performance. An overview of existing strategies for reverberation robust ASR is provided at the end.

2.1 HMM-Based ASR

ASR is a pattern classification task. The aim is to process unknown speech signal, i.e. an utterance, on the input and transcribe it into the sequence of words on the output. When speaking, we encode the message as a sequence of sounds. Speech recognition system models these sounds by distinct classes and tries to estimate their

[1] However, several CS methods (mainly the decoder-based ones) presented in Section 3 do not assume any specific kind of distortion.

correct sequence.

Figure 2.1 shows the function blocks and data flow of a typical ASR system. When building a recognizer, the language and acoustic models are first trained using text and annotated speech databases. Speech signal contains a lot of redundant information that is not useful for decoding the message, so the recorded signal is usually not used directly for classification. In conventional recognition systems, the continuous speech waveform is first converted into a sequence of overlapped frames, and from each frame a vector of descriptive parameters, also called features or observations, is extracted.

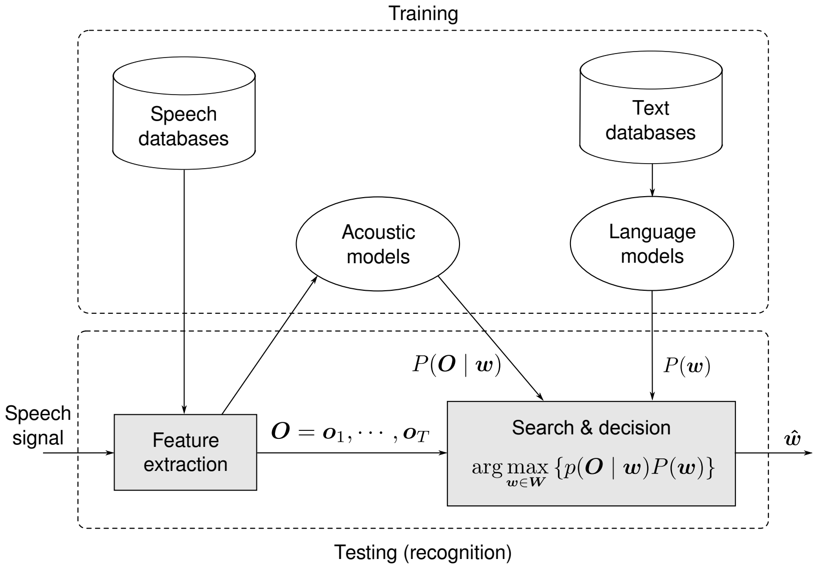

Training
Speech databases
Text databases
Acoustic models
Language models
$P(\boldsymbol{O} \mid \boldsymbol{w})$
$P(\boldsymbol{w})$
Speech signal
Feature extraction
$\boldsymbol{O} = \boldsymbol{o}_1, \cdots, \boldsymbol{o}_T$
Search & decision
$\arg\max_{\boldsymbol{w} \in \boldsymbol{W}} \{p(\boldsymbol{O} \mid \boldsymbol{w}) P(\boldsymbol{w})\}$
$\hat{\boldsymbol{w}}$
Testing (recognition)

Figure 2.1: Block diagram of a speech recognition system with the data flow between training and recognition.

Let $\boldsymbol{O} = \boldsymbol{o}_1, \cdots, \boldsymbol{o}_T$ be the sequence of observation vectors, where $\boldsymbol{o}_t$ is a vector observed at time t, and $\boldsymbol{w} = w_1, \ldots, w_N$ be a sequence of words from all possible word sequences $\boldsymbol{W}$ which recognition system can hypothesize. The speech recognition problem may then be expressed as finding the word sequence $\hat{\boldsymbol{w}}$, which maximizes the posterior probability given the sequence of observation vectors $\boldsymbol{O}$:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \boldsymbol{W}} \{P(\boldsymbol{w} \mid \boldsymbol{O})\}. \tag{2.1}$$

This probability is not computed directly, but using the Bayes' rule, Equation (2.1) is converted into the product of the likelihood $p(\boldsymbol{O} \mid \boldsymbol{w})$ and the prior probability $P(\boldsymbol{w})$, normalized by the evidence $p(\boldsymbol{O})$ as

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w} \in \boldsymbol{W}} \left\{ \frac{p(\boldsymbol{O} \mid \boldsymbol{w}) P(\boldsymbol{w})}{p(\boldsymbol{O})} \right\}. \tag{2.2}$$

To estimate the probability of the observation vector $p(\boldsymbol{O})$ in the denominator may be very costly and it is usually omitted, since the selected sequence of words does not depend on it, resulting in

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}\in\boldsymbol{W}} \left\{ p(\boldsymbol{O} \mid \boldsymbol{w}) P(\boldsymbol{w}) \right\}. \tag{2.3}$$

2.1.1 Feature Extraction

Apart from the message, speech signal carries a lot of details that are not needed to decode the content, such as pitch, accent, emotional state, speaking rate, and also distorting effects from the channel or environment. The motivation for feature extraction is to remove irrelevant information from the signal and to reduce the complexity of the classification problem.

Over the years many feature extraction methods have been developed, most of them using the same underlying principle, where a single feature vector is extracted from the envelope of the time-frequency pattern in the Short-Time Fourier Transform (STFT) domain. Although there have been attempts to change this paradigm [14], STFT-based features extracted in the frame-by-frame fashion remain dominant in current ASR systems. The systems used is this thesis also employ this kind of features, namely Mel-Frequency Cepstral Coefficients (MFCC) [15], which are considered a de-facto standard, and a variant of them, called FF [9, 16], which were used as a basis for the new feature extraction method for the REMOS concept (see Section 5).

In Figure 2.2 a flow chart showing the generation of MFCC and FF features is shown. As may be seen, the first steps are the same for both methods. The incoming speech signal is divided into a sequence of overlapping frames (usually 20-30 ms long with an overlap 10 ms). Then, each frame is windowed and transformed into the frequency domain using the Discrete Fourier Transform (DFT). The square of the magnitude of the DFT coefficients is taken and multiplied by a series of overlapping triangular weighting functions following the mel scale. These filters are imitating the human auditory system by having better resolution in the low frequencies than in the high ones. Mel-spectral (melspec) coefficients are computed as the energy in each of the mel filters. Human auditory system is further approximated when dynamic range of melspec energies is compressed by applying the logarithm, resulting into the so-called log mel-spectral (logmelspec) coefficients.

As discussed later, the acoustic model is based on HMMs, which are probabilistic finite state machines. The output probability distribution functions in each state are usually modeled as Gaussian mixtures with diagonal covariance matrices. This

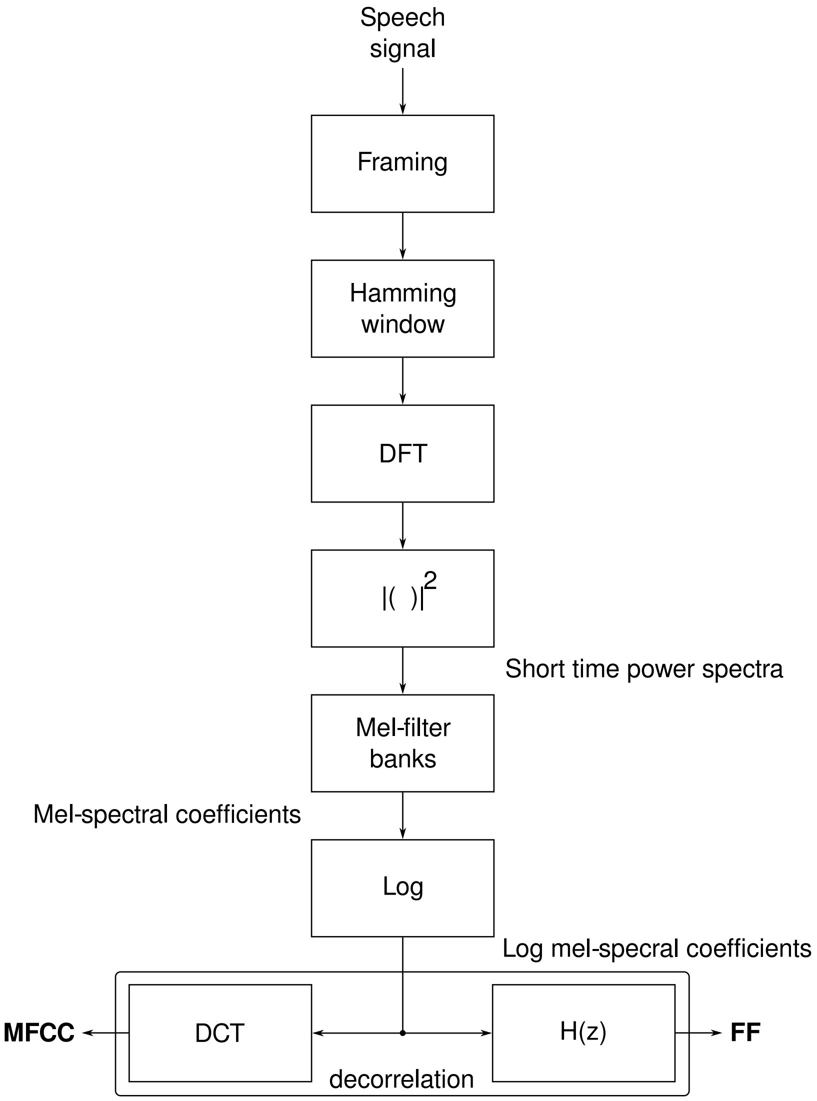


Figure 2.2: Block diagram of the feature extraction for MFCC and FF.

implies, that the elements of the observation vectors are uncorrelated, which is not true in case of the logmelspec coefficients. To obtain a set of decorrelated parameters, a linear transformation is usually applied on the logmelspec features. At this point, MFCC and FF computation differ.

MFCC

In case of MFCC, the DCT is applied to decorrelate the spectral parameters. Doing this transformation, we actually perform the frequency analysis of the logmelspec features for a given frame, convert them to so called cepstral domain, and obtain cepstral coefficients. Usually only a set of coefficients, ranging from 12 to 20 is taken for recognition [12]. Low-order coefficients contribute more to the class separability than the high-order ones [17] and are less prone to noise. The 0-th MFCC coefficient represents the sum of logmelspec coefficients (energies) of the analyzed frame and is often replaced by more robust measures, e.g. the log energy of the frame.

FF

Decorrelation in case of FF is achieved by filtering the sequence of logmelspec energies in the spectral domain by a simple Finite Impulse Response (FIR) filter. This filter is designed in such way, that the variance of the cepstral coefficients is equalized. The usual filter, proposed in the original work [16], has a transfer function

$$H(z) = z - z^{-1}. \tag{2.4}$$

The impulse response of this filter is $h(k) = \{1, 0, -1\}$, so the filtering operation consists of a subtraction of the two bands adjacent to the current one. During filtering zeros are assumed before the first and after the last logmelspec coefficient. Therefore, after the filtering operation, each endpoint contains the energy of its neighboring band. Often, full frame energy is added to the feature vector, so these endpoint coefficients may be removed, since they become redundant. Unlike MFCC the FF coefficients remain in the frequency domain, so their physical interpretation is more intuitive.

It has been shown in [18] that information about temporal evolution of the spectral coefficients increases the discrimination of phonemes. The dynamic features, that are the first and second time derivatives of the static MFCC or FF coefficients are therefore usually included into the feature vector.

2.1.2 Acoustic Modeling and Decoding

The exact computation of the joint conditional probability $p(\boldsymbol{O} \mid \boldsymbol{w})$ and the prior probability $P(\boldsymbol{w})$ is not practicable. In the real world systems they are approximated by the scores determined by the acoustic and language models respectively, as shown in Figure 2.1. The language model is an important, often indispensable, part of the ASR system. However, in this thesis we tackle the acoustic distortion problem, so language modeling is omitted in further discussion.

In most state-of-the-art ASR systems HMMs are used to represent the acoustic model. Words (or smaller units, e.g. phonemes) are typically modeled as a sequence of states $\boldsymbol{s} = s_1, \ldots, s_{n_s}$, where each state represents a small part of a word. The problem of finding the word sequence that maximizes the posterior probability in

Equation (2.1) is then equivalent to a problem of finding the best state sequence $\hat{\boldsymbol{s}}$ from the set of all possible sequences $\boldsymbol{S}$ as

$$\begin{aligned} \hat{\boldsymbol{s}} &= \arg\max_{\boldsymbol{s}\in\boldsymbol{S}} \{P(\boldsymbol{s} \mid \boldsymbol{O})\} \\ &= \arg\max_{\boldsymbol{s}\in\boldsymbol{S}} \{P(s_1, \ldots, s_{n_s} \mid o_1, \ldots, o_n)\} \,. \end{aligned} \tag{2.5}$$

Applying the Bayes' rule and omitting the denominator we get

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}\in\boldsymbol{S}} \{P(o_1, \ldots, o_n \mid s_1, \ldots, s_n) P(s_1, \ldots, s_n)\} \,. \tag{2.6}$$

The evaluation of this equation may be simplified by making the following assumptions:

- A first-order Markov process is assumed, so the probability of the current state s_i depends only upon the previous state s_{i-1}.
- Assuming the state-conditional independence, the current observation vector o_i depends only on the current state s_i and does not depend on the previous ones.

Under these assumptions Equation (2.3) becomes

$$\hat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}\in\boldsymbol{S}} \left\{ \prod_{i=1}^{n} P(o_i \mid s_i) \prod_{i=1}^{n} P(s_i \mid s_{i-1}) P(s_i) \right\} . \tag{2.7}$$

Two very effective algorithms are used to implement the modeling in practice. The first one is the Baum-Welch algorithm [19, 20] which is applied in training to find the HMM parameters that best fit the models to the annotated speech data in terms of maximum likelihood. The second one is the Viterbi algorithm [21] used in testing to find the most likely sequence of states, hence words, by effectively solving Equation (2.7).

2.2 ASR in Reverberant Environments

The acoustic model represents the training data. The key problem in robust speech recognition comes from the fact that ASR systems assume to observe similar data in testing as in training, which is rarely the case in reality. A speech recognition system trained on clean speech can hardly perform well if the speech signal is distorted, unless some measures are taken to reduce the mismatch. In enclosed spaces when recording is made using distant-talking microphones, reverberation is the main source of distortion. It is created when acoustic waves reflected by the walls and objects arrive to the microphone attenuated and with different delays, introducing undesirable and unpredictable interferences. The reverberation distortion is usually modeled through the linear convolution between the RIR $h(n)$ and the clean speech signal $x(n)$ as

$$y(n) = h(n) * x(n). \tag{2.8}$$

The linear distortion from the acoustic channel can not be easily canceled or attenuated in the feature domain as it is routinely done in ASR for the linear distortions produced in the electric channel (microphone, amplifiers, telephone network, etc.), since the duration of the RIR is usually much longer than the electrical channel impulse responses and encompasses several consecutive phones.

An example of the RIR measured in the UPC smart room may be seen in Figure 2.3. It may be split into 3 parts. The first dominant peak corresponds to the direct wave. This wave arrives to the microphone after the initial delay which can be derived from the Speaker to Microphone Distance (SMD) and the sound velocity. Follows the interval up to 50 100 ms of early reflections and the response ends with a dense reverberation tail (late reflections). Typical acoustic measures used to characterize the level of reverberation are the reverberation time T_{60} and the "definition" D_{50} ("Deutlichkeit") [22]. The reverberation time is defined as the time needed for the sound level to decay by 60 dB after the sound source is turned off. "Deutlichkeit" is defined as the ratio of the energy of the direct sound and early reflections arriving within the first 50 ms to the energy of the whole RIR, i.e.,

$$D_{50} = \frac{\int_0^{50ms} h^2(t)dt}{\int_0^{\infty} h^2(t)dt}. \tag{2.9}$$

It is primarily used as an objective measure of speech intelligibility, but it has also been applied to predict the ASR accuracy [23, 24].

Reverberation smears the speech signal over time, as phonemes are masked by the energy from the previous speech segments. Consequently, the phonetic information is degraded. This degradation is difficult to estimate, because resulting effect depends on both, the acoustic channel characteristic (which varies with position and orientation of the speaker, position of the objects, or temperature in the room), as well as on the uttered speech signal.

In Figure 2.4 we illustrate this on the spectrogram of the utterance "one nine zero seven", which was recorded with the close-talking microphone and reverberated using the RIR from Figure 2.3. The word and phoneme boundaries can be identified more easily in the close-talking case. Due to reverberation, energy from strong vowels is

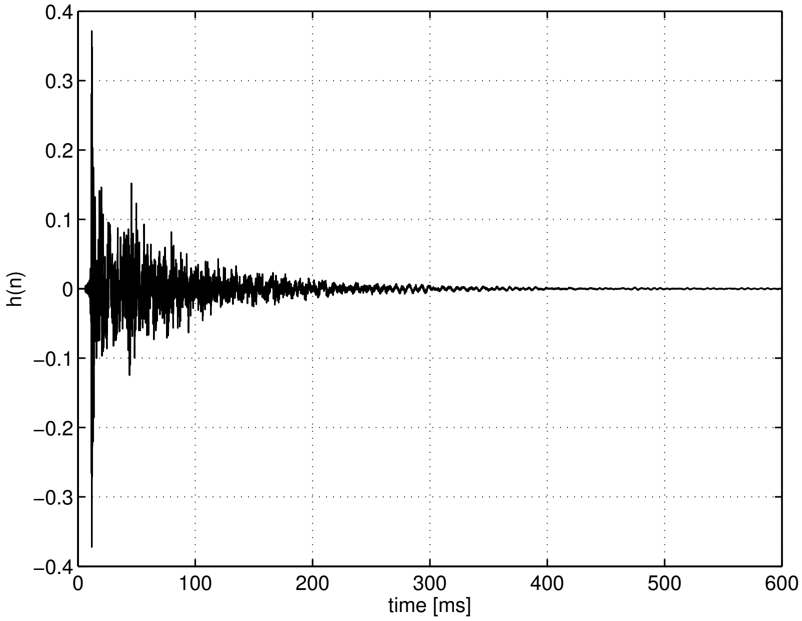


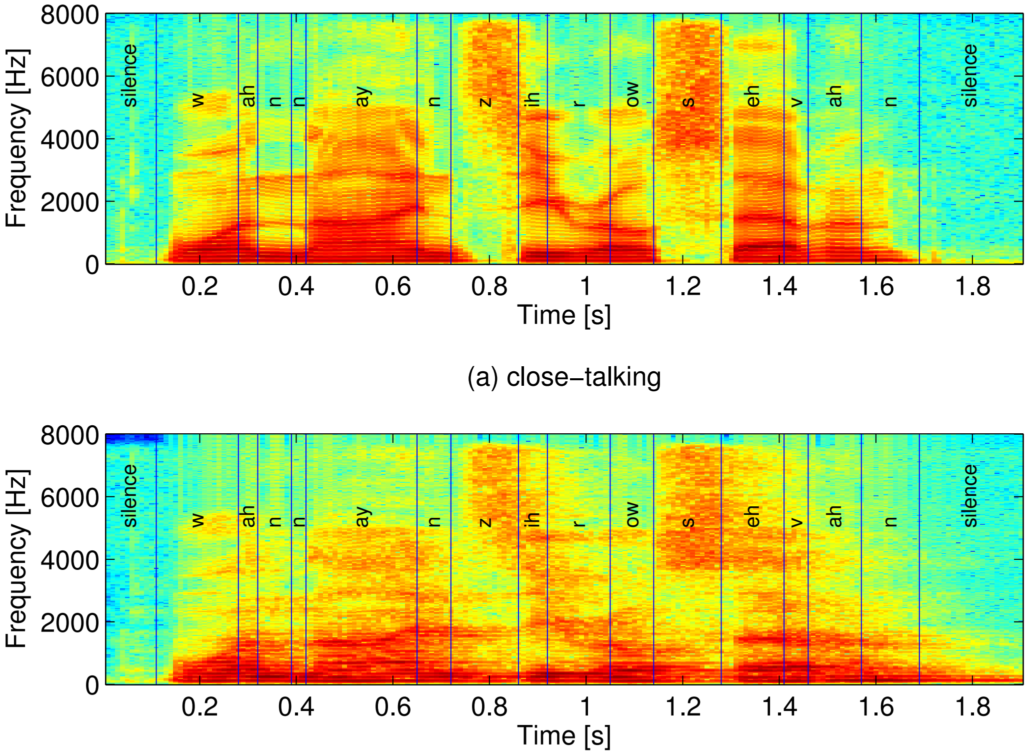
Figure 2.3: The RIR measured in the UPC smart room with reverberation time around 500 ms and with the SMD 1.4 m.

masking weaker consonants. This may be clearly seen in segments around 0.8 and 1.2 s. These segments correspond to the fricative consonants /z/ and /s/, and are filled up in the low frequency bands with energy from preceding voiced sounds /ay/ and /ow/. Also, the formants appear flattened, and their transitions are less pronounced.

Interestingly, not all parts of the RIR harm the ASR performance equally. In [25] effects of different parts of the RIR on the ASR accuracy were investigated and it was shown that early reflections and the late tail are not harming the speech recognition. On the other hand, the middle part between approximately 70 ms and 2/3 of T_{60} was identified as the most harming one. Also, the low frequency reverberations approximately between 250 and 2500 Hz disturb speech recognition more than the higher ones.

2.2.1 The Effects of Reverberation on the ASR Performance

We have defined reverberation and described how it affects the speech signal. In the following we will demonstrate on a Large Vocabulary Continuous Speech Recognition (LVCSR) task, how these signal distortions influence the ASR performance. Experiment was made using the RWTH Aachen University open source ASR system [26], and Catalan Speecon and FreeSpeech databases. For training, approximately 121 hours of data from both databases were selected, using only close-talking



(a) close−talking

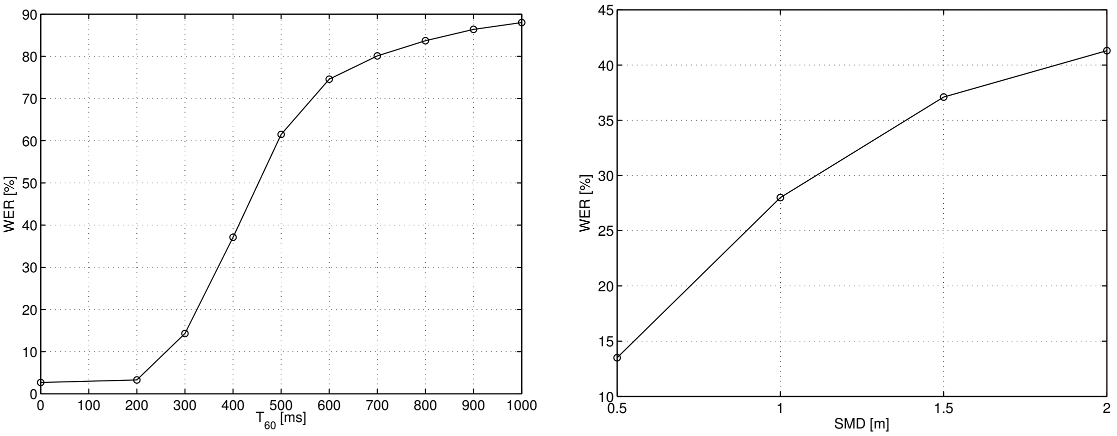
(b) reverberated

Figure 2.4: The spectrograms for the utterance "one nine zero seven" for (a) close-talking recording and (b) reverberated speech.

recordings. In the testing a 1.5 hour long subset of FreeSpeech database was used and convolved with simulated RIRs as explained later.

Speech signal was framed applying 25 ms long Hamming window with 10 ms overlap. The basic feature vector of 16 MFCC was extended by a voicedness feature [27]. The Mean and Variance Normalization (MVN) was applied on the cepstral coefficients and the speaker adaptation was handled by the fast Vocal Tract Length Normalization (VTLN). The temporal context is preserved by concatenating the features of 9 consecutive frames. Prior to the acoustic model training, Linear Discriminant Analysis (LDA) was applied in order to reduce the dimensionality and increase the class separability [28]. The acoustic model was using HMMs and emission probabilities were modeled by continuous Gaussian mixtures using one globally pooled diagonal covariance matrix.

In the testing, close-talking recordings were convolved with a set of simulated RIRs. The simulation was made assuming a room with the dimensions 4 m x 5 m x 2.5 m. The reverberation time T_{60} and SMD were varied. The microphone was



(a) Reverberation time

(b) Distance to the microphone

Figure 2.5: WER as a function of (a) the reverberation time T_{60} with the fixed SMD=1.5 m and (b) the SMD with the fixed T_{60}=400 ms.

located in the middle of the shorter wall, 2 m above the ground. The assumed speaker was 1.7 m tall speaking towards the microphone. To generate the RIRs the image-source model method implementation from [29, 30] was used, applying the default absorption-coefficient weights.

The recognition results for different reverberation times and SMDs are shown in Figure 2.5. The graph on the left shows for the fixed SMD case how the WER grows with increasing T_{60}. While there is not almost any performance degradation for reverberation times below 200 ms, the error rate grows steadily for longer times, until the results of the recognition system become practically useless. Similar behavior may be observed when the reverberation time is fixed and the distance between the speaker and the microphone increases. On the left we may see that if the speaker is close to the microphone WER is rather small. This is because the direct wave is stronger with respect to the harming reflections and the effect of reverberation is less significant.

These recognition results clearly demonstrate that speech recognition technology can be extremely fragile when it is deployed in conditions different from those assumed in training. Some robust techniques are definitely needed to make the system useful in reverberant scenarios with distant-talking microphones. Many techniques have been explored to solve this problem. In the next section we briefly overview some of them.

2.3 State-of-the-art of Reverberation-Robust ASR

Research on reverberation-robust ASR received significant attention over the past years and several comprehensive overviews have been published recently in [17,31–33]. Approaches to dealing with the reverberation problem may be broadly classified into four categories:

- signal-level dereverberation techniques, and
- reverberation-robust feature extraction methods, both operating in the front-end, or
- robust acoustic modeling and adaptation techniques operating in the back-end, and
- decoder-based approaches like REMOS which is discussed in Chapter 5.

The first set of techniques aims to increase the robustness of the recognition system by removing the reverberation from the speech signal or features. This is done in the front-end of the ASR system usually by spatial filtering (beamforming), inverse filtering, spectrum or feature enhancement. Alternatively, the back-end techniques aim to adjust the parameters of the acoustic model to the distorted speech, or to increase robustness by modifications in the decoding process to account for the reverberation distortion. Naturally, techniques from different categories may also be combined for additional improvements, as was done for example in [34].

2.3.1 Front-End Methods

The central problem to robust ASR with distant-talking microphones is the mismatch between training and testing signals due to varying acoustic conditions. With the increasing SMD increases also the power of distortion captured by the microphone relative to the power of the desired signal. Recall, that not all parts of RIR harm the recognition performance equally, and that early reflections may even help the recognition (Subsection 2.2). Let's hence rewrite Equation (2.8) to the following form:

$$y(n) = \sum_{\tau=0}^{T_h} h(\tau)x(n-\tau) = \sum_{\tau=0}^{T_e} h(\tau)x(n-\tau) + \sum_{\tau=T_e+1}^{T_h} h(\tau)x(n-\tau), \tag{2.10}$$

where T_h is the length of the whole RIR and T_e is the end of the interval with non-harming early reflections. Ideally, to obtain a good estimate of the clean speech we would like to suppress the second component corresponding to the late reverberation. However, the separation of these components, or the estimation of the late reverberation are very difficult, because of high non-stationarity and correlation of both terms.

Beamforming

Beamforming [6] is a technique that can be applied in multi-microphone environments and is a case of spatial filtering. The key idea is to steer the main lobe of the beamformer towards the desired source (the origin of the speech) and enhance the energy coming from that direction by suppressing the sounds coming from others. By doing so, the late reflections (the second term in Equation (2.10)) which take longer than the direct path are in fact suppressed.

There are several issues in microphone array processing that limit the beamforming performance when applied in ASR:

- most of the multi-microphone processing techniques require the source position and speaker orientation to be known [35],
- precise calibration of microphones is required to avoid different amplitudes and phase responses of the individual microphones in the array [36, 37],
- position of the sensors must be known and fixed, and
- narrow band signals and far-field propagation are generally assumed in the array processing theory, which is usually not true for speech.

Nevertheless, improvements in recognition performance have been reported for different beamforming strategies, for example using the basic delay-and-sum beamformer [38,39], maximum kurtosis beamformer [40], likelihood maximizing beamforming [41], as well as with emerging approach in the area of far-field audio and speech processing based on spherical microphone arrays [7], or with many others [17, 42].

Inverse Filtering and Spectrum Enhancement

If $h(n)$ was known and unchanged along the duration of utterance, an inverse filter could be designed to reverse the reverberation effects. To ensure the existence of such filter the impulse response is required to be be minimum phase, which is not the case in most acoustic environments. However, in [43] it was shown that exact inverse filtering can be achieved with multiple microphones. The problem is that there are not many practical applications in ASR where RIRs would be known, so the filter must be estimated from the speech signal. This process is called blind deconvolution. Different blind deconvolution methods are discussed in [44–46], or in more recent publications [47–49].

An interesting dereverberation approach was presented in [50]. As reasoned there, speech is a non-stationary signal with signal energy varying over a wide dynamic range both in temporal and spectral domains. The relative degradation of speech due to reverberation may therefore vary even among segments within the same frame. By identifying those relatively less distorted segments and enhancing them, overall enhancement of reverberated speech may be achieved. In [50] this was done by weighting Linear Prediction (LP) residual signal using the weight function derived from the characteristics of reverberant speech in differently distorted regions.

In [51] the authors suggest to model the reverberated speech as a convolution of the clean speech and RIR in the Gammatone spectral domain. Applying the Non-negative Matrix Factorization (NMF) framework and assuming non-negative spectral coefficients, reverberated spectra are decomposed into clean and RIR spectra using the least-squares error criterion. Clean spectra is than used to extract conventional features. Remarkable WER reductions were reported for both, matched and unmatched training.

Reverberation-Robust Feature Extraction

Another strategy that may be used to increase the robustness of the ASR systems is to use speech representations insensitive to reverberation. Ideally, features extracted from the clean signal would be similar to those extracted from the reverberated speech. Conventional frame-based techniques fail to perform in reverberant conditions, because the reverberation noise is non-stationary and the effects span over several consecutive frames. For this reason methods like RelAtive SpecTrAl processing (RASTA) [52], that have been successfully applied in cases when speech was transmitted over the channel with a short impulse response fail.

When dealing with reverberation, it is important to account for the long-term acoustic context. The first and second time derivatives of the static coefficients are usually used to capture the temporal changes in the spectra, but often the covered time-range is not wide enough to capture the reverberation effects. Furthermore, in reverberated speech the delta coefficients extracted in the usual way tend to keep the constant value for a long time. A new scheme for calculating the dynamic features that solves this problem was proposed in [53].

Studies on the perception of speech by human listeners revealed the importance of the slow changes in the speech spectrum for speech intelligibility [54, 55]. This observation suggests that ASR robustness might be improved by focusing on temporal structure of the speech signal that appears as low-frequency (below 16 Hz). In [56] this concept was adopted and modulation spectrogram emphasizing the temporal structure at very low frequencies was used to extract the features.

In [31,57] a feature extraction method called Harmonicity-based Feature Analysis (HFA) was presented. It is based on the idea that harmonic components of the voiced speech spectrum are less affected by reverberation. On the other hand, unvoiced parts are strongly corrupted by the energy spread from the previous segments. The harmonic components of the voiced spectra are therefore used to synthesize a purely harmonic signal and corrupted parts are replaced by a noise floor.

As describe before, to compute MFCC coefficients, the DCT is applied to the logmelspec energies to obtain decorrelated coefficients. The authors in [58] suggest to replace the DCT by the kernel Principal Component Analysis (PCA). After the transformation the main speech element is projected into low-order features, while noise or reverberant elements are projected into the high-order ones.

2.3.2 Back-End Methods

The strategy of front-end methods is to remove the distortion from the speech signal, so it would be more similar to the undistorted signals that were used to train the HMMs. An alternative approach is to use the models that reflect the statistical properties of the reverberated speech. The parameters of those models may be found in different ways. The most conventional one is to train the HMMs on reverberated data; alternatively, clean models may be adapted to the new acoustic conditions.

Matched Training

It is a well known fact that training and testing in matched conditions leads to a better recognition performance. If the parameters of acoustic model are trained on data which was collected in the target environment, recognition accuracy of such system increases. But, to record a speech database separately for each environment is costly and impractical. Alternatively, reverberated training data may be generated by convolution of the clean speech with the RIRs measured in the target room, as it was done for example in [59]. This way the data collection effort is significantly reduced.

The RIR is a very unstable parameter that depends on many factors, like the relative position of the speaker and microphone, position of the objects in the room, or temperature. It is impossible to collect all RIRs and so only a few are measured. However, this approach may be justified, because even if the training data is synthesized in that way, the recognition performance is only slightly lower compared to the case when naturally reverberated signals are used [60].

The data collection effort may be further reduced by using artificially generated RIRs as it was done in [61]. In that work RIRs were designed using only two high-level acoustic properties of the target reverberant environment, namely the Direct-to-Reverberant Ratio (DRR) and the reverberation time T_{60}. The approach was further extended in [62], where a set of acoustic models was trained for different a priori-defined reverberation conditions. During recognition, the full-band reverberation time was estimated from a speech utterance and used to select the best acoustic model.

Model Adaptation

Large amount of training data is needed to obtain well-trained HMMs. It is not always convenient to collect or synthesize it, and to retrain the models from scratch for each environment. Alternatively, the parameters of well-trained clean-speech models may be adapted to new conditions, using a relatively small set of annotated data recorded in the target environment. Well known methods used for speaker and additive noise adaptation like Maximum-Likelihood Linear Regression (MLLR) [63] and Maximum A Posteriori (MAP) [64] adaptation, or Parallel Model Combination (PMC) [65, 66] or Vector Taylor Series (VTS) [67] can be used to adapt the models to reverberant environments, but the performance improvement is often insufficient for long reverberation times. This is mainly because they assume the additive nature of the noise and intra-frame distortion, which is not the case of reverberation. Furthermore, even if we make some approximations and consider reverberation as additive noise, there is still the problem of extreme non-stationarity of the noise, which makes the noise parameter estimation very difficult.

All adaptation methods tailored to reverberation are using a similar concept. They estimate the energy contributions from the previous frames or states and use it to adapt parameters of Gaussian Mixture Model (GMM)s. For example, in [68] the mean of the energy parameter at each state is adapted by adding the energy of the state itself and contributions from the proceeding states. If MFCC are used, cepstral coefficients are first transformed back to the spectral domain where they are adapted in a similar fashion as the energy coefficient. The weight of the contribution from each state is estimated from the reverberation time, assuming an exponentially decaying shape of RIR. The T_{60} is estimated iteratively from the previous utterance, searching for the value that maximizes the likelihood of the already decoded sequence. A forced alignment is used to reduce the computational cost.

In [69] the means of HMMs are adjusted during the decoding using a state-dependent estimate of the late reverberation determined by joint use of a feature-domain reverberation model and an optimum partial state sequence from the clean-speech model. The reverberation model [8] is room-specific and captures the statistical properties of the measured RIRs. It may be estimated independently of the speech HMMs, what facilitates the system deployment in different environments.

While the previous two approaches are using measured or approximated RIRs to estimate the adaptation coefficients, in [70] the energy contribution of proceeding states is estimated in a maximum-likelihood manner from a few seconds of transcribed adaptation data. This method, as well as others [71, 72], is capable to compensate both, additive noise and reverberation.

2.4 Summary

The work in this thesis was developed in the context of the UPC smart room focusing on ASR in reverberant environments. Therefore, in this chapter we have first briefly described the basic components of HMM-based speech recognition systems, explained the reverberation phenomenon, and demonstrated on a LVCSR task its impact on the ASR performance. While the ASR performance may be satisfactory in clean, controlled conditions, it decreases significantly when the same acoustic models are applied to ASR in reverberant environment. It is therefore important to increase the robustness of the ASR systems, if we want to use them in applications with distant-talking microphones. Several alternatives proposed in the literature to solve this problem were listed in the last section of this chapter.

Conventional reverberation-robust techniques, in general, try to remove the distorting effects from the signal, or diminish them by modifications in the acoustic model. In the following part of this thesis we investigate an alternative concept, which may be used in parallel with them. It is applicable to scenarios with multiple microphones and exploits the spatial diversity. The motivation behind this approach may be illustrated by the following example. In the results of the experiment with increasing SMD (Figure 2.5b) we observed, that WER grows if the distance between a speaker and a microphone increases. Imagine a scenario with multiple microphones placed around the room recording the speech signal in parallel. If we had to pick up one microphone for recognition, the results of the experiment suggest that the best choice would be the closest one to the speaker, which in our experiment recorded the least distorted signal. If the speaker moves, the best microphone probably changes. How to do this selection automatically and optimally for ASR is the problem to solve in CS, which is discussed in the following chapter.

Chapter 3

Channel Selection

Assume a practical, cost-effective and unconstrained multi-microphone scenario, where the microphones are arbitrarily located and show a variety of characteristics. For instance, a reverberant room as the one shown in Figure 3.1, where some microphones are hanging on the walls, others standing on the table, or they may be built in personal communication devices of the meeting participants. Moreover, some of them may be omnidirectional, others directional or noise-canceling, etc. In such situation, where the positions of the microphones are either not known or fixed, the application of commonly used multi-microphone approaches, like array processing [6], becomes difficult.

An alternative is provided by CS. Before any processing, the degree of signal distortion differs among the channels, depending on the speaker position and microphone characteristics. Even if speech enhancement is applied, the processed speech signals will not be distorted equally, so some of them may be decoded with less recognition errors than others. Consequently, the ASR system may benefit if signals of higher quality are selected for further processing. To do so, a measure of distortion, or a measure of how well recorded or enhanced signals fit the set of acoustic models of the ASR system is needed.

Ideally, we would select the channel that leads to the highest recognition accuracy. As the WER is unknown during recognition, the main problem is to develop a measure, that allows to rank the channels in a way as close as possible to the WER based ranking. In this chapter we classify and describe the CS measures found in the literature. Then several new techniques developed in this thesis are presented.

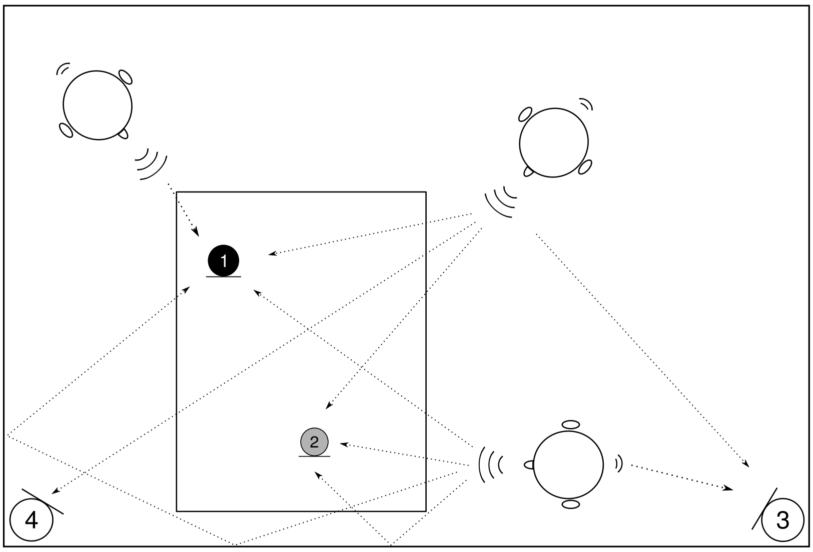


Figure 3.1: An example of unconstrained multi-microphone scenario with several speakers. Some direct and reflected waves are illustrated by the arrows.

3.1 State-of-the-art of CS

The CS techniques may be classified into two groups as signal-based and decoder-based, depending on the way how the measure that is used to make the decision is extracted. Follows the detailed description of the methods from both groups which were published in literature so far.

3.1.1 Signal-Based CS Methods

The signal-based measures are extracted from the signal or channel characteristics. The CS methods using these measures operate in the front-end and the decoder of the ASR system is not involved in the measure extraction. If the clean speech is used to train the acoustic models, we can hypothesize that the least distorted signal leads to the highest ASR accuracy and select it for recognition. The advantage of this kind of methods over the decoder-based ones is their lower computational complexity. The channel can be selected before the signal enters the classification part of the ASR system, so recognition is made only once. The disadvantage is that the CS measure is tailored to a specific kind of distortion and may fail in different conditions. In this thesis, we primarily assume reverberant environments, but many of developed CS methods used also in presence of other kind of distortion.

Actually, only two signal-based CS methods were reported in the literature. One is using Signal to Noise Ratio (SNR), a well known signal quality measure, the other one is using the cross-correlation among signals from different microphones in a microphone array, and is applied to reduce the number of channels that are used for the beamforming.

Energy and Signal-to-Noise Ratio

A straightforward way to identify the least distorting channel could be the signal energy. A strong signal indicates that the sound was uttered with the speaker close and oriented towards the microphone, therefore the direct wave is presumably stronger relative to the reverberation. This very simple approach may achieve good results, but one strong assumption must be made. In multi-microphone scenarios, attenuation in the electrical path among microphones varies for reasons like different wire length, varying volume set on preamplifier, different microphone type, etc. If we want to use signal energy as a reliable indicator of the signal quality, a perfect calibration of all microphones is needed, which is not a trivial task.

The problem of calibration could be avoided if the energy of the speech signal was normalized, for example, by the energy of the noise in the silent portions (assuming that some additive noise is present). This leads us to a SNR. CS based on this measure was evaluated in [73] and [74]. If speech is recorded by distant-talking microphones, reverberation is often the dominant source of distortion. A problem associated to the use of the SNR is that it does not properly reflect that kind of distortion. Furthermore, an accurate SNR measurement can be hardly obtained, since the boundaries between the speech signal and the silent portions, where the noise power can be estimated, are less clear after the smearing effect of reverberation. Another disadvantage of energy-based measures in general is that they do not consider the specific characteristics of the speech signal (only its energy).

Multichannel Cross-Correlation Coefficient

One of the advantages of CS is that it does not require a spatial structure of the microphone set, what simplifies the deployment and reduces the cost of the system. CS may also be combined with beamforming and used to reduce the number of channels. Although, in theory, a higher number of microphones in the array should lead to a better beamforming performance, in practice, it was shown that the use of all possible channels does not always increase the ASR accuracy [39, 73–75].

The CS method using Multichannel Cross-Correlation Coefficient (MCCC) [76] as a measure for identifying reliable channels from the microphone array was proposed in [75]. The basic idea of this approach is to treat the channel that is uncorrelated with the others as unreliable, and select only a subset of microphones with the most correlated signals. The experiments in [75] show, that for a given configuration the number of channels could be reduced by half without significant loss in the recognition performance.

The assumption to treat the channel uncorrelated with others as unreliable can be justified when applied on a microphone array, where sensors are placed close to each other, but it may fail in unconstrained scenarios. For example, assume a situation with three microphones, two placed at a distance and one close to the speaker. Signals in the distant microphones will probably be more correlated with each other than with the signal in the close-talking one, which would then get discarded. Also, since at least one microphone pair is needed to extract this cross-correlation measure, it is not possible to select only one channel. Note, this is the only CS method we have found in the literature, that has not been implemented and included in the evaluation.

3.1.2 Decoder-Based CS Methods

The decoder-based approaches do not estimate the degree of signal quality using some signal-processing measure, but the estimation process includes some kind of classification, which may be directly related to the decoding part of the recognition system (e.g. by using likelihood, or posterior probabilities). As that implies performing a classification for each channel, CS methods based on these measures are computationally more demanding. With the signal-based measures we want to ascertain which is the cleanest signal (channel), but our real objective is to minimize WER. Intuitively, a measure extracted from the decoder can be more correlated with the WER than a measure extracted from the signal, what can be an advantage of the decoder-based methods.

Likelihood

The first, straight forward, decoder-based measure based on acoustic likelihood was presented in [77]. In that implementation, signals from all channels were passed to the recognizer and the channel giving the maximum acoustic likelihood was selected. In the following, we will outline a problem related to this measure and explain why it should not be used for CS.

In conventional ASR systems the Bayes' rule is applied to compute the posterior probability as show in Equation (2.2). The probability of the observation vector $p(O)$ in the denominator is usually omitted since it does not depend on the selected sequence of words and works only as a scaling factor. Thus, the use of non-normalized likelihood scores is not a problem when comparing competing word sequences form a single channel.

If there are several parallel channels, the probability of the observation vector $p(O_m)$ is different for each stream. The posterior probability for the multi-channel case is then defined as

$$P(w \mid O_m) = \frac{p(O_m \mid w)P(w)}{p(O_m)}, \tag{3.1}$$

where $m = 1, \cdots, M$ is the channel index. Obviously, the probability of the observation vector can not be neglected if we want to compare competing word sequences from different channels. It may happen that in some channels the word sequences are decoded with less errors, but the corresponding non-normalized likelihood scores are smaller than in the competing streams. This is the reason why normalization is needed and non-normalized scores, usually provided by ASR systems, can not be used as a reliable indicator of channel quality. In this thesis we address this issue by using a pairwise likelihood normalization across channels, which is discussed in Section 3.2.4, or by applying a normalization factor derived from the N-best list of each channel as described in Section 3.2.5.

Feature Normalization

In [39, 73], some feature normalization technique (e.g. mean and variance normalization [78, 79] or histogram equalization [80, 81]) is first applied to each channel as illustrated in Figure 3.2. Then, both the original and the normalized feature streams are recognized, and the channel with the smallest difference between the recognized word sequences from the original and the normalized version is selected. The underlying assumption is that the higher the distance between the two recognized word sequences, the more distorted is the signal. A drawback of this method is again the computational complexity. To extract the measure, a full decoding has to be run for both the original and the normalized stream for each channel.

Class Separability

In [82], the CS based on a class separability measure was introduced. Class separability is a common concept in pattern recognition, used for example in the well-known linear discriminant analysis [83] to find the linear projection matrix. In the context of CS, instead of a projection matrix, we aim at finding the channel where the class separability measure is maximized, as

$$\hat{C} = \arg\max_m \{trace(\boldsymbol{S}_w^{-1}(m)\boldsymbol{S}_b(m))\}, \tag{3.2}$$

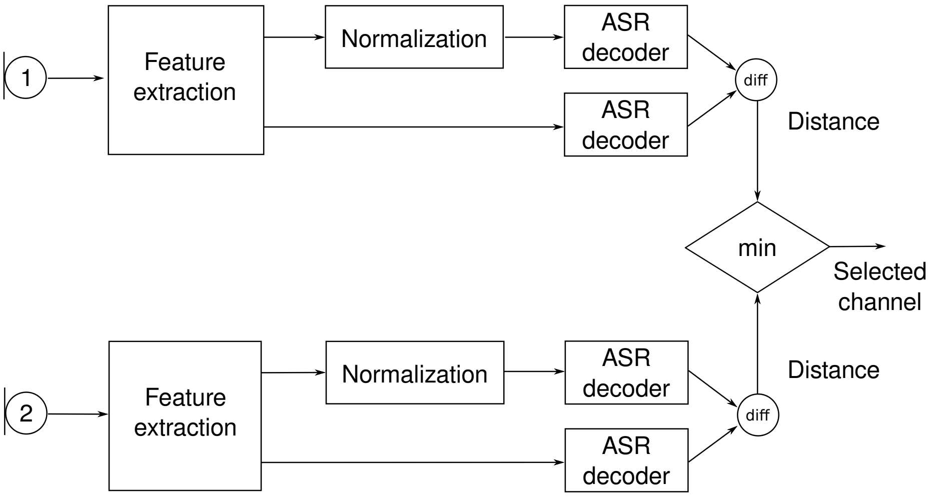


Figure 3.2: Block diagram of CS method based using feature normalization. Example for 2 channels.

where $m = 1, \cdots, M$ is the channel index. For a given microphone the between class scatter matrix $\boldsymbol{S_b}$ is estimated as

$$\boldsymbol{S}_b = \sum_{i=1}^{N_c} N_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T, \tag{3.3}$$

and the within-class scatter matrix $\boldsymbol{S_w}$ as

$$\boldsymbol{S}_w = \sum_{i=1}^{N_c} \sum_{j=1}^{N_i} (\boldsymbol{x}_{ij} - \boldsymbol{\mu}_{ij})(\boldsymbol{x}_{ij} - \boldsymbol{\mu}_{ij})^T]. \tag{3.4}$$

where N_c is the total number of classes and N_i denotes the number of samples (feature vectors) in class i. The mean vector for the i^{th} class is defined by $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}$ defines the mean vector over all classes.

As indicated in the original work [82], the choice of the class units for the estimation of the matrices in Equations (3.3) and (3.4) is not trivial. These classes may either correspond to the units used by the recognition system (e.g. phonemes, tri-phones or words), or they may be different (the so called stand-alone approach), derived by merge and split training, as it was done in the original work. The stand-alone approach, surprisingly, let to better results there. The authors reason that this might be due to the high number of classes (phonemes were used there), and because the time boundaries to separate the frames into the classes were extracted only from the alignments of one channel, resulting in a possible mismatch when applied to other channels.

3.2 New CS Methods

In this section we present new CS methods developed in this thesis. First, we describe three signal-based approaches, in particular, we discuss the possibility to use the information about the position and orientation of the speaker, to extract the CS measure from the RIR, or from the sub-band energy envelope of the speech signal. Then, we present two decoder-based methods, one using the pairwise normalization of the likelihood across channels, and the other one using the N-best lists to extract the likelihood normalization factor.

3.2.1 Position and Orientation

Speech should be less distorted by reverberation if the microphone is closer to the speaker. In [84], we discussed the possibility to use information about the relative position and orientation of speaker and microphone for CS. For instance, the closest microphone may be selected by measuring the time of arrival of the waveform. However, as was shown in [84], the information about the orientation is even more important, mainly due to the attenuation of the signal by the head of the speaker, and the fact that speech used in training is usually recorded by a microphone in front of the speaker. Both position and orientation may be estimated either from multi-microphone audio processing, multi-camera video processing, or a combination of both. In any case, CS would have to rely on the output of another system, that may not always provide accurate measures and the knowledge about the positions of the microphones is needed, what puts additional demands on the system deployment. In Section 4.2.2 we demonstrate how position and orientation estimation errors influence the performance of the ASR in the UPC smart room.

3.2.2 RIR-Based Measures

Assuming constant conditions in the room, the RIR can be used to describe the propagation between the acoustic source and a given microphone. In [25], relations between different parts of RIR and the WER of the ASR system were investigated. Authors showed that there are certain components of the RIR that harm speech recognition more than others. Consequently, if there was a feature extracted from the RIR that is correlated with WER, it could be used to predict the recognition performance before the speech recognition takes place.

In [85] we presented a methodology to identify relevant measures for CS, assuming an exact knowledge of the RIR. The process is outlined by a block diagram in Figure

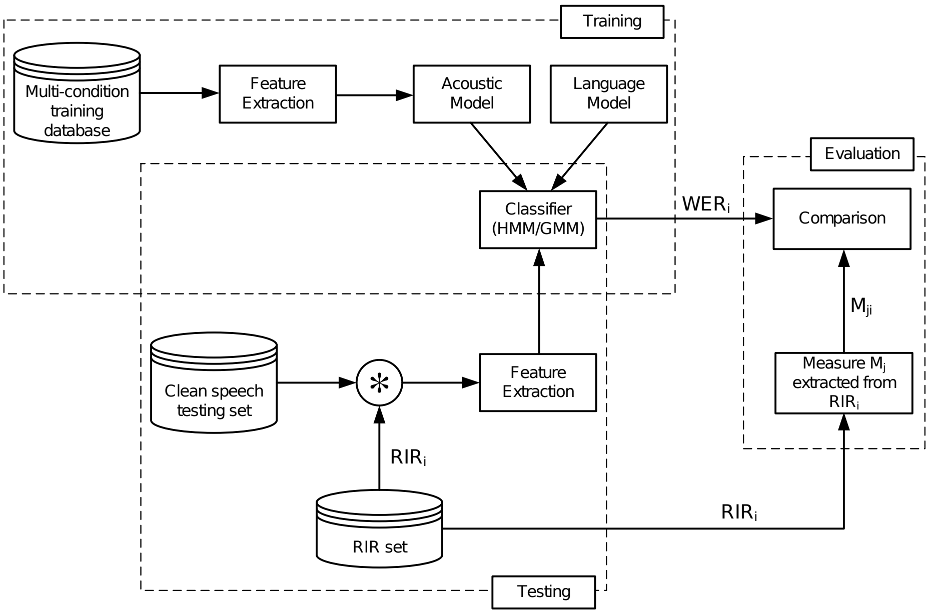


Figure 3.3: Block diagram of evaluation of the different RIR-based measures.

3.3. Let's denote by WER_i the obtained WER corresponding to the i^{th} RIR. Now we can choose a particular measure M_j, compute its values M_{ji} from every RIR_i and compare (correlate) those values M_{ji} with the corresponding values of WER_i. In this way, we can see the relation between each of the defined RIR measures M_j and the speech recognition rate, and choose the most relevant one(s). Then, such measure(s) can be used for selecting the best microphone before entering the recognition system.

As described in Section 2.2, RIR can be split into 3 parts: direct sound, early reflections, and late reflections. Using the described methodology, relations among speech recognition accuracy and different RIR-based measures M_j was investigated, namely:

- energy of the whole RIR
- energy of direct wave and early reflections normalized by energy of whole RIR
- energy of late reflections normalized by energy of whole RIR
- ratio between energies of early and late reflections

We found out that measure calculated as a ratio between energy of the late reflections (those harming the recognition most) and energy of the whole RIR showed the highest correlation index between the parameter and WER. This observation may be interpreted as lower the energy of late reflections normalized by global energy, lower

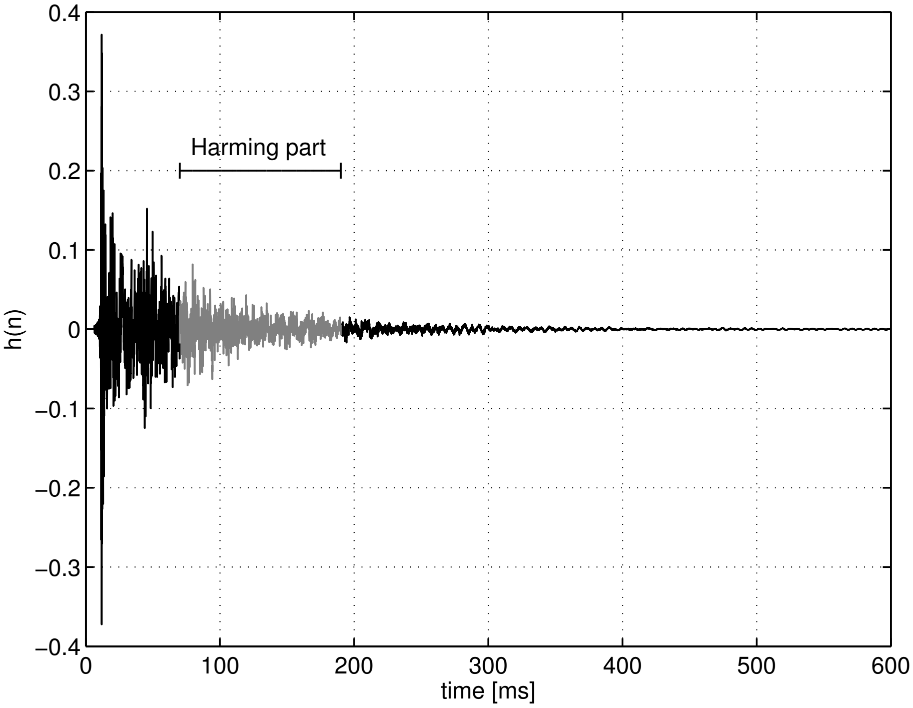


Figure 3.4: Harming and non-harming parts of the RIR.

the WER. It means that the microphone where this quotient of energies is the lowest will be chosen as the most suitable for recognition. Channel C with the lowest ratio value may be selected for further processing as:

$$C = \arg\min_m \frac{\sum_{t=50ms}^{190ms} h_m^2(t)}{\sum_{t=0ms}^{T} h_m^2(t)}, \tag{3.5}$$

where $h_m(t)$ is the RIR corresponding to channel m and T is the length of the RIR. Corresponding intervals of the RIR are highlighted in Figure 3.4.

The main problem of this approach is the RIR estimation, or the direct estimation of the measures derived from it. As RIR may change while a speech utterance is being produced, that estimation should be made online or directly from speech. Though this may be too difficult in quickly changing environments, it has been done, for instance, in [86] for a similar measure, the well known DRR.

3.2.3 Envelope-Variance Measure

To avoid the aforementioned problems of RIR estimation, we proposed a new CS method in [84], where the measure of distortion is extracted directly from the speech signal. It is a well known fact that reverberation smooths the time sequence of speech energy values, also called speech intensity envelope, so the effect of reverberation may be observed as a reduction in the dynamic range of that envelope [54]. Herewith, we define the sub-band envelopes as the time sequences of non-linearly compressed Filter-Bank Energies (FBE). Speech signal is first framed and windowed, and then the energy in mel-scaled sub-bands is calculated for each frame, like is done when extracting the MFCC.

To remove the short term effects (e.g. different electrical gain and impulse response of the microphone) from the signal in each channel, the mean value should be subtracted in the log domain from each sub-band, i.e.

$$\hat{x}_m(k,l) = e^{log[x_m(k,l)] - \mu_{log[x_m(k)]}}, \tag{3.6}$$

where $x_m(k,l)$ is the sequence of sub-band FBEs in channel m, k is the sub-band index and l is the time frame index. The mean $\mu_{x_m}(k)$ is estimated by a time average in each sub-band along the whole speech segment (utterance, in our case). An illustrating example of the envelope for clean and reverberated speech for a given sub-band is shown in Figure 3.5. Note the strong reduction in amplitude range of the reverberated signal in comparison to the clean speech signal.

After the mean normalization, the sequence of FBE is compressed applying a cube root function, as it was done for example in [87], and a variance measure is calculated for each sub-band and channel:

$$V_m(k) = Var[\hat{x}_m(k,l)^{1/3}]. \tag{3.7}$$

The cube root compression function in Equation (3.7) is preferred to the conventionally used logarithm, because very small values in the silent portions may lead to extremely large negative values after the log operation, which would distort the variance estimation. Conceptually, this variance measure is alike the modulation index in [54]. There are many similarities between the extraction process of this CS measure and the standard speech feature extraction algorithms, so the additional computational effort needed by CS may be very low. More importantly, if standard ASR features are used for CS, the distortion estimation takes place in the same speech representation domain, what may make it more effective.

Note that from Equation (3.7) results a set of measure vectors. Each vector corresponds to one channel and consists of the estimates of the variance of compressed FBE in each sub-band. Using all this information, it is possible to select a different channel for each sub-band, but the reconstruction of the signals would be a complex problem, and it is not considered here. Instead, we select the same channel for all sub-bands which shows the maximum weighted average variance over all sub-bands.

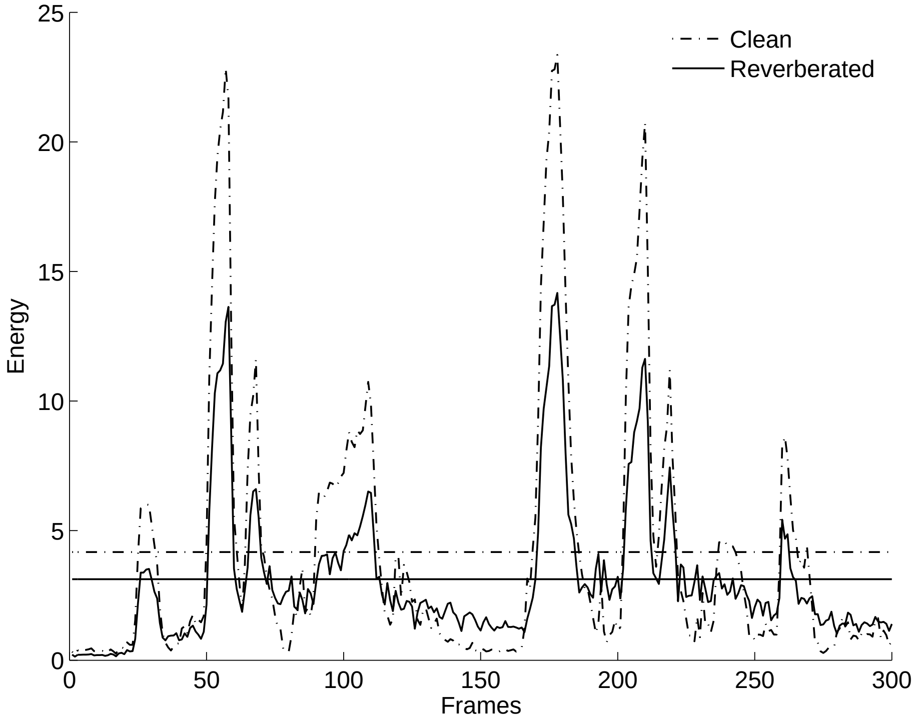


Figure 3.5: Sequence of sub-band FBE of clean and reverberated speech. Straight lines mark the mean values.

To do so, the variance is first scaled in each sub-band, to be in the range between 0 and 1, by dividing it by the maximum for that sub-band over all channels, and then a specific weight $\omega_m(k)$ is applied to each channel and sub-band i.e.

$$C = \arg\max_m \sum_k \omega_m(k) \frac{V_m(k)}{\max_m(V_m(k))} \tag{3.8}$$

The channel C whose measure is the highest is selected as the least distorted one.

The degree of distortion (reverberation, noise) may vary from one sub-band to another, so in principle a non-uniform weighting of the sub-band measures may lead to a higher ASR accuracy. A development data set may be used to tune the weights, or, alternatively, they may be related to the amount of sub-band distortion and estimated directly from the signal.

3.2.4 Pairwise Likelihood Normalization

If the posterior probability of the hypothesized word sequence in Equation (3.1) was available for each channel, it could be used as a CS measure. However, the estimation of the observation vector probability $p(O_m)$ requires either some assumptions or approximations [88], and it is not usually computed by the ASR system. To avoid this problem, we proposed in [89] an alternative solution based on the ratio of likelihoods.

Let's first discuss the case with 2 channels. To simplify the explanation, we will assume equal prior probabilities for all word sequences and neglect them in (3.1). Assume we know the decoded sequences w_1, w_2 and their corresponding likelihoods $p(O_1 \mid w_1)$, $p(O_2 \mid w_2)$ for both channels as illustrated in Figure 3.6a. If the word sequences are the same, the channel is selected randomly and no further CS related processing is needed. If they are different, we may take the sequence w_1 from the first channel and compute the likelihood of the observation from the second channel $p(O_2 \mid w_1)$ and vice versa, $p(O_1 \mid w_2)$, as shown in Figure 3.6b. Then, by dividing the likelihood $p(O_1 \mid w_1)$ by the likelihood corresponding to the alternative sequence $p(O_1 \mid w_2)$, we get a measure of confidence about the correctness of the decoded sequence w_1 in that channel 1. After computing the corresponding likelihood ratio for channel 2, the channel with maximum ratio (confidence) is selected.

For the general multi-channel case, by using addition of likelihood ratios across channels, the selection criterion can be expressed as

$$C = \arg\max_m \sum_i \frac{p(O_m \mid w_m)}{p(O_m \mid w_i)}, \tag{3.9}$$

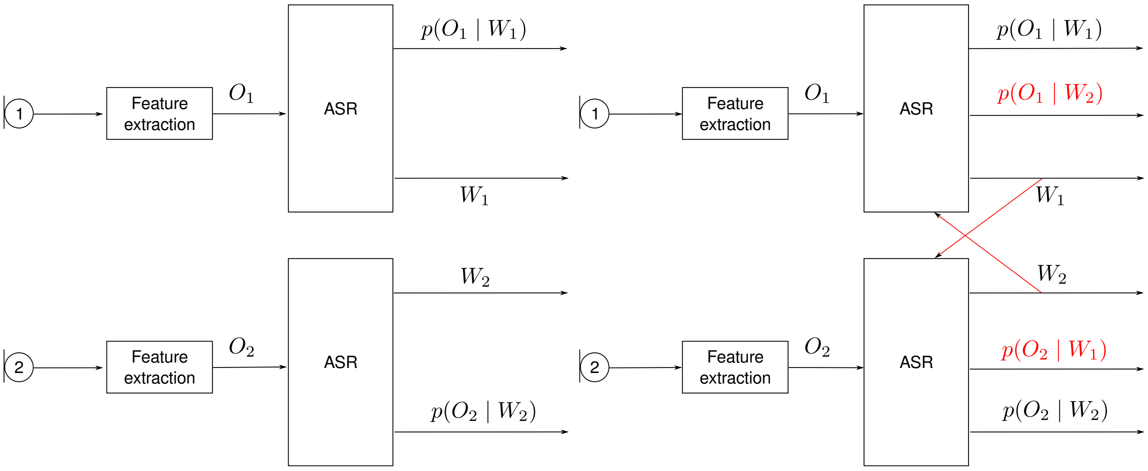
where $m = 1, \cdots, M$ and $i = 1, \cdots, M$ are channel indexes, and C is the selected channel. By doing this, we are selecting the channel which shows in average the maximum confidence regarding to its decoded word sequence. A computational drawback of this method is that in case all the M decoded sequences are different, the likelihood of the alternative hypothesis has to be calculated $M - 1$ times.

3.2.5 N-Best Hypothesis

The lack of normalization is also the key problem for confidence measuring, so many solutions may be found in that area [88, 90]. In this CS method, which we published in [91], the N-best list approach is adapted and applied as follows. It is a well known fact [92], that $p(O)$ may be computed as

$$p(O) = \sum_{w \in \Omega} p(O \mid w) P(w), \tag{3.10}$$

where Ω is the set of all possible word sequences for O. Apparently, without any constraints this enumeration is not feasible, so some approximations are required.



(a) The first step

(b) The second step

Figure 3.6: Two steps in the extraction process for pairwise likelihood normalization.

Let w^n be the n^{th} hypothesis in the N-best list. The $p(O)$ may then be approximated by the finite sum

$$p(O) \approx \sum_{n=1}^{N} p(O \mid w^n) P(w^n), \tag{3.11}$$

as it was done for example in [93] or [94].

Finally, based on the above reasoning the CS measure in our N-best approach is computed as

$$C_m = \frac{p(O_m \mid w_m^1)^{1/\alpha} P(w_m^1)}{\sum_{n=1}^{N} p(O_m \mid w_m^n)^{1/\alpha} P(w_m^n)}, \tag{3.12}$$

where n is the hypothesis index in the N-best list of channel m. The acoustic model likelihoods $p(O_m \mid w_m)$ usually have a very large dynamic range. An appropriate scaling factor α must be applied to them, otherwise the summations are often dominated by the largest value. The value of α can be estimated a-priori using a development corpus. Another option is to set it equal to the number of frames, as we did in this work. If the acoustic model likelihoods provided by the recognition system are in the log scale, setting α to this value is equivalent to divide the log likelihoods by the number of frames, which results in an average log-likelihood per frame.

3.3 Summary

In this chapter we introduced the basic concepts of CS. State-of-the-art methods were reviewed and categorized. Also, several new CS techniques developed in this thesis were presented.

The advantage of the signal-based methods is that the channel can be selected before the signal enters the classification part of the ASR system, so recognition is made only once. Contrary, in the extraction process of the decoder-based methods at least one classification is performed for each channel, so the CS methods based on these measures are computationally more demanding.

The signal-based techniques discussed here do not require any training phase. Instead they make assumptions about the back-end of the ASR system. Presuming clean speech is used in training, we can hypothesize that the least distorted signal leads to the highest ASR accuracy and select it for recognition. However, this strategy may not be optimal if the acoustic model is trained on the distorted data. Furthermore, signal-based measures are extracted practically at the beginning of the speech recognition process, many steps before the word sequence is hypothesized. Intuitively, one does not expect a high correlation with WER.

The decoder-based measures, on the other hand, are extracted, or at least use some information, from the decoder. Hence, in principle, they should be more correlated with the WER and lead to a better recognition performance than the signal-based ones. In the next chapter, we will compare the methods experimentally on a digit string recognition task.

Chapter 4

Evaluation of Channel Selection Measures

In this chapter the experimental evaluation of both, the CS measures already reported in the literature, and the new methods that were developed in this thesis is presented. We first describe the experimental setup, ASR system, and used databases in Section 4.1. Then in Section 4.2 the results for the first database, where close-talking recordings were convolved with measured RIRs to generate the reverberated speech signals, are presented for the individual CS measures. Some methods are evaluated further in Section 4.3, where we show how their performance depends on the amount of data that is used for the measure estimation. In Section 4.4 we demonstrate that the parallel combination of different CS measures allows further recognition rate improvements. We also propose a serial combination of signal- and decoder-based measures and show that it reduces the computational load without any significant loss in the recognition performance.

The results for the second database, where distant-talking microphone recordings are used instead of convolved signals to further evaluate some CS methods and their combinations in real conditions, are presented in Section 4.5. In Section 4.6 we analyze in more detail the CS method based on the N-best lists and show how its performance depends on the size and content of the list. Finally, in Section 4.7 we test the real-time CS client implementation. It is based on the Envelope Variance (EV) measure developed in this thesis and the tests were performed with the real distant-talking microphone recordings of several moving speakers in the UPC smart room. The chapter is summarized in Section 4.8.

4.1 The Experimental Setup

There are different ways how CS may be integrated into ASR. In our experiments we use the following strategy. We assume that a given speaker's utterance is recorded by several microphones. After the utterance ends, different CS measures are extracted, and signal from one channel is selected for recognition. The objective is to minimize WER, so a good CS measure will indicate signals that lead to minimum recognition error.

4.1.1 Databases

The experiments were conducted with two different setups[1]. Two databases were used: TIDigits, a well known database of connected digits in English [95], and the Meeting Recorder Digit (MRD) corpus [96]. The MRD corpus is a collection of connected digit strings recorded in a real meeting room at the International Computer Science Institute (ICSI) as a part of the ICSI Meeting corpus data collection [97].

In the first setup, the original close-talk recordings from TIDigits were downsampled to 16 kHz and convolved with a set of real RIRs, which were measured in the UPC smart room. The details about the RIR database may be found in Annex A. As shown in Figure 4.1, in that room there are 6 T-shaped omni-directional microphone clusters with 4 microphones per cluster installed on the walls. In our experiments we selected only the microphones in the middle of each cluster. There are 16 positions in the RIR database. Only a subset of RIRs corresponding to seven different positions and four orientations of the speaker was used for testing. The original position number is kept in the figure. Only the utterances from adult speakers (8700 files in total) were included in the experiments.

The advantage of using convolved signals is the controlled experimental setup, which simplifies the evaluation of the tested CS measures. The MRD corpus, on the other hand, was used to test the measures also in real conditions. In this corpus, the sequences of digits were read by the meeting participants and recorded in parallel using 4 distant-talking microphones that were placed on the table in the meeting room. There are 29 speakers who, in summary, read 2790 utterances over 22 sessions (not all speakers participated in all sessions). Both, native and non-native speakers were included in the tests.

[1] Excluding the real-time CS client tests, which are described separately in Section 4.7

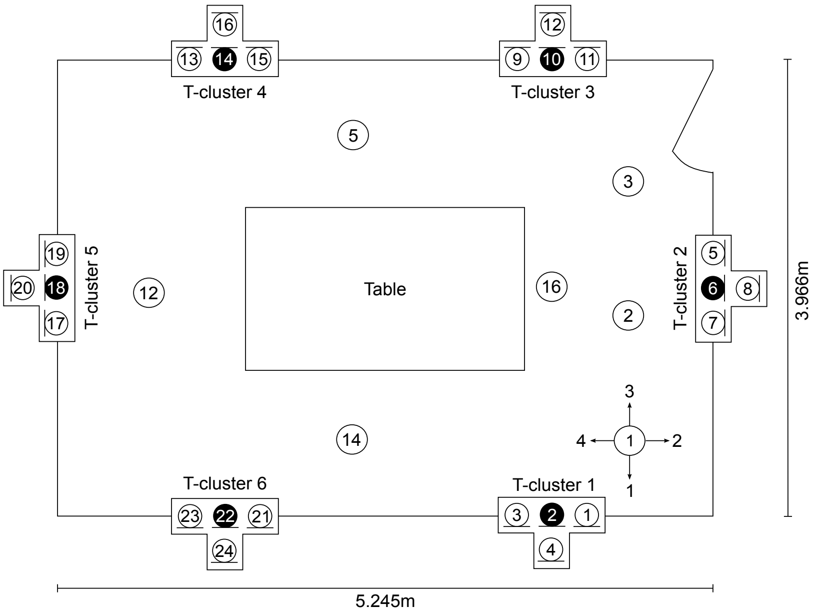


Figure 4.1: The UPC smart room with 6 highlighted microphones used in the experiments, and 7 acoustic source positions and 4 considered orientations of the speaker used for testing.

4.1.2 The ASR System

A continuous HMM-based system (Hidden Markov Model ToolKit (HTK) [98]) was used, applying the setup commonly used for TIDigits. The 11 models for words (digits zero, oh, one, ..., nine) have 16 states, the silence model has 3 states, and a short pause model 1 state which is shared with the middle state of the silence model. There are 3 Gaussians per state for the words, and 6 for the silence model. Standard MFCC features were extracted from 20 mel-frequency bands. The feature vector consisted of 12 cepstral coefficients without the 0^{th} coefficient, frame energy, delta and acceleration features. The size of the vector was therefore 39. Frame length was 25 ms and frame shift 10 ms. Optionally, MVN of the baseline features was used.

The acoustic model was trained in two different ways, either using the original close-talking recordings from the standard training set of TIDigits, or, those recordings were convolved with randomly selected RIRs from the UPC smart room[2], or from the room where the MRD corpus was recorded. We will refer to these cases as clean and matched training respectively. The recognition WERs of the system, when trained and tested with clean TIDigits, are 0.7% using the baseline features (MFCC

[2] Only RIRs from positions that were not used in testing (Figure 4.1) were used to generate reverberated data for training. All positions and assumed orientations of the speaker from the database are shown in Figure A.1. The microphones used in training and testing were the same.

+ energy $+\Delta + \Delta\Delta$), and 0.6% when MVN is applied.

4.2 Experiments with Convolved Signals

In the experiments with convolved signals all 8700 utterances from the testing database were used repeatedly to generate reverberated speech for each position and orientation. The WERs for every microphone for convolved TIDigits are presented in Table 4.1. As expected in the distant-talking microphone environment, a significant ASR performance degradation may be observed compared to the clean speech case. In the UPC smart room the selected microphones are numbered 2, 6, 10, 14, 18 and 22. Each presented WER is calculated over all positions, orientations and speakers. In other words, it shows the performance of the ASR system as if only that particular microphone was present in the room. If the average WER over all microphones was calculated for a given configuration, it would show the recognition performance of the system as if all microphones were present in the room, but the signals for recognition were selected randomly. As expected, when MVN is applied to the baseline features, the WER decreases. Even more significant WER reduction may be observed when the acoustic model is trained in the matched conditions.

4.2.1 Recognition Results with Signal-Based Measures

The recognition results in terms of WER for TIDigits convolved with the RIRs from the UPC smart room, are shown in Table 4.2 for CS with the presented signal-based measures: signal-to-noise ratio (SNR), the measure extracted from the room impulse response (RIR), envelope variance (EV), distance (D), and orientation (O). The case of random channel selection (RND) is also included for comparison purposes. Note that the WER for random CS is equal to the average of WERs from the individual microphones in Table 4.1. The CS performance was evaluated on the baseline features (MFCC + energy $+\Delta + \Delta\Delta$), and also applying MVN to them. A single utterance

Table 4.1: WER using a single microphone for convolved TIDigits.

Training	Features	2	6	10	14	18	22
Clean	Baseline	30.7	30.9	31.5	28.1	26	27.2
	+ MVN	23.5	22.9	24.5	22.1	21.4	21.6
Matched	Baseline	7.2	7	6.7	6.3	6.8	6.1
	+ MVN	6.2	6	6.2	5.9	6	5.9

was used to extract the CS measure when the extraction was made from the speech signal (i.e. EV, and SNR), so a different channel could be selected for each utterance. If the measure is independent from the speech (i.e. RIR, distance and orientation) the same channel was selected for all utterances for a given position and orientation.

We observe that all techniques, except SNR, perform much better than the random CS, for both, the clean and matched training. The relative improvement with respect to the random case is shown in parenthesis. The improvement in the matched training case is consistent, but smaller compared to the clean training. This may be because the signal-based CS methods in the tests assume that the least distorted speech leads to the best recognition result, which may not be true if the acoustic model is trained with the reverberated signals.

As expected, when MVN is applied to the baseline features, the recognition accuracy increases. The relative improvement with respect to the random selection in the clean training case also increases (for all methods except SNR). This may be due to the fact that normalization attenuates the short term effects of the microphone transfer function as well as other stationary noises. Consequently the reverberation effects become more dominant in the measures and the least distorted channel may be identified more accurately. The same effect does not appear in the matched training case. Although, thanks to the normalization, we may select the least distorted signal more accurately, it is less probable that the cleanest signal will lead the best recognition result when the acoustic model is trained with reverberated speech.

The performance of CS based on the EV measure is almost the same as the performance of the RIR-based method (a similar performance was reported on a large vocabulary continuous speech recognition task in [84]). Note that, whereas the method based on the RIR assumes knowledge of channel characteristics, the EV-based measure does not make any strong assumption.

In our implementation, the National Institute of Standards and Technology (NIST) SPQA tool [99], the same as in [73], was used to estimate the SNR for each utterance. The channel with the largest SNR was selected for recognition. CS based on SNR does not work at all. The performance is similar, or worse than the random CS. Reverberation is the main source of distortion in this setup, and measuring SNR as if additive noise was present is not very meaningful.

If only the measure of distance between speaker and microphone is used, some degradation in the recognition performance may be observed in comparison to other measures in the UPC smart room. When information about orientation of the speaker is used, and the channel with the most direct orientation of the speaker to the mi-

Table 4.2: CS performance in terms of WER for convolved TIDigits using signal-based measures. The relative improvement with respect to RND is shown in parenthesis.

	Clean		Matched	
CS method	Baseline	+ MVN	Baseline	+ MVN
RND	29.1	22.7	6.7	6
RIR	25.4 (12.7%)	19.3 (15%)	6.2 (7.5%)	5.6 (6.7%)
SNR	28.1 (3.4%)	22.7 (0%)	6.7 (0%)	6.1 (-1.7%)
EV	25.5 (12.4%)	19.2 (15.4%)	6.1 (9%)	5.5 (8.3%)
D	27.7 (4.8%)	20.2 (11%)	6.6 (1.5%)	5.8 (3.3%)
O	25.6 (12%)	19.1 (15.9%)	6.2 (7.5%)	5.6 (6.7%)

crophone is selected, the resulting WER is similar to that of RIR- and EV-based methods. An exact knowledge of the position and orientation was assumed when calculating the results presented in Table 4.2. In real conditions those measures must be estimated, so some errors will be introduced and consequently WER will increase. This problem is discussed in the following section.

4.2.2 Position and Orientation-Based CS

In this section we further evaluate the performance of the position and orientation-based CS methods assuming that the measures used to select the channel were estimated with some errors. The correct position of the speaker was shifted in an arbitrary direction by adding a randomly generated numbers to the known coordinates. The numbers followed a normal distribution with the zero mean and the standard deviation varying from 0 to 50 cm. Resulting mean localization estimation error or such system, calculated as the mean of the Euclidean distances between the correct and shifted positions, is shown in Figure 4.2. As expected, it linearly grows with the increasing standard deviation.

The recognition performance in terms of WER of such system in the UPC smart room may be seen in Figure 4.3. The results are displayed for the clean and matched training, and using the baseline and normalized features. For example, if we assume a system with the mean localization estimation error around 33 cm, which corresponds to the case with the standard deviation 25 cm in Figure 4.2, the resulting WER in the matched training case using the normalized features is 5.8%. We may also observe that WER is constant or does not grow very much when increasing the standard deviation. Actually, even if the position is estimated without any errors (the first

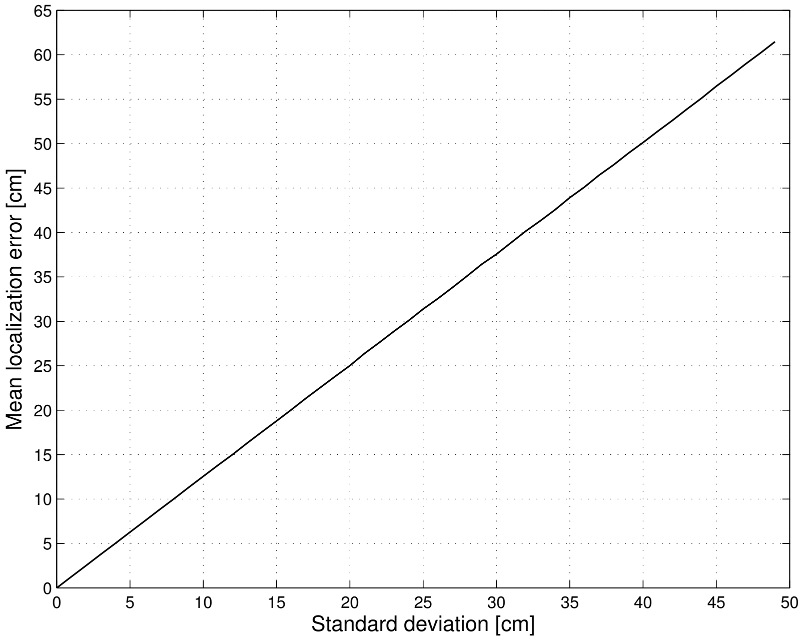
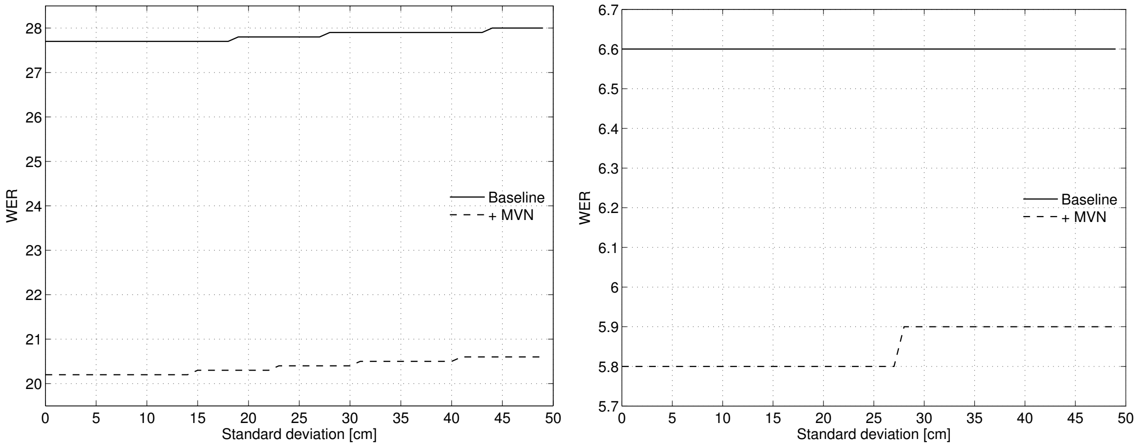


Figure 4.2: Emulated mean localization estimation error.

point of the graph) the WER is already quite high, close to the random selection case as shown for the distance measure (D) in Table 4.2. This may be caused by the fact, that many times the closest microphone is located behind the speaker and so the signal is attenuated by the speaker's head. Also, the database used in training was recorded (as usually) with the front microphones. Therefore, the orientation measure in this case may be a more reliable indicator of channel quality.



(a) Clean training

(b) Matched training

Figure 4.3: The position-based CS performance in terms of WER varying the position estimation error for TIDigits convolved with the RIRs from the UPC smart room.

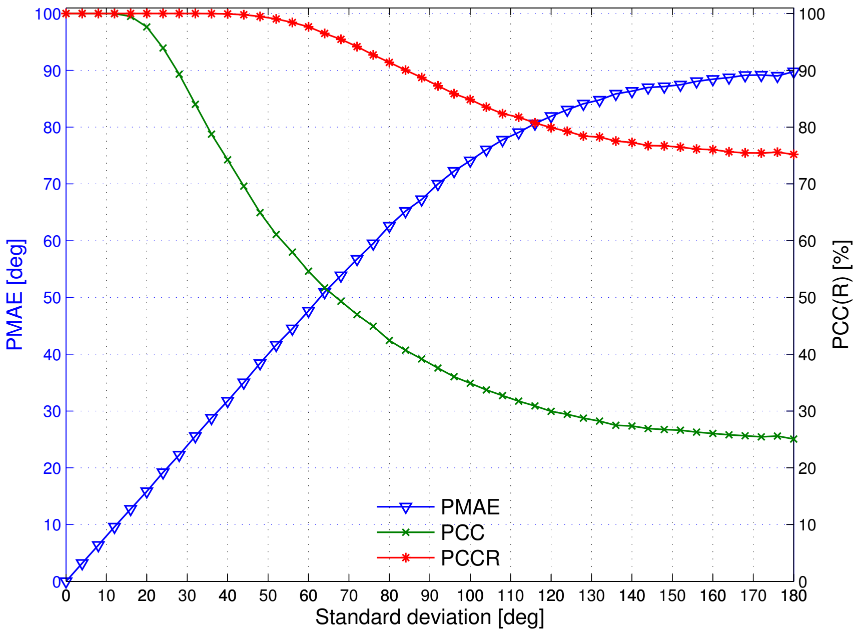


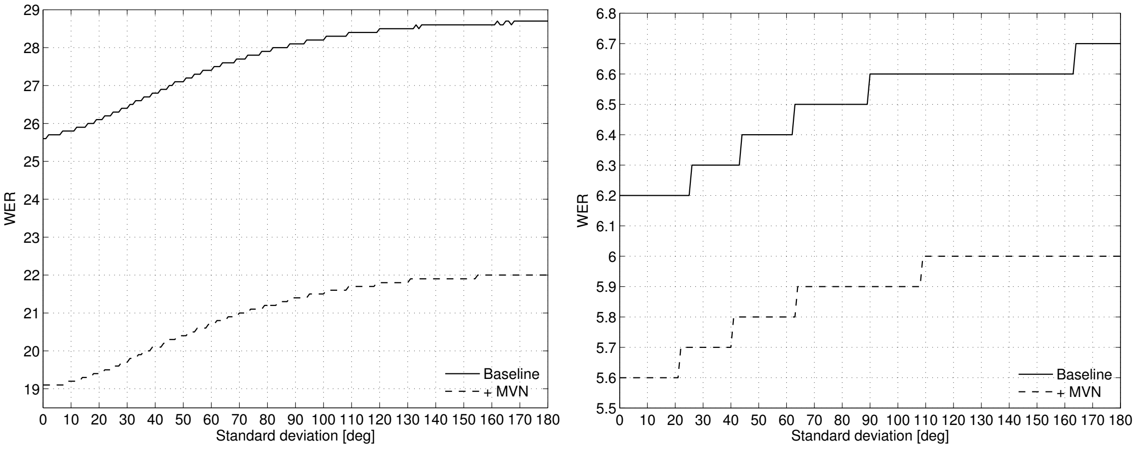
Figure 4.4: Metrics of the emulated head orientation estimation system.

In the experiments with the orientation-based CS method we emulate the system where the estimation of an angle has an error, which follows a normal distribution with the zero mean and the standard deviation varying from 0 to 180 degrees. There are tree basic metrics, agreed by the Computers in the Human Interaction Loop (CHIL) consortium, used to evaluate the head orientation estimation systems [35, 100]:

- Pan Mean Average Error (PMAE) [degrees]: the precision of the head orientation angle estimation.
- Pan Correct Classification (PCC) [%]: the ability of the system to correctly classify the head position within 8 classes spanning 45 degrees each.
- Pan Correct Classification within a Range (PCCR) [%]: the ability of the system to correctly classify the head position within 8 classes spanning 45 degrees each, allowing a classification error of ±1 adjacent class.

The metrics computed for the emulated system are shown in Figure 4.4. The PMAE grows linearly at the beginning until saturates when the standard deviation reaches 180 degrees. As expected the PCC curve decreases faster that the PCCR one, since the classification error span is much smaller. These curves also saturate at 180 degrees.

In Figure 4.5 it is shown for the clean and matched training, and using the baseline and normalized features, how the recognition performance in the UPC smart room



(a) Clean training

(b) Matched training

Figure 4.5: The orientation-based CS performance in terms of WER varying the orientation estimation error for TIDigits convolved with the RIRs from the UPC smart room.

changes, if the orientation-based CS is used and the angle estimation is not perfect. Let's assume, for example, the head orientation estimation system with the PMAE equal to 40 degrees, and PCC and PCCR equal to 62% and 99% respectively, which corresponds to the standard deviation 50 degrees in Figure 4.4. The resulting WER for the matched training case using the baseline features is 6.4%. If normalized features are used the WER is 5.8%.

In [100] the head orientation estimation system developed in the same UPC smart room was presented. In the reported experiments the PMAE of that system was equal to 11.8 degrees, which is similar to the error of the emulated system with the standard deviation 15 degrees. The reported PCC metric was lower, equal almost to 76.87% compared to 100% in the emulated case, but the PCCR is almost the same. This suggests that a very good CS performance may be expected with such system in the UPC smart room. According to this emulation it could be very close to the case when the exact speaker's orientation is known.

4.2.3 Recognition Results with Decoder-Based Measures

In Table 4.3 the recognition results in terms of WER for the CS methods using non-normalized likelihood (L), normalized likelihood using the pairwise normalization (NL), feature normalization (FN) class separability (CLS), and N-best lists (NB) are presented for convolved TIDigits. Results for random selection (RND) are the same as in the case of signal-based measures (see Table 4.2). The feature extraction and normalization scheme used in the CS measure extraction and for ASR were the same. So, if MVN was applied on the baseline features for recognition, the same normalized features were used to compute the class separability measure.

The performance of the likelihood related CS methods is every case better than that of the random selection. Again, the relative improvement with respect to the random case is shown in parenthesis. When normalized likelihood is used with the clean training, only a slight, but consistent, reduction in terms of WER may be observed with respect to the non-normalized one. However, when the acoustic model is trained in matched conditions, the relative improvement of normalized likelihood with respect to the random case is 15% or more, while the non-normalized likelihood-based CS performs almost as if the channel was selected randomly.

As described in Section 3.1.1, in the case of CS based on feature normalization, some normalization method is applied for each channel, and the speech recognition hypothesis before and after feature normalization are compared. For that purpose we used our baseline features and applied MVN[3]. The normalized streams were used to obtain the WERs displayed in the table, since they offer a better ASR accuracy than the non-normalized ones (the baseline). If the minimum distance between two compared hypothesized word sequences was the same in two or more channels, the channel was selected randomly. The CS method based on feature normalization performs the best when the normalized features with matched training are used, but in the clean training case it is surpassed by other methods, including the signal-based EV measure (see Table 4.2).

The performance of the class separability-based CS method in this setup is the worst among all tested methods, and in the matched training case it is even worse than the random selection. Possible reason for this could be that the selection decision was made on an utterance basis and, in many cases, the amount of data was not sufficient to estimate the separability measure reliably. In our implementation for the consistency we used the same classes as those used by the recognizer, i.e. digits. To get the time boundaries to separate the frames into the classes, we first applied the recognition system to signals in each channel, so the alignments are channel specific. To calculate the between class scatter matrix in Equation (3.3), at least 2 different digits (classes) have to be present in the utterance. For all utterances which contain only one kind of word, the channel has to be selected randomly. Even if the stand-alone approach was used as recommended in the original work [82] (i.e. the separation measure was calculated from classes different to those used by the recognizer), most

[3] Cepstral mean subtraction was also tested, but the performance is slightly worse.

Table 4.3: CS performance in terms of WER for convolved TIDigits using decoder-based measures.

	Clean		Matched	
CS method	Baseline	+ MVN	Baseline	+ MVN
RND	29.1	22.7	6.7	6
L	26.9 (7.6%)	20.3 (10.6%)	6.5 (3%)	5.9 (1.7%)
NL	26.6 (8.6%)	19.5 (14.1%)	5.5 (17.9%)	5.1 (15%)
FN	x	19.6 (13.7%)	x	4.6 (23.3%)
CLS	26.7 (8.2%)	22.5 (0.9%)	7.3 (-9%)	7.2 (-20%)
NB	25.4 (12.7%)	17.6 (22.5%)	5.4 (19.4%)	4.9 (18.3%)

of the utterances are so short that it would not be possible to obtain the reliable estimates anyway. To compensate for this fact, and to keep the comparison with other CS methods objective, we designed an experiment (presented in Section 4.3), where a number of utterances are used to estimate the CS measure.

The technique using N-best lists, which was developed in this thesis, is the best in average. It outperforms all the methods, in all cases, except of the feature normalization-based one, when the acoustic model was trained in matched conditions and MVN was applied to the baseline features. To extract the measure, N-best lists with N=40 were generated. The reason why this value was used, and the dependency of the ASR performance on the size and content of the list are discussed in Section 4.6.

4.2.4 Discussion

If we compare the whole set of results, we see that the CS methods using the signal-based measures perform better in the clean training case than the decoder-based ones (for all the methods except the one using N-best lists, which performs very well for all tested conditions). This is quite a surprising result, since one would expect a measure extracted from the classifier to be more correlated with the WER than a measure extracted from the signal. A possible reason for the poor performance of the class separability measure, the insufficient amount of data used to compute it, has already been suggested above. In the next section we will analyze this hypothesis for the class separability measure and also for other techniques by looking what happens if more data than one utterance is used to extract it. The TIDigits signals convolved with the UPC smart room RIRs will be used for this purpose.

4.3 Relation Between the ASR Performance and Amount of Data Used to Estimate the CS Measure

The CS measures extracted from the RIR, or related to the position and orientation of the speaker do not depend on the speech signal. Therefore, unless the room conditions change, the selected channel is the same for all utterances. The other measures are extracted directly from the speech signal or from the decoder output, so they change with each new utterance. The quality of the estimation of these measures may depend on the amount of available speech data. We have already explained how this affects the class separability-based CS measure, but the rather low performance of other decoder-based methods when the acoustic model is trained on clean speech suggests, they may suffer from a similar problem.

To investigate this further, we designed an experiment where utterances were joined into groups of increasing size. The CS measure was extracted from all files in a given group, so using progressively more and more data, and the same channel was selected for all utterances in that group. We started from one utterance and increased the group size until there was only one group with all utterances. There were 8700 files in the testing set and the groups were of equal size (factors of 8700). This means we had 8700 groups containing 1 utterance, 4350 groups with 2 utterances, 2900 groups with 3 utterances, etc. In the following, we will describe the implementation details for each CS technique in this experiment.

For the likelihood-based technique each utterance was decoded individually and the log likelihood for a group or utterances was calculated as the sum of acoustic log likelihoods of the individual utterances. Before summation, the likelihoods were normalized by the number of frames to compensate for the different utterance lengths. The channel with the highest accumulated score was selected for that group.

The normalized likelihood for a group or utterances was calculated as the sum of scores of the individual utterances. Similarly to the case of non-normalized likelihood, a normalization by the number of frames was applied to compensate for the different utterance lengths. The N-best measure for a group or utterances was extracted in a similar way.

In the case of feature normalization-based CS, the measure for a group of utterances was calculated as the sum of the individual distances between the decoded sequences from the original and the normalized feature streams, accumulating over all utterances of the group.

To extract the class separability measure for a group of utterances, the word alignments were first obtained for each channel and utterance. Then, the feature vectors from the group of utterances were pooled into their corresponding classes and the measure was computed.

To compute the EV measure reverberated utterances were first joined on the signal level, and the measure was extracted from these long files.

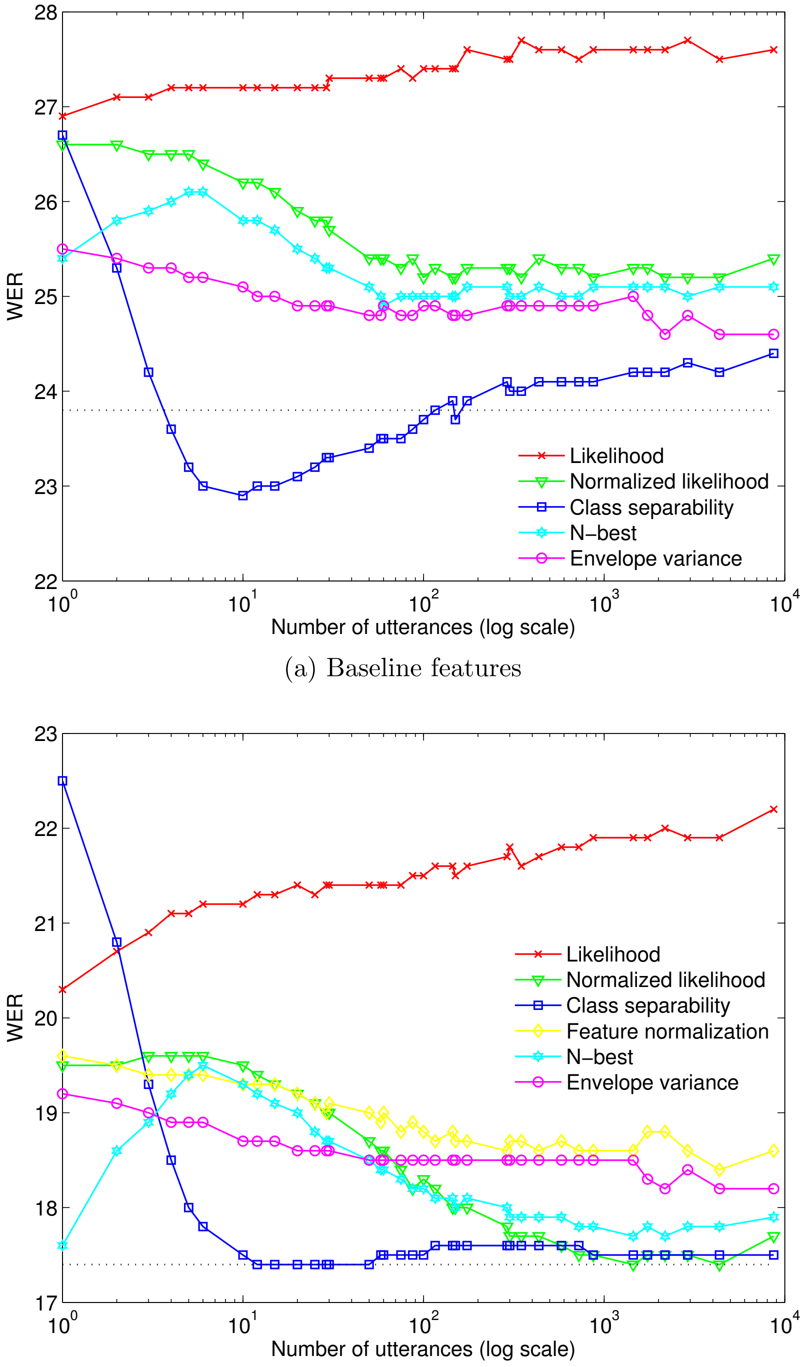
The resulting WERs for the different CS measures and clean training may be observed in Figure 4.6. On the x-axis there is the number of utterances in a group in log scale. Results are presented for the baseline features, and also for the case of applying MVN to them. When the measure is extracted combining all 8700 utterances, there is only one group and so only one channel is selected for a given position and orientation. Therefore, the best possible CS performance in such case is when the channel which gives the lowest WER in average for all utterances is chosen. Horizontal dotted lines in Figure 4.6 represent those WERs.

We can see that if likelihood is used without normalization, the ASR performance does not improve even if the number of utterances increases. Although more data are used to estimate the measure, the WER grows. The pairwise normalization clearly helps. A consistent WER decrease may be observed when the size of the data set for the measure computation increases. The same goes for the CS method based on feature normalization as shown in Figure 4.6b.

Increasing the amount of data to estimate the measure clearly helps also in case of the class separability measure. The WER decreases rapidly right from the beginning and even falls below the horizontal line for the variant with baseline features (Figure 4.6a). This fact may be confusing on the first sight, but there is a simple explanation. In fact, the horizontal dotted line, as explained above, represents the WER of the channel that gives the best WER in average for all 8700 utterances. If the best microphone in average is selected for a smaller group of utterances, a better choice can be made for that group, and consequently overall WER may be lower.

We may further observe in Figure 4.6 that increasing the amount of data for the estimation of EV measure improves the recognition performance only slightly. This is because, unlike the most of the decoder-based methods, the EV-based CS already works very well when the measure is extracted from a single utterance.

The N-best measure works the best from all the methods if only one utterance is used to extract it. If the number of utterances increases, the WER grows. This is probably because to extract the measure for a group of utterances we sum the measures from different utterances without any proper normalization. This may be a problem if probability of the observation in the denominator in Equation (3.11) varies significantly among utterances.



(a) Baseline features

(b) Baseline features + MVN

Figure 4.6: CS performance in terms of WER for convolved TIDigits and the clean training when the channel is selected per group of utterances.

When the acoustic model is trained in matched conditions, the performance for

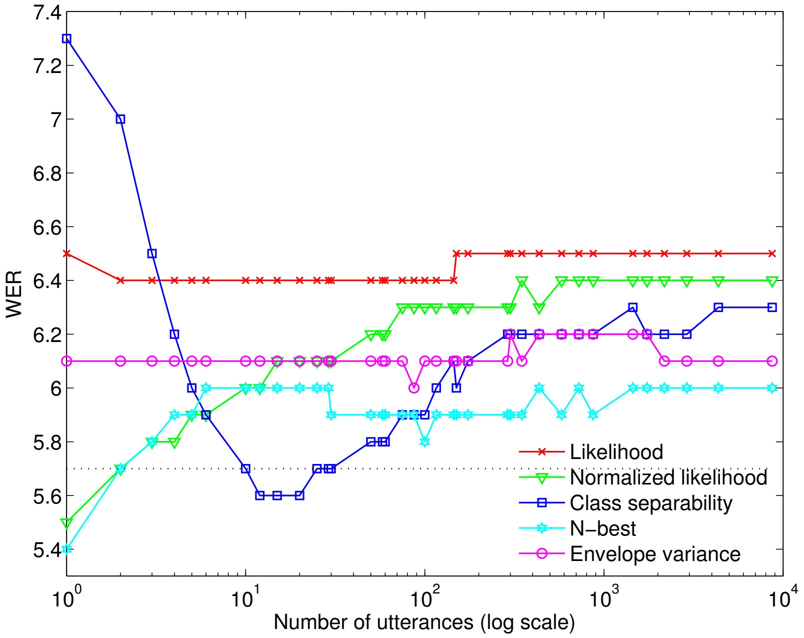


Figure 4.7: CS performance in terms of WER for the convolved TIDigits, baseline features and matched training when the channel is selected per group of utterances.

the decoder-based CS methods is better than for the signal-based ones, with the exception of the non-normalized likelihood (L) and class separability measure (CLS). As we may see in Figure 4.7 for the matched training and baseline features, increasing the amount of data to estimate the class separability measure clearly helps also in this case. This confirms the hypothesis we have made above that a sufficient amount of data to estimate the separation measure is a crucial requirement with this method. The performance of the non-normalized likelihood does not improve if the number of utterances increases, nor does the the EV-based method.

The methods based on the normalized likelihood and N-best lists perform well with the matched training, even if only one utterance is used to extract the measure. The corresponding performance curves start below the dotted line. Since for the group of utterances only the best channel in average may be selected, the WER grows.

4.3.1 Discussion

In real environments, the CS measure should be able to track the changes in a room or in the speaker's position or orientation. If the changes are slow or not very significant, the algorithm can wait, collect more speech data, and make the measurement more robust; but if it waits too much, the recognition delay may become unacceptable. So, ideally want to have a technique that requires a small amount of data to make an estimation of the measure and still leads to a good recognition performance.

Two CS methods, both developed in this thesis, based on the EV and N-best lists meet this requirement. The measure using N-best lists is computationally more expensive, but provides the best results for almost all tested ASR configurations. Alternatively, the computational simplicity and a good performance of the EV measure, even when a small amount of data is used to extract it, make it the best candidate for applications where a low latency and computational load are required, e.g. in embedded systems, for command-type applications, or scenarios where we can not afford to store and recognize speech samples from all streams.

The normalized-likelihood- and feature normalization-based measures perform better if the acoustic model is trained in matched conditions, but not so well in the clean training case, unless sufficient data is used to estimate the measure. The class separability measure may provide significant recognition performance reduction, like we may observe for example in Figure 4.6a, but it always requires more speech data to reliably estimate the separability measure. This increases the recognition delay. However, with some measure estimation tracking mechanism, it could still be employed in those scenarios which require short recognition delays. In an initial phase we could accumulate sufficient amount of speech data to get a reliable estimate of the quality measure and then, in the working phase, this estimate could be updated with each new utterance to track the changes in the room. Some decaying weighting function could be applied, to give more importance to the most recent utterances. This way we could avoid in the working phase the recognition latency problem, since the CS decision would be updated with each new utterance, we would track up to some extend the changes in the room conditions.

4.4 Combination of CS Measures

CS may benefit from combination of various methods in two ways. Firstly, since the CS measures are extracted from different domains, they may be complementary and their combination could increase the robustness of the CS system. We will refer to this case as a parallel combination. Secondly, the decoder-based methods are computationally expensive. This may be a problem if number of channels is high and fast system response is required. Therefore, to reduce the number of channels, some computationally cheap signal-based CS method may be first applied in the front-end and in the next step more precise selection can be made on the reduced channel set using the decoder-based methods. This case will be referred as a serial combination

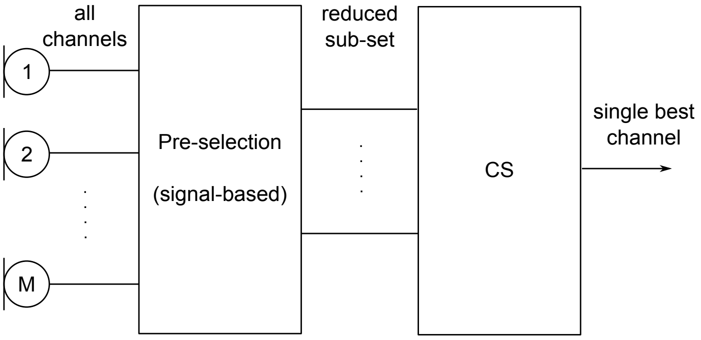


Figure 4.8: Block diagram of the serial combination of signal- and decoder-based methods.

and its block diagram is shown in Figure 4.8.

4.4.1 Parallel Combination

In the parallel combination all CS measures are extracted for all channels. Usually the measures are in different scales. In this work we applied a simple combination strategy, where we first rank the channels separately for each measure, and then select the channel with the highest ranking in average as

$$C = \arg\min_{m} \sum_{i} r_m(i), \tag{4.1}$$

where $r_m(i)$ is the ranking position of the channel m according to the measure i.

The performance of CS methods with convolved TIDigits using the parallel combination and normalized features is shown in Table 4.4. We may see that in the clean training case, combined measures perform better than any one of them alone (see Tables 4.2 and 4.3), suggesting their complementarity. The performance of the feature normalization-based method alone in the matched-training case is slightly better than when used in combination, and it is surpassed only when the class separability-based measure is excluded.

Training the acoustic models in matched conditions, as expected, improves the recognition performance. Furthermore, when the parallel combination of CS measures is used, additional consistent improvement higher than 20% relative to the random selection case may be observed. Interestingly, very similar relative improvements may be seen for the clean training case.

Table 4.4: CS performance in terms of WER using parallel combination for the clean and matched training, and normalized features.

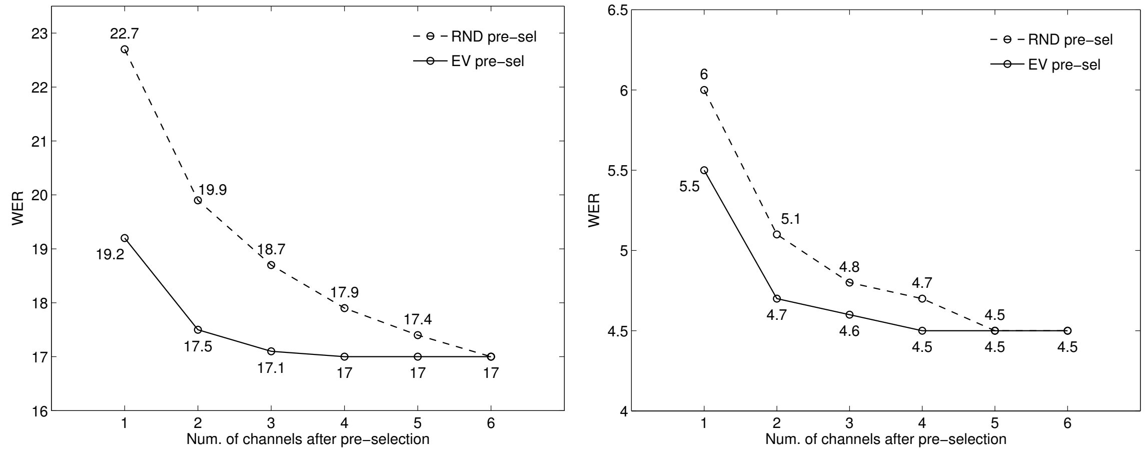
CS methods	Clean	Matched
RND	22.7	6
NL + FN + CLS	18.2 (19.8%)	4.8 (20%)
NL + FN + CLS + NB	17.3 (23.8%)	4.7 (21.7%)
NL + FN + CLS + NB + EV	16.7 (26.4%)	4.6 (23.3%)
NL + FN + NB + EV	17 (25.1%)	4.5 (25%)

4.4.2 Serial Combination

The performance of CS methods in serial combination, applying MVN to the baseline features and using the acoustic models trained in the clean and matched conditions, is shown in Figure 4.9. Each point corresponds to the number of channels after the pre-selection step. There are 2 curves for each setup, one for the random pre-selection, and the other for the case when the channel is pre-selected using the EV method. A single channel is selected from the subset in the second step using the parallel combination of several measures (NL + FN + NB + EV).

The first point of the graph (from the left) shows the recognition performance if only one channel is pre-selected, either randomly, or using the EV-based CS method. The WERs are the same as in Table 4.2 for the RND and EV case. The last point of the graph is the variant without pre-selection, when all available channels are passed to the second step, so the WERs are the same as in the last line of Table 4.4. We can observe for the clean training case (Figure 4.9a) that if EV is used to pre-select the channels, the computation load may be reduced almost by half (only 3 channels are used after pre-selection), and the relative WER increase is only 0.6% compared to the case when all channels are used. On the other hand, if channels are pre-selected randomly, the relative WER increase is 10%. When the acoustic model is trained in the matched conditions (Figure 4.9b) and we pre-select only 2 channels using the EV measure, the computational load may be lowered by 2/3, and the relative WER increase would be only 4% while for the random pre-selection it would be 13%.

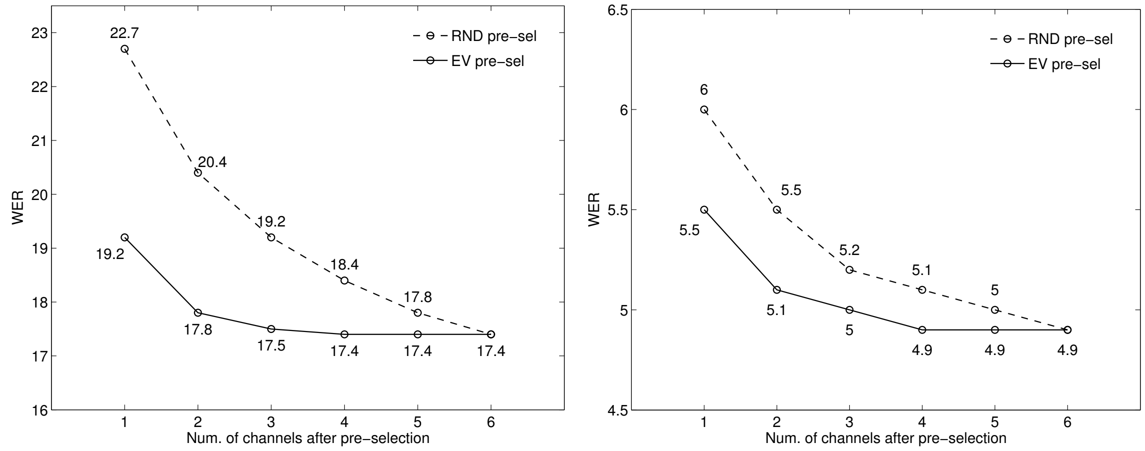
Similar behavior may be observed for the other configurations as well. For example, as shown in Figure 4.10a, when the EV- and N-best lists-based measures are used in combination, we may observe for the clean training that if half of the channels is pre-selected using the EV-based measure, the relative WER reduction is only 0.5% compared to the case when all channels are used. If the pre-selection is random, the relative WER reduction is 10%. In the matched training case reducing the number of channels by 2/3 leads to 4% and 12% relative WER increase for the EV and random pre-selection case respectively.



(a) Clean training

(b) Matched training

Figure 4.9: The CS performance in terms of WER using serial combination of the measures for convolved TIDigits. The parallel combination of various CS measures (NL+FN+NB+EV) was applied to select the channel in the second step. Normalized features were used to extract the WERs.



(a) Clean training

(b) Matched training

Figure 4.10: The CS performance in terms of WER using serial combination of the measures for convolved TIDigits. The parallel combination of the N-best list and EV-based measures was applied to select the channel in the second step. Normalized features were used to extract the WERs.

Table 4.5: WER using a single microphone for the MRD corpus.

Training	Features	6	7	E	F
Clean	Baseline	20.1	29.4	26.6	20.7
	+ MVN	15.9	21.5	21.1	17.4
Matched	Baseline	14	20.7	18.6	15.3
	+ MVN	14	16.4	17.1	15.6

4.5 Experiments with Real Distant-Talking Microphone Recordings

In this section we evaluate selected CS methods using the real distant-talking microphone recordings from the MRD corpus. Our knowledge about the recording conditions of this corpus is limited. We know there are 4 parallel channels, but do not have any additional information about the exact position or orientation of the speakers, neither about the characteristic of the acoustic channels in terms of their RIRs. Notice that the reverberated speech is not obtained by means of convolution, since it is recorded from distant-talking microphones.

The WERs for every microphone are presented in Table 4.5. Each WER shows the performance of the ASR system as if only that particular microphone was present in the room. There are 4 microphones labeled as 6, 7, E, and F in the MRD corpus. If the average WER was computed over all microphones for a given configuration, it would show the recognition performance of the system as if all the microphones were present in the room, but the signals for recognition were selected randomly. As expected, when the acoustic model trained in matched conditions is used, the WER decreases. Applying MVN to the baseline features improves significantly the ASR performance in the clean training case, but not as much in the matched training case.

4.5.1 Recognition Results for Selected CS Measures

The recognition results for a few CS methods using the envelope variance (EV), normalized likelihood (NL), feature normalization (FN), and N-best lists (NB) measures are presented for the MRD corpus in Table 4.6. Results for random selection (RND) are also included for reference.

Both, signal- and decoder-based measures are shown together. The encouraging results from the previous experiments with convolved signals are confirmed. We can see that the relative improvement with respect to the random selection case is very

Table 4.6: CS performance in terms of WER for the MRD corpus

	Clean		Matched	
CS method	Baseline	+ MVN	Baseline	+ MVN
RND	24.2	19	17.2	15.8
EV	13.2 (45.5%)	12.4 (34.7%)	10.5 (39%)	12.8 (19%)
NL	16.4 (32.2%)	13.9 (26.8%)	12.1 (29.7%)	12.7 (19.6%)
FN	x	12.8 (32.6%)	x	11.6 (26.6%)
NB	14.9 (38.4%)	12.4 (34.7%)	10.3 (40.1%)	11.2 (29.1%)

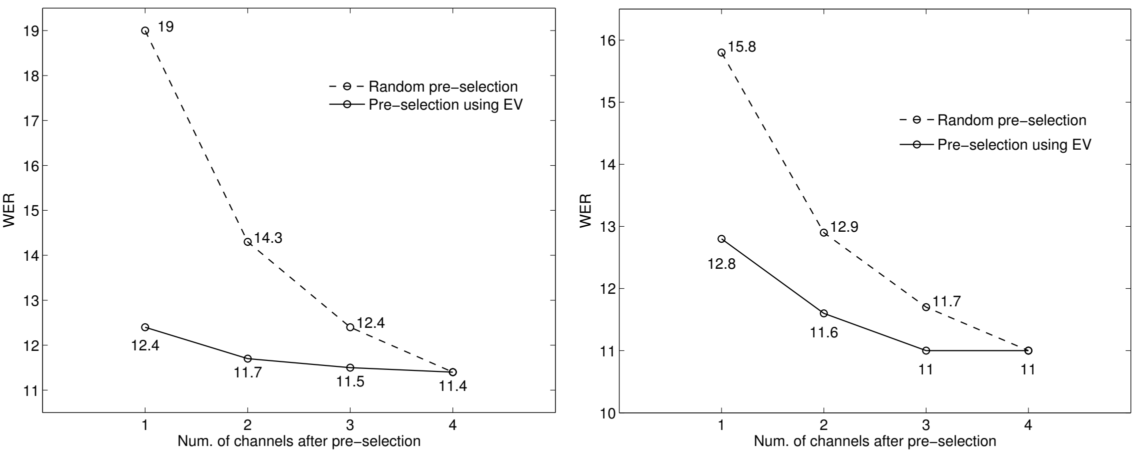
high already for individual measures. Using the EV-based CS method developed in this thesis, it may be more than 45%. Another method developed here based on the N-best lists leads to the highest relative improvement in average across all configurations. Again we may see that even if CS is combined with other robust ASR methods, matched training and feature normalization in this case, it consistently brings further recognition performance improvements. Interestingly, the WER after CS for the matched training and baseline features is lower than when MVN is applied.

4.5.2 Recognition Results for Combined CS Measures

The performance of CS measures in serial and parallel combination for the MRD corpus is shown in Figure 4.11 using the normalized features and models trained in the clean and matched conditions. As it was in the similar experiment with convolved signals, which was presented in Section 4.4.2, each point corresponds to the number of channels after the pre-selection step. There are 2 curves in each graph, one for the random pre-selection, and the other for the case when the channel is pre-selected using the EV method. A single channel is selected from the subset in the second step using the parallel combination of all CS measures from Table 4.6 (NL + FN + NB + EV).

The last point of the graph is the variant without pre-selection when all available channels are passed to the second step, and so only the parallel combination of the measures is used to select the channel. The performance of combined measures is better than when they are applied individually.

The first point of the graph shows the recognition performance if only one channel is pre-selected, either randomly, or using the EV-based CS method. The WERs are the same as in Table 4.6 for the RND and EV case. We can observe that if EV is used to pre-select the channels we may reduce the computational load by 25% almost without any loss in the recognition performance.



(a) Clean training

(b) Matched training

Figure 4.11: The CS performance in terms of WER using serial combination of the measures for the MRD corpus. Parallel combination of various CS measures (NL+FN+NB+EV) was used to select the channel in the second step. Normalized features were used to extract the WERs.

4.6 The Size and Structure of the N-Best Lists

The number of hypothesis N included in the computation of the CS measure from the N-best lists in Equation (3.12) is a free parameter. In theory, the more hypothesis are used, the more precise is the approximation of $p(O)$ in Equation (3.11). However, the objective in CS is not to estimate the posterior probability with the highest precision, but rather to minimize WER. Also, from a practical point of view, we do not want to generate large and computationally costly N-best lists.

An important factor is what kind of hypothesis is included in the CS measure estimation. The N-best algorithm may generate a lot of hypothesis that have the same word sequence as the first one, and differ only in the presence and position of the silence label[4]. The calculation of the CS measure from such a redundant list is not very efficient, because it means that we are aiming to maximize a ratio of likelihoods using almost the same word sequences.

In Figure 4.12 we show how increasing the number of hypothesis in the estimation of the N-best CS measure influences the recognition performance for the TIDigits convolved with the RIRs from the UPC smart room. Results are presented for both,

[4] For example, 'silence one two' is the same word sequence as 'one silence two'.

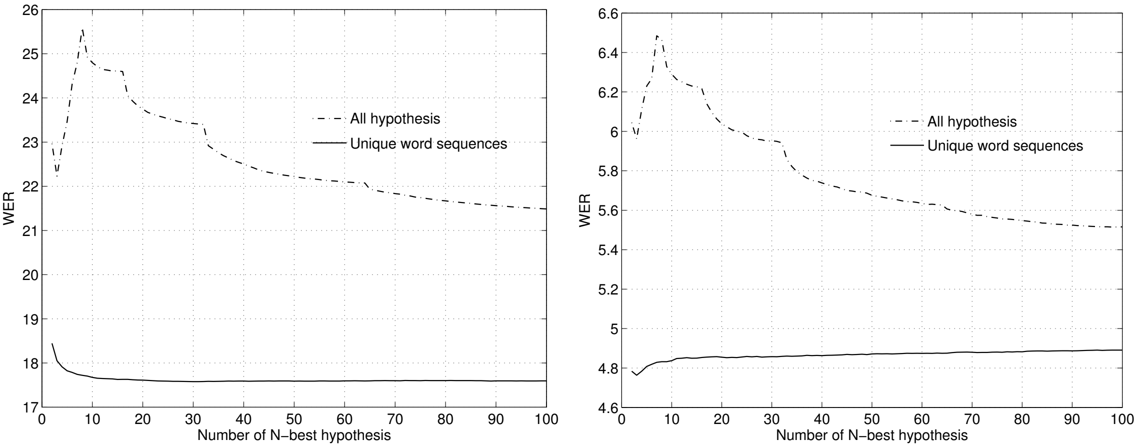
the clean and matched training. Only the results for normalized features are shown. Similar behavior may be observed for the baseline features. There are two curves presented in each graph. One curve corresponds to the N-best measure extracted using all hypothesis in the N-best list, while the other corresponds to the case when the generation of the N-best list was constrained to give only the unique hypothesis, i.e. when the redundancy of word sequences is avoided. As we can observe, the latter variant performs much better. Even if only 2 hypothesis (the first point of the graph) are used to extract the CS measure, the WER is lower than for any single channel in Table 4.1.

When the number of hypothesis increases in the clean training case, the WER decreases even more, until it saturates. Contrarily, the WER for the variant with all the hypothesis decreases much slower, and even grows at the beginning. The filtering of word sequences also helps in the matched training case, however, when the number of filtered hypothesis grows, the WER does not decrease. It actually slightly grows, until it saturates. In this case, it would be better to use the shorter N-best list, but to avoid the specific tuning for each setup, the N-best lists of the same size (N=40) were used in all the experiments where this CS measure was applied.

For the MRD corpus a similar behavior may be observed, as shown in Figure 4.13. Again, only the results for normalized features are shown. Notice, that as in the previous case with convolved TIDigits, even if only 2 hypothesis are used to extract the CS measure (the first point of the graph), the WER is lower than for any single channel in Table 4.5. In this setup, increasing the number of hypothesis leads to the WER reduction even for the matched training case.

4.7 Evaluation of the Real-Time CS Implementation in the Scenario with Moving Speakers

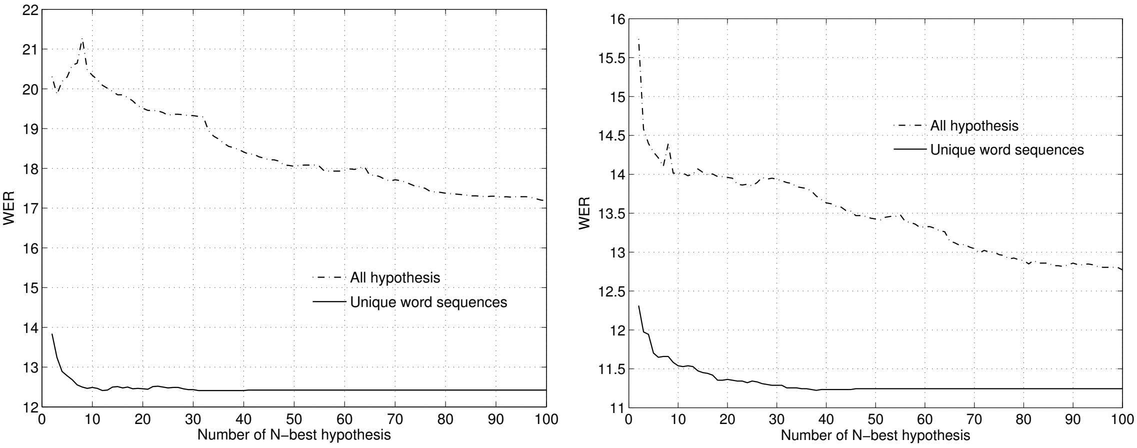
To test the real-time CS client implementation, which is described in Appendix B, but also to evaluate the CS paradigm as close as possible to the real-world conditions, we designed the following experiment. Five utterances were randomly selected for each adult speaker from the testing set of the TIDigits database, resulting in a set of 565 utterances. These were concatenated into a single long file, inserting one second of silence between them. The long file was reproduced from the loudspeaker that was held by a person moving around the UPC smart room, as if the speech was uttered by the real speakers. The position or orientation usually changed after 5 utterances (which corresponded to the same speaker), otherwise the loudspeaker moved only



(a) Clean training

(b) Matched training

Figure 4.12: CS performance in terms of WER for convolved TIDigits, using different numbers of N-best hypothesis, and all or unique word sequences. Normalized features were used to extract the WERs.



(a) Clean training

(b) Matched training

Figure 4.13: CS performance in terms of WER for the MRD corpus, using different numbers of N-best hypothesis, and all or unique word sequences. Normalized features were used to extract the WERs.

slightly.

As in the off-line experiments, only 6 microphones from the middle of each T-shaped cluster were used. For the convenience the ASR was not performed online, however, this does not affect the conclusions regarding the CS performance. The whole session was recorded by each microphone and stored to the hard-drive. In the same time, the real-time CS client was selecting the microphone and creating the 7^{th} stream that was also stored on the hard-drive. Finally, the files were sliced using the

Table 4.7: WER using a single microphone in the experiment with moving speakers.

Features	Training	2	6	10	14	18	22
Baseline	Clean	39.8	41.2	40.3	38.5	38.3	36.1
	Matched	27	26.1	30.7	31.1	34.4	30
+ MVN	Clean	30.7	27.5	30.5	26.8	31.1	27.8
	Matched	31.1	28.9	30.1	27.6	28.9	28.5

known time-stamps of the original utterances and the stream from each channel was recognized using the same ASR system as in the off-line experiments.

The WERs for each channel, using the baseline features (MFCC + Δ + $\Delta\Delta$), or applying MVN to them, and testing with clean and reverberated acoustic model are shown in Table 4.7. The MVN improves the recognition performance significantly. This is probably because it compensates for the stationary noise that is generated by the cooling fans of the servers in the adjacent room.

The same acoustic models as in the experiments with artificially reverberated signals were used (see Section 4.1). Matched training improves the recognition performance in case of the baseline features, but not as much as in the tests with convolved speech (see Table 4.1). This is probably because the speaker was not stationary and the signals in testing were recorded by distant-talking microphones, not reverberated by convolution with the measured RIR. Actually, when MVN is applied and the acoustic model trained in matched conditions is used, the recognition performance is even slightly worse in average.

If we use the same procedure to recognize the speech signal from the real-time CS client (the 7^{th} stream), we get significant improvements, as shown in Table 4.8. The improvement relative to the random case (RND), shown in the brackets, is more than 15% when testing with the clean speech models. Even if we use matched training or apply MVN, we still get almost 10% relative improvement. Note also, that after CS the WER is lower than if we selected the best single channel, as may be seen in Table 4.7.

When the off-line version of the CS algorithm is applied like in the experiments in Section 4.2.1, similar WERs are observed. By this experiment we not only tested and proved that the implementation of the the real-time CS is correct, but also shown that CS works well if applied in conditions similar to the real meeting scenario. Simple yet efficient and robust EV-based CS method developed in this thesis may further improve the recognition performance in this scenario, even when combined with other

Table 4.8: The performance of real-time CS in terms of WER.

Features	Training	RND	$\mathrm{EV}_{\mathrm{RT}}$
Baseline	Clean	39	34.3 (12.1%)
	Matched	29.9	25.3 (15.4%)
+ MVN	Clean	29.1	24.6 (15.5%)
	Matched	29.2	26.3 (9.9%)

methods for robust ASR, like MVN.

4.8 Summary

In this chapter we compared all signal- and decoder-based CS measures. Both clean speech data convolved with recorded RIRs as well as real distant-talking microphone recordings were used in the tests. The experiments show that CS may provide large recognition improvements, in some cases up to almost 46% relative, compared to the case when the channel is selected randomly. If it is used in combination with other robust ASR techniques, like matched training or mean and variance normalization, it was observed that the recognition improvements from both approaches are cumulative up to some extent.

We also evaluated the combination of different CS measures and showed that simple combinations of signal and decoder-based methods lead to further WER reduction. Also, the combination may help to reduce the computational load by half with just a slight loss in terms of WER. In fact, the computationally cheap EV signal-based method is used at the front to reduce the number of input channels, so the more expensive decoding operation does not have to be made for all channels.

Two CS methods, both developed in this thesis, based on the EV and N-best lists showed a very good performance for all experimental setups. The N-best lists-based method is computationally more expensive, but provides the best results in average. Alternatively, a computationally simple EV measure often performed better than other complex decoder-based approaches and is a good candidate for the applications where low recognition delay and computation load are required, as was demonstrated in the tests of the real-time CS client implementation.

Chapter 5

Integration of Partly Decorrelated Features into REMOS

The main topic discussed in this thesis so far was CS. We have shown that it may help to improve the ASR performance significantly, but indeed there is room for further improvements. The contribution of CS depends strongly on the conditions in the room. There may be situations when all distant microphones record similarly distorted signals, so the space for improvement by selecting a channel is small. Also, even if we were always able to select the channel with the minimum WER, we would hardly reach the performance of the recognition system using clean speech. Additional techniques are therefore needed to further improve the robustness of the ASR systems.

In this chapter we introduce a new feature extraction method into REMOS [8], a generic framework for robust distant-talking speech recognition. We start with brief description of REMOS, and explain the restrictions imposed on the speech features due to the complexity of the optimization problem in the decoding. Because of those restrictions, only highly correlated logmelspec (logarithmic mel-spectral) energies had been used before in REMOS. This limited the ASR performance because the assumption of diagonal covariance matrices in the HMMs output probability density functions was not met. To overcome this, a new set of partly decorrelated features derived from FF [9, 16] was proposed.

This work is a result of collaboration with the colleagues from the chair of Multimedia Communications and Signal Processing at the University of Erlangen-Nuremberg, Erlangen, Germany, where the original REMOS framework was developed. The idea to integrate FF into REMOS comes from the author of this thesis, but the main credit belongs to Roland Maas, who derived the mathematical formulations, as well as conducted and described the experiments. The text in this chapter comes mostly from the joint publication [101].

5.1 Introduction to REMOS

The main reason for a low ASR performance in presence of additive noise or reverberation is a mismatch between the training and testing conditions. Usually different single or multi-microphone speech enhancement techniques are used to reduce the distortion, or the adaptation of the speech features and/or acoustic models of the ASR system is applied to compensate for the mismatch. REMOS belongs to the latter group.

A common, straight-forward way to reduce the mismatch is to train the acoustic models on the matched data [102]. The disadvantage is that different conditions require a costly retraining. Multi-condition style of training may be used [103], but the performance will not be as good as in the matched case. Various HMM-based adaptation methods have been presented, for example in [70] and [68], but they make the conditional independence assumption, i.e. that the current output vector depends only on the current state, which holds even less in the presence of reverberation.

REMOS tries to overcome these limitations. It is based on the combination of two kinds of models, a HMM network describing the clean speech, and a reverberation model describing the effect of reverberation directly in the feature domain. An extended version of the Viterbi algorithm is used during the recognition to determine the most likely contributions of the clean speech HMM and the reverberation model to the current reverberant observation. Since both models can be estimated independently, only a re-estimation of the reverberation model is required in case the reverberation conditions change, which can be done very efficiently. Costly data collection from the new environment and retraining of the speech acoustic models are not necessary.

5.1.1 Review of the REMOS Concept

Following notational distinction will be used in this section: Every vector $\mathbf{v}$ without the explicit subscript "mel" or "FF" is meant to be in the logmelspec domain, whereas the corresponding vectors $\mathbf{v}_{\text{mel}}$ and $\mathbf{v}_{\text{FF}}$ denote the melspec and frequency filtered representation of $\mathbf{v}$, respectively. Furthermore, the operator "exp" applied to vectors is meant componentwise. The number of mel-channels is denoted by L. Single-Gaussian probability density functions with mean-vector μ and covariance matrix $\mathbf{C}$ are abbreviated by $\mathcal{N}(\mu, \mathbf{C})$.

The REMOS framework consists of two major elements: a clean speech HMM and a Reverberation Model (RVM). The RVM is room specific and captures the following information, whose role will be explained later:

- M melspec-feature vectors

$$\mu_{\mathbf{h}_{\text{mel}}(0)}, \ldots, \mu_{\mathbf{h}_{\text{mel}}(M-1)} \in \mathbb{R}^L$$

  being a statistical description of the room impulse response partitioned into M frames,

- a probability density function $f_{\mathbf{h}(0)} = \mathcal{N}(\mu_{\mathbf{h}(0)}, \mathbf{C}_{\mathbf{h}(0)})$ describing the early part $\mathbf{h}(0, k)$ of the room impulse response in the logmelspec domain,

- and a probability density function $f_{\mathbf{a}} = \mathcal{N}(\mu_{\mathbf{a}}, \mathbf{C}_{\mathbf{a}})$ capturing the weighting $\mathbf{a}(k)$ of the late reverberation estimation in the logmelspec domain.

A detailed description of RVM estimation techniques can be found in [8].

The combination of both the HMM and the RVM is based on the so-called "melspec-convolution" which assumes that the reverberant observation vector $\mathbf{x}_{\text{mel}}(k)$ is given in the melspec domain by

$$\mathbf{x}_{\text{mel}}(k) = \sum_{m=0}^{M-1} \mathbf{h}_{\text{mel}}(m, k) \odot \mathbf{s}_{\text{mel}}(k-m),$$

where $\mathbf{h}_{\text{mel}}(m, k)$ and $\mathbf{s}_{\text{mel}}(k-m)$ denote the melspec representation of the room impulse response and the clean speech frames, respectively, and $\odot$ denotes the Hadamard product. During recognition, REMOS estimates both contributions $\mathbf{h}_{\text{mel}}(m, k)$ and $\mathbf{s}_{\text{mel}}(k-m)$ based on only the observation $\mathbf{x}_{\text{mel}}(k)$, the HMM, and the RVM. To this end, the melspec convolution is first of all simplified in the following way:

$$\mathbf{x}_{\text{mel}}(k) = \mathbf{h}_{\text{mel}}(0, k) \odot \mathbf{s}_{\text{mel}}(k) + \mathbf{a}_{\text{mel}}(k) \odot \widehat{\mathbf{x}}_{r,\text{mel}}(k), \tag{5.1}$$

where

$$\widehat{\mathbf{x}}_{r,\text{mel}}(k) = \sum_{m=1}^{M-1} \mu_{\mathbf{h}_{\text{mel}}(m)} \odot \mathbf{s}_{\text{mel}}(k-m) \tag{5.2}$$

is an approximation of the late reverberant component

$$\sum_{m=1}^{M-1} \mathbf{h}_{\text{mel}}(m, k) \odot \mathbf{s}_{\text{mel}}(k-m).$$

By transforming (5.1) to the logmelspec domain, we obtain a description for the observed reverberant feature vector sequence $\mathbf{x}(k)$:

$$\exp\big(\mathbf{x}(k)\big) = \exp\big(\mathbf{h}(0, k) + \mathbf{s}(k)\big) + \exp\big(\mathbf{a}(k) + \widehat{\mathbf{x}}_r(k)\big). \tag{5.3}$$

For recognition, an extended version of the Viterbi algorithm is employed to determine the most likely contributions of the HMM, i.e., $\mathbf{s}(k)$, as well as of the RVM, i.e., $\mathbf{h}(0,k)$ and $\mathbf{a}(k)$. At each step of the extended Viterbi algorithm, the Viterbi score is therefore weighted by the outcome of the following inner optimization problem:

$$\begin{aligned} \max_{\mathbf{s}(k),\mathbf{h}(0,k),\mathbf{a}(k)} \; & f_{\mathbf{s}}\big(\mathbf{s}(k)\big) \cdot f_{\mathbf{h}(0)}\big(\mathbf{h}(0,k)\big) \cdot f_{\mathbf{a}}\big(\mathbf{a}(k)\big) \\ & \text{subject to (5.3)}, \end{aligned} \tag{5.4}$$

where $f_{\mathbf{s}}$ is the single-Gaussian output density of the current HMM state and the late reverberation $\widehat{\mathbf{x}}_r(k)$ is calculated by using estimates of $\mathbf{s}(k-m)$, $m = 1, ..., M-1$, cf. (5.2), known from former Viterbi steps [8].

Once the optimization problem (5.4) is solved, the calculated optimal contributions $\widehat{\mathbf{s}}(k)$, $\widehat{\mathbf{h}}(0,k)$, and $\widehat{\mathbf{a}}(k)$ are inserted into the objective functions of (5.4), i.e., the pdfs of the HMM and RVM, to obtain the model score

$$O_j(k) = f_{\mathbf{s}}\big(\widehat{\mathbf{s}}(k)\big) \cdot f_{\mathbf{h}(0)}\big(\widehat{\mathbf{h}}(0,k)\big) \cdot f_{\mathbf{a}}\big(\widehat{\mathbf{a}}(k)\big),$$

by which the Viterbi score is weighted.

5.2 Extension to Partly Decorrelated Features

Conventional ASR systems usually employ a set of energies extracted from frequency bands distributed in a mel-scale at some stage of the feature extraction process. A non-linear operation (usually logarithm) is applied to compress the large range of the amplitudes of the spectral measurements, resulting in logmelspec features. As mentioned before, their decorrelation is needed, which is usually achieved by some linear transformation. The most common transformation is the DCT leading to the well known MFCC representation. FF is an alternative way to decorrelate the features with comparable performance to MFCC. The transformation to the decorrelated domain is achieved using a simple FIR filter, usually of order 2 (instead of DCT), which simplifies integration into REMOS. In the following, this transformation will be denoted by a matrix $\mathbf{S} \in \mathbb{R}^{L \times L}$, i.e.,

$$\mathbf{v}_{\mathrm{FF}} = \mathbf{S}\mathbf{v}.$$

However, before explaining the detailed structure of $\mathbf{S}$, we discuss the influence of this feature transformation on the optimization problem (5.4).

5.2.1 Global Solution of the Inner Optimization Problem

One of the main issues in the REMOS decoding is the treatment of the inner optimization problem (5.4). In order to obtain reliable clean-speech estimates $\widehat{\mathbf{s}}(k)$, one should aim for determining a global solution to (5.4). State-of-the-art optimization algorithms cannot assure convergence to a global optimum but only to an arbitrary local one. In this section, we therefore focus on the mathematical examination of the inner optimization problem to determine a global solution while pointing out some restrictions to the form of $\mathbf{S}$.

5.2.1.1 Reformulation

First of all, we repeat the reformulation steps according to [104] and equivalently decompose (5.4) into two subproblems:

$$\max_{\mathbf{x}_0(k),\mathbf{a}(k)} \; f_{\mathbf{x}_0}\big(\mathbf{x}_0(k)\big) \cdot f_{\mathbf{a}(k)}\big(\mathbf{a}(k)\big)$$

$$\text{s.t. } \exp\big(\mathbf{x}(k)\big) = \exp\big(\mathbf{x}_0(k)\big) + \exp\big(\mathbf{a}(k) + \widehat{\mathbf{x}}_r(k)\big) \tag{5.5}$$

and

$$\max_{\mathbf{s}(k),\mathbf{h}(0,k)} \; f_{\mathbf{s}}\big(\mathbf{s}(k)\big) \cdot f_{\mathbf{h}(0)}\big(\mathbf{h}(0,k)\big)$$

$$\text{s.t.} \mathbf{x}_0(k) = \mathbf{s}(k) + \mathbf{h}(0,k), \tag{5.6}$$

where $\mathbf{x}_0(k) = \mathbf{s}(k) + \mathbf{h}(0,k)$ denotes the direct sound component and $f_{\mathbf{x}_0} = \mathcal{N}(\mu_{\mathbf{x}_0}, \mathbf{C}_{\mathbf{x}_0})$ is the corresponding probability density function. As (5.6) is explicitly solvable, we henceforth focus on the solution of (5.5). For sake of readability, we normalize the following quantities and drop the vector indices k:

$$\begin{aligned} \mathbf{u} &= \mathbf{x}_0 - \mathbf{x}, \\ \mathbf{w} &= \mathbf{a} + \widehat{\mathbf{x}}_r - \mathbf{x}, \\ \mu_{\mathbf{u}} &= \mu_{\mathbf{x}_0} - \mathbf{x}, \\ \mu_{\mathbf{w}} &= \mu_{\mathbf{a}} + \widehat{\mathbf{x}}_r - \mathbf{x}. \end{aligned} \tag{5.7}$$

The objective function to be maximized can be reformulated such that we obtain a quadratic functional to minimize. Hence, (5.5) becomes

$$\begin{aligned} \min_{\mathbf{u},\mathbf{w}\in\mathbb{R}^L} \quad & \frac{1}{2}(\mathbf{u} - \mu_{\mathbf{u}})^T \mathbf{C}_{\mathbf{x}_0}^{-1} (\mathbf{u} - \mu_{\mathbf{u}}) \\ & + \frac{1}{2}(\mathbf{w} - \mu_{\mathbf{w}})^T \mathbf{C}_{\mathbf{a}}^{-1} (\mathbf{w} - \mu_{\mathbf{w}}). \end{aligned}$$

$$\text{s.t. } \exp(\mathbf{u}) + \exp(\mathbf{w}) = 1. \tag{5.8}$$

5.2.1.2 Decomposition

In contrast the previous REMOS implementation, we now allow the logmelspec covariance matrices $\mathbf{C}_{\mathbf{x}_0}^{-1}$ and $\mathbf{C}_{\mathbf{a}}^{-1}$ to be non-diagonal, i.e., we assume diagonality for the corresponding matrices $\mathbf{C}_{\mathbf{x}_0,\mathrm{FF}}^{-1}$ and $\mathbf{C}_{\mathbf{a},\mathrm{FF}}^{-1}$ in the FF domain. These matrices are related by

$$\begin{aligned}\mathbf{C}_{\mathbf{x}_0}^{-1} &= \mathbf{S}^T\mathbf{C}_{\mathbf{x}_0,\mathrm{FF}}^{-1}\mathbf{S},\\ \mathbf{C}_{\mathbf{a}}^{-1} &= \mathbf{S}^T\mathbf{C}_{\mathbf{a},\mathrm{FF}}^{-1}\mathbf{S}.\end{aligned}$$

In order to determine a global solution of the problem (5.8), all its local solutions have to be calculated. This is computationally extremely demanding since (5.8) can have up to 2^L local minima [105], i.e., more than 16 million minima for $L = 24$. To face this problem, we restrict the form of $\mathbf{S}$ in order to split (5.8) into small subproblems.

Thus, $\mathbf{S}$ is considered to be a block matrix under permutation, designed in such way that if two logmelspec channels are used to calculate a transformed coefficient, neither of those channels is used again in combination with other logmelspec channel. This that the pairs are exclusive, i.e.,

$$\mathbf{SP} = \mathring{\mathbf{S}} = \begin{pmatrix}\mathring{\mathbf{S}}_1 & & \\ & \ddots & \\ & & \mathring{\mathbf{S}}_{L/2}\end{pmatrix} \in \mathbb{R}^{L\times L} \tag{5.9}$$

with

$$\mathring{\mathbf{S}}_i \in \mathbb{R}^{2\times 2},\ i = 1,\cdots,L/2,$$

and $\mathbf{P} \in \mathbb{R}^{L\times L}$ being an appropriate permutation matrix. Hence, the logmelspec covariance matrices can as well be transformed to a block structure by permutation:

$$\mathbf{P}^T\mathbf{C}_{\mathbf{x}_0}^{-1}\mathbf{P} = \mathring{\mathbf{C}}_{\mathbf{x}_0}^{-1} = \begin{pmatrix}\mathring{\mathbf{C}}_{\mathbf{x}_0,1}^{-1} & & \\ & \ddots & \\ & & \mathring{\mathbf{C}}_{\mathbf{x}_0,L/2}^{-1}\end{pmatrix} \in \mathbb{R}^{L\times L}$$

with

$$\mathring{\mathbf{C}}_{\mathbf{x}_0,i}^{-1} \in \mathbb{R}^{2\times 2},\ i = 1,\cdots,L/2.$$

We define $\mathring{\mathbf{C}}_{\mathbf{a}}^{-1}$ in the same way. Analogously, each of the quantities $\mathbf{u}$, $\mu_{\mathbf{u}}$, and $\mu_{\mathbf{w}}$ are permuted according to

$$\mathbf{P}^T\mathbf{w} = \mathring{\mathbf{w}} = \begin{pmatrix}\mathring{\mathbf{w}}_1\\ \vdots \\ \mathring{\mathbf{w}}_{L/2}\end{pmatrix} \in \mathbb{R}^L,\ \mathring{\mathbf{w}}_i \in \mathbb{R}^2, \tag{5.10}$$

which finally allows the decomposition of (5.8) into $L/2$ subproblems for each pair of coupled logmelspec channels:

$$\begin{aligned} \max_{\mathring{\mathbf{u}}_i, \mathring{\mathbf{w}}_i \in \mathbb{R}^2} \quad & \frac{1}{2}(\mathring{\mathbf{u}}_i - \mathring{\mu}_{\mathbf{u},i})^T \mathring{\mathbf{C}}_{\mathbf{x}_0,i}^{-1} (\mathring{\mathbf{u}}_i - \mathring{\mu}_{\mathbf{u},i}) \\ & + \frac{1}{2}(\mathring{\mathbf{w}}_i - \mathring{\mu}_{\mathbf{w},i})^T \mathring{\mathbf{C}}_{\mathbf{a},i}^{-1} (\mathring{\mathbf{w}}_i - \mathring{\mu}_{\mathbf{w},i}) \\ & \text{s.t. } \exp(\mathring{\mathbf{u}}_i) + \exp(\mathring{\mathbf{w}}_i) = 1. \end{aligned} \tag{5.11}$$

Note that, due to the decomposition of (5.8) into several lower dimensional optimization problems (5.11), the number of local minima is reduced to 4 for each $i = 1, ..., L/2$, whereas the overall problem (5.8) would have had more than 16 million local minima for $L = 24$.

The decomposed problem (5.11) is now numerically solved by piecewise linear approximation of the non-linear constrained as proposed in [104]. Once this has been effected for each $i = 1, ..., L/2$, the permutation (5.10) and the normalization (5.7) are undone, and hence a numerical global solution $\widehat{\mathbf{x}}_0$, $\widehat{\mathbf{a}}$ of the optimization problem (5.5) is obtained.

However, the restriction (5.9) we imposed to the form of $\mathbf{S}$ influences the way features are extracted. In the following, we introduce the FF approach and show it can be easily modified to meet the constraint (5.9). In contrast, a DCT matrix cannot be adapted to fulfill the condition (5.9). For this reason, implementation of MFCCs is not feasible.

5.2.2 Frequency Filtering in REMOS

The main goal of FF is to decorrelate the sequence of logmelspec energies. As explained before in Subsection 2.1.1, decorrelation is achieved by filtering the sequence of logmelspec energies in the spectral domain with a simple filter (2.4). Recall that the impulse response of the filter is $h(k) = \{1, 0, -1\}$, so the filtering operation consists of a subtraction of the two bands adjacent to the current one, and during filtering, zeros are assumed before the first and after the last logmelspec coefficient.

Thus, the transformation matrix **S** using filter (2.4) for 8 channels looks as

$$\mathbf{S_{FF}} = \begin{pmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{5.12}$$

As explained before, **S** has to be designed in such way, that it meets the constraints in (5.9). In particular, A.) the size of the matrix is $L \times L$, i.e., the number of coefficients is not reduced after the transformation, and B.) when 2 logmelspec channels are used in the transformation to calculate some coefficient, neither of those two channels may be used with any other logmelspec channel in the pair again.

The transformation matrix in the original FF implementation (5.12) does not meet the second constraint, because there are logmelspec channels that appear in the calculation of transformed coefficients in several pairs. For example, the fourth channel is used with the second to calculate one FF coefficient and it is used again with the sixth channel to calculate another FF coefficient. This problem may be solved by removing lines from the matrix in such way that only exclusive pairs remain, resulting in

$$\mathbf{S'_{FF}} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}.$$

Since the number of features in the transformed domain has to be the same as in the original one to meet (5.9), the suppressed lines must be replaced. One option is to simply insert zeros, but preliminary tests showed that the discrimination among classes would be decreased. Instead, we replace the missing values by the average of logmelspec coefficients adjacent to the current channel in order to include the information about the original shape of the spectra. The rows in the resulting transformation matrix are orthogonal to each other and the matrix for 8 channels has the following form:

$$\mathbf{S} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}. \tag{5.13}$$

Averaging may be seen as if the filter with impulse response $h(k) = \{1, 0, 1\}$ was applied.

In Fig. 5.1, a graphical interpretation of the transformation above is shown. The full band is actually split into several independent bands. For each band, which includes 4 logmelspec features, 4 new FF based features, coupled in a band, are derived by a regular linear transformation according to (5.13). The FF_- coefficients are calculated using the filter (2.4) and are measuring the slope of the spectra. The complementary coefficients FF_+ are related the spectral amplitude.

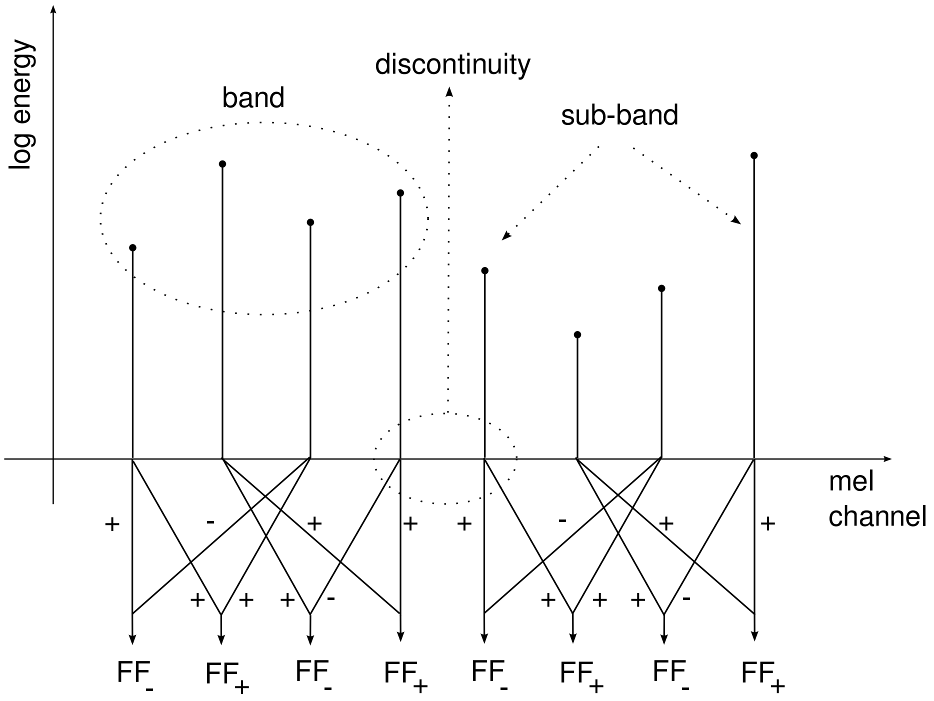


Figure 5.1: Transformation from logmelspec domain into partly decorrelated FF domain.

5.3 Experiments

Experiments with the connected digit recognition task were carried out to analyze the performance of REMOS using the new feature extraction method. This task is chosen for evaluation since the probability of the current digit can be assumed to be independent of the preceding digits so that the recognition rate is solely determined by the quality of the acoustic model of REMOS.

5.3.1 Experimental Setup

To get the reverberant test data, the clean-speech TIDigits data were convolved with different RIRs measured at different loudspeaker and microphone positions in five rooms with the characteristics given in Table 5.1. A strict separation of training and test data is maintained in all experiments both for speech and RIRs. Each test utterance is convolved with an RIR selected randomly from a number of measured RIRs in order to simulate changes of the RIR during the test.

The REMOS-based recognizer is implemented by extending the decoding routines of HTK [98]. The speech data is sampled at 20 kHz. 16-state word-level HMMs with single-Gaussian densities serve as clean-speech models. Two different versions of REMOS are compared: the former REMOS in the logmelspec domain as well as the new REMOS with the proposed FF based features extracted from the logmelspec domain using transformation similar to (5.13). Both types have been tested with 16 and 24 mel-bands.

To obtain the baseline results, we employed an HTK recognizer trained on matched reverberant data with 13 MFCCs as well as 13 delta coefficients and 3 Gaussian output densities per HMM state. Furthermore, we applied the MLLR technique [98] to two clean speech recognizers based on A) 13 MFCCs, 13 deltas, 3 Gaussians, and B) 24 MFCCs with single Gaussian densities, respectively. Hereby, only the mean vectors of the output densities are transformed using 44 matched reverberant utterances and a regression class tree of 32 base classes.

Further details about experimental setup can be found in [8]. Note that although the task is the same, connected digits, the setup used here is different to the one used in the previous chapters to evaluate the CS performance.

5.3.2 Experimental Results

Table 5.2 compares the word accuracy of the different REMOS versions with a state-of-art recognizer trained on clean and on matched reverberant data, respectively. In

Room	Type	T_{60}	d	SRR
R1	lab	300 ms	2.0 m	4 dB
R2	conf. room	600 ms	2.0 m	0.5 dB
R3	conf. room	780 ms	2.0 m	0.5 dB
R4	studio	700 ms	4.1 m	-4 dB
R5	lecture room	900 ms	4.0 m	-4 dB

Table 5.1: Summary of room characteristics: T_{60} is the reverberation time, d is the distance between speaker and microphone, and SRR denotes the Signal-to-Reverberation-Ratio.

all cases, one can observe a performance degradation with increasing reverberation time. For all rooms the REMOS implementations show a significant improvement compared to the clean-speech and the MLLR (1G+MFCC) adapted recognizer. In room R5, they even outperform the MLLR adaptation of the 3G+MFCC+delta system.

The 24 and 16 band versions perform quite similar in combination with logmelspec features. However, stepping towards FF based features and reducing the number of sub-bands from 24 to 16 brings a remarkable relative decrease in WER of up to 29%. This clearly underlines the benefit of the proposed modified FF method. Concerning the additional gain by sub-band reduction, one can think of three possible reasons:
1.) The high-indexed cepstral coefficients, which are intrinsically noisy, are less weighted by the FF operation [16].
2.) Given the frequency filter (2.4), the correlation of the new features may be lower since the sub-bands get further apart in frequency.
3.) The number of bands with 4 features decreases from 6 to 4, so there are less discontinuities (see Fig. 5.1) in the final transformation.

We would like to underline that, although REMOS is only based on static features and HMMs are trained on clean speech with single-Gaussian densities, it almost reaches the performance of the state-of-the-art recognizer (3G+MFCC+delta) trained on the matched reverberant data in the most reverberant room R5.

5.4 Summary

In this chapter, we presented the incorporation of partly decorrelated features based on FF into the REMOS concept. Modifications were made on both sides, in the feature extraction, as well as inside the inner optimization problem of REMOS. Con-

Training	Parameters	R1	R2	R3	R4	R5
clean	1G+MFCC	85.7	62.5	60.4	52.5	47.7
clean+MLLR	1G+MFCC	88.7	84.1	81.2	83.6	70.4
REMOS	1G+logmel, 24 bands	90.5	90.5	88.9	87.5	88
REMOS	1G+logmel, 16 bands	90.7	89.8	89.1	88.2	87.7
REMOS	1G+proposed FF, 24 bands	91.9	91.8	90.8	89.3	90.2
REMOS	1G+proposed FF, 16 bands	93.4	92.2	91.6	90.5	90
clean+MLLR	3G+MFCC+Δ	96.6	94.8	91.2	92.5	83.4
matched rev.	3G+MFCC+Δ	98.2	96.6	95	95.8	91.7

Table 5.2: Comparison of word accuracies in % for rooms R1 to R5 and different recognizers.

nected digit recognition experiments confirm that recognition strongly benefits from new partly decorrelated features and show a significant and consistent improvement in WER of up to 29% compared to the former logmelspec implementation.

REMOS can efficiently be adapted to changing acoustic conditions by simply re-estimating the RVM, whereas a matched trained recognizer would have to be retrained, which is in general computationally very demanding. Based on the experimental results we can furthermore hypothesize that the HMMs' conditional independence assumption is more inaccurate when the reverberation time increases. Even matched reverberant training cannot compensate this restriction to an arbitrary extend.

Those facts indicate the limitations of conventional HMM-based recognizers in reverberant environments and, at the same time, the need of reverberation modeling techniques such as REMOS. There are several options for further improvements by extending REMOS to multi-Gaussian densities and dynamic features.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis addressed the problem of ASR with distant-talking microphones in a room environment, where reverberation is the dominant source of distortion. Several solutions for both, single- and multi-microphone setups were described. Two main research lines were followed: Channel Selection (CS) and feature extraction in the framework of REMOS.

The CS is based on the idea that in multi-microphone scenarios the degree of signal distortion differs among the channels, depending on the microphone position and characteristics. Even if speech enhancement is applied, the processed speech signals will not be distorted equally, so some of them may be decoded with less recognition errors than others. Consequently, the ASR system may benefit if signals of higher quality are selected for further processing. To do so, a measure of distortion, or a measure of how well recorded or enhanced signals fit the set of acoustic models of the ASR system is needed.

In this thesis a detailed overview of the CS measures reported in the literature was made. They were categorized as signal- and decoder-based. The decoder-based measures work in close cooperation with the decoder, so they should better reflect the decoder's preference than the signal-based ones. The advantage of the signal-based measures is that the channel can be selected before the signal enters the classification part of the ASR system, so recognition is made only once. Contrary, in the extraction process of the decoder-based measures at least one classification is performed for each channel, so the CS methods using these measures are computationally more demanding.

Almost all existing CS measures were compared experimentally, what has not been done so far. Several novel techniques from both categories were proposed and tested.

In particular, the signal-based measures using the relative position and orientation of the speaker and microphone [84], measures extracted from the RIR [85], or from the speech sub-band envelopes (the EV measure) [84, 106]; and two decoder-based measures, one using the acoustic likelihood normalized across channels [89], and the other extracted from the N-best lists [91].

The EV measure in many experiments matched or outperformed existing methods both, in terms of recognition performance of the ASR system and computational complexity. The decoder-based technique using the N-best lists is computationally more expensive, but it provided the best results in average from all the techniques. Very good recognition results were achieved for both methods even if the measure was extracted only from a short utterance, so they may be applied in scenarios where low recognition delay is required. Alternative decoder-based approaches usually require longer speech segments to reach a similar performance.

We also proposed and tested the combination of various CS measures in two ways: parallel and serial combination. If different measures are combined in parallel, further ASR performance improvement may be achieved, surpassing the performance of the individual measures. The serial combination may be used to reduce the computational load. The CS methods using the decoder-based measures perform well, but the computational load may be very high if the measure is extracted for a large number of channels. Therefore, we proposed to use a less complex signal-based measure first in the front-end to select only few channels, and in the next step make a more precise selection from the reduced channel set using the decoder-based methods or their combinations. When we used a computationally cheap EV method in the pre-selection and combined methods in the second step, we showed that the computational load could be reduced almost by half with only a slight loss in the recognition performance.

A real-time CS client using the EV measure was implemented in the UPC smart room. When tested with the real distant-talking microphone recordings and moving speakers, a significant recognition improvement was achieved for different ASR system configurations. Also, it was observed in all experiments that if CS was applied jointly with other robust ASR techniques, like matched training or mean and variance normalization, the recognition improvements from both approaches were cumulative up to some extend.

The second topic reported in this thesis was developed in the REMOS framework, which is usually applied in single-microphone scenarios. A new feature extraction method based on FF was designed, and published in [101], to replace the logarithmic mel-spectral features in the former REMOS implementation. The elements of logarithmic mel-spectral feature vector are highly correlated, what violates the assumption made in the HMM-based systems, and results in a recognition performance loss. The use of standard ASR features, like MFCC is restricted due to the complexity of the inner optimization problem inside REMOS. The new feature extraction method provided consistent recognition error reduction for all tested conditions.

6.2 Future work

In this thesis, the CS measures were applied in the following way. We assumed that an utterance was acquired by several microphones. Then, different CS measures were extracted for each channel, and only one signal, that presumably leads to the lowest WER, was selected for recognition. The concept however can be exploited further by working separately for each sub-band. In fact, it may happen that particular frequency bands are less distorted for a given microphone than for the others. Therefore, it would make sense to select and combine signals coming from the least distorted frequency bands. There are several possibilities to combine them. It may be done at the signal level, the feature level, or at the level of HMM emission probabilities in a multi-stream way. Hence, not only the spatial, but also the spectral diversity would be exploited.

When selecting the best microphone using the signal-based CS methods it was assumed that the least distorted signal is the best one for recognition. This assumption is reasonable if the acoustic models are trained with the clean speech. But, if models are trained with noisy data, the use of the least distorted signal may not be optimal, and a better procedure may be envisaged. To find a more suitable CS criterion, tailored to a particular ASR system, a small amount of transcribed training data could be used prior to the system deployment. Then, ideas from the feature adaptation methods could be borrowed. For instance, feature histograms in each channel could be computed, and, assuming known transcriptions of the training data, the system could learn the histogram shape corresponding to the lowest average WER. In testing, the channel with the histogram shape most similar to that one would be selected for recognition. This is just an example of what in general could be called adaptive signal-based CS methods. The main advantage of the signal-based methods over the decoder-based ones, i.e its lower computational cost, would remain, since a decoding for each stream is not necessary, but at the same time, thanks to the training phase, the CS criterion would be closely coupled to the recognition system.

In previous works it was shown that microphone array processing can benefit from

CS. The CS methods that were proposed here, however, have not been evaluated when working jointly with microphone arrays yet. Furthermore, there are many single- and multi-microphone noise and reverberation reduction methods that could be be used together with CS and benefit from the combination in two ways. Firstly, the best of the available signals in the multi-microphone environment would be used by those methods, what may help to achieve better performance. Secondly, the computational load would be reduced, if we apply those methods, which can be computationally expensive, to a reduced channel set.

Additionally, relatively high quality microphones were used in the experiments in this work. However, CS could be applied in even more difficult scenarios, where the speech signals are recorded by microphones with different characteristics, like those in laptops and smart-phones of the meeting participants. Nowadays, there is often more than one microphone per person available in a meeting, considering almost everybody owns at least one communication device. Usually those microphones are not used for the ASR in the room, and CS could be one possibility to take advantage of their presence and integrate them into the ASR processing chain; for instance, simply by selecting the microphone that is the closest one to the active speaker. It would be interesting to study the potential of such approach. The author of this thesis is not aware of any speech database that could be used for that purpose, so a data collection would be required first.

Speech is not the only audio signal that has to be recognized. Acoustic event classification has recently received an increased interest of the research community. Although classification of signals like door slam, steps, or laughter, in general, is less susceptible to effects of noise and reverberation than speech, it could still benefit from CS, for example in scenarios where the acoustic events occur in greater scales (e.g. in corridors, or halls) or in different rooms, so more microphones are required to capture the signal.

In summary, CS is a paradigm applicable to many multi-microphone scenarios. Only a specific case (ASR in a meeting room with distant microphones) was investigated in this thesis, but it was demonstrated by different experiments that CS may lead to significant improvements of ASR performance often with very low computational effort. Furthermore, those improvements proved to be cumulative when combined with other robust ASR methods. In the future, CS could be applied to even more difficult scenarios and used not only with ASR, but also in classification of other audio signals. Also, the possibility to rank and combine different sub-band signals in order to take advantage of space-frequency dependent distortion, is appealing.

Appendix A

The UPC Smart Room Impulse Response Database

A.1 The UPC Smart Room description

Presented multi-microphone RIR database was recorded in the UPC smart room [107]. This room was developed by the UPC Speech Processing Group, which has been involved in several projects with focus on applications related to the meetings and seminars in the smart rooms, such as CHIL project founded by European Union, or ACESCA, SAPIRE and SARAI projects founded by the Spanish Government.

The UPC smart room is a multi-modal room with multiple audio and video sensors. This database contains RIRs recorded by 24 microphones, which are installed in 6 T-shaped clusters and placed on the walls as illustrated in Figure A.1. Exact positions of the microphones in meters are listed in Table A.1. The order of coordinates is indicated in the left-up corner of Figure A.1.

A.2 Measurement of RIRs

A RIR describes the wave propagation between the source and microphone. To estimate it, an excitation signal is emitted from the desired position and recorded by the microphone. In theory, an ideal signal for this purpose is the Dirac impulse, but in practice it is technically difficult to generate it. Hence, noises and signal sweeps are usually used instead. In this work we used the so-called "logarithmic sweep", which is a sweep signal with instantaneous frequency varying exponentially with time. The spectrogram of this signal is shown in Figure A.2.

The original sweep signal $s(t)$ is recorded in the distant talking microphone dis-

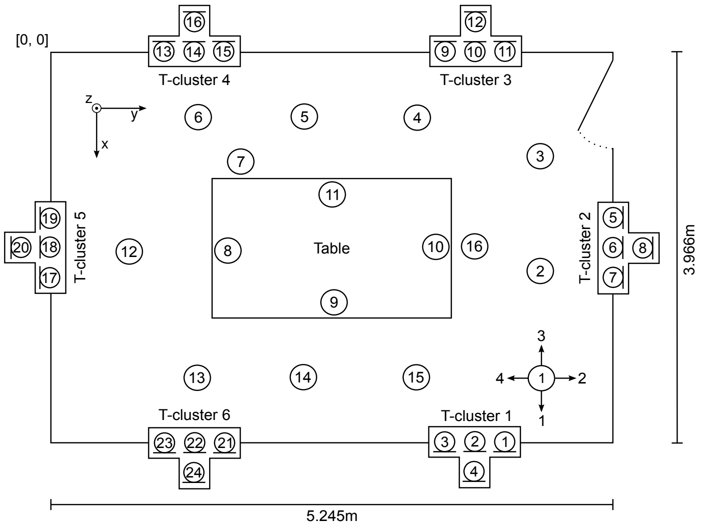


Figure A.1: The UPC smart room with 24 microphones grouped in 6 T-shape clusters. Approximate positions of RIR measurements and assumed orientations of the speaker are shown. The drawing is not to scale.

torted by reverberation and additive noise. This may be may be described as

$$r(t) = s(t) * h(t) + n(t), \tag{A.1}$$

where $h(t)$ is the RIR we want to estimate and $n(t)$ is the additive noise present in the room. Transforming Equation (A.1) into the frequency domain, we may write

$$R(t) = S(f) \cdot H(f) + N(f). \tag{A.2}$$

Neglecting the additive noise, the room transfer function can be extracted as

$$h(t) = \Im^{-1}\{H(f)\} = \Im^{-1}\{\frac{R(f)}{S(f)}\}, \tag{A.3}$$

where $\Im^{-1}\{.\}$ denotes the inverse Fourier transform.

The RIRs extracted neglecting the additive noise were used in the experiments in this thesis. However, the recording in the “silent“ room, which is also part of the database, may be used to remove the additive noise from the measured RIR if needed.

The measurement equipment consisted of a loudspeaker and a laptop with a sound card with sound IO. The excitation signal was emitted from the loudspeaker held by a person on 16 different positions are 4 orientations, and recorded by 24 microphones as shown in Figure A.1. Coordinates of the positions are listed in Table A.2. There are 1536 RIRs available in the database. The sampling frequency was 44100 kHz and the signal was recorded with the precision 16 bits.

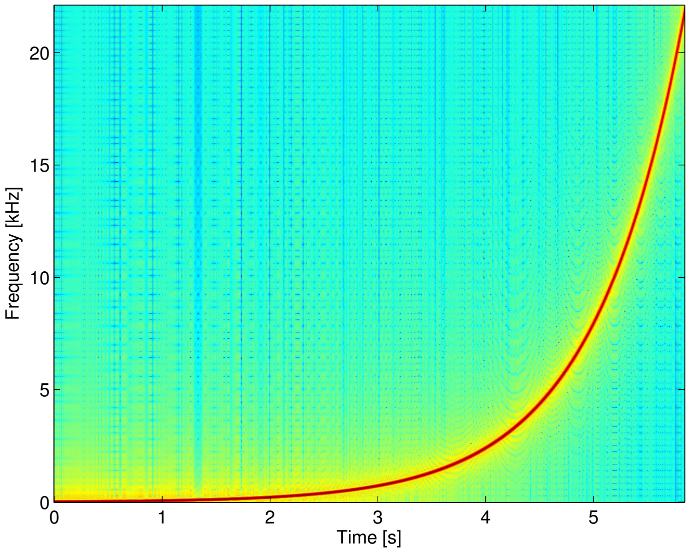


Figure A.2: Spectrogram of the excitation sweep signal with sampling frequency 44100 kHz, frame length 25 ms and frame shift 10 ms.

The measuring software and microphones were not perfectly synchronized, so the exact moment when the excitation signal was emitted is not known. Therefore, the initial delay before the arrival of the direct wave multiplied by the sound velocity can not be used to estimate the distance between the speaker and microphone.

Note also, that similar procedure using the "logarithmic sweep" signal to estimate RIRs was applied in the same room also in other works [31, 108].

A.3 Statistics

To show an overall statistics of the collected RIRs, the reverberation time T_{60} and DRR have been extracted from the database. The distributions of these parameters are shown in Figures A.3 and A.4 respectively.

The reverberation time T_{60} was estimated using the Schroeder integral [109] computed from a full-band RIR as

$$L(t) = 10 \log \int_t^\infty h^2(\tau) d\tau. \tag{A.4}$$

The DRR was computed as a ratio between the energy of direct wave and early reflections and the energy of the reverberation tail as

$$DRR = 10 \log \frac{\sum_{t=0}^{50ms} h^2(t)}{\sum_{t=50ms}^{T} h^2(t)}, \tag{A.5}$$

where $h(t)$ is a RIR of the duration T.

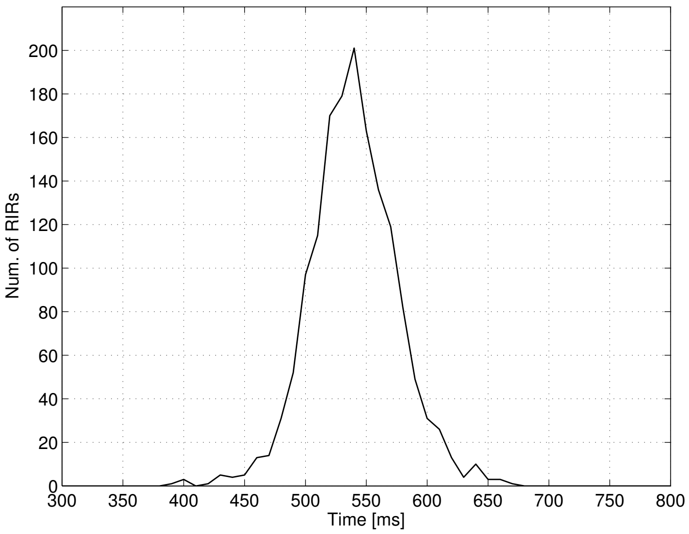


Figure A.3: Histogram of the full-band reverberation time T_{60}.

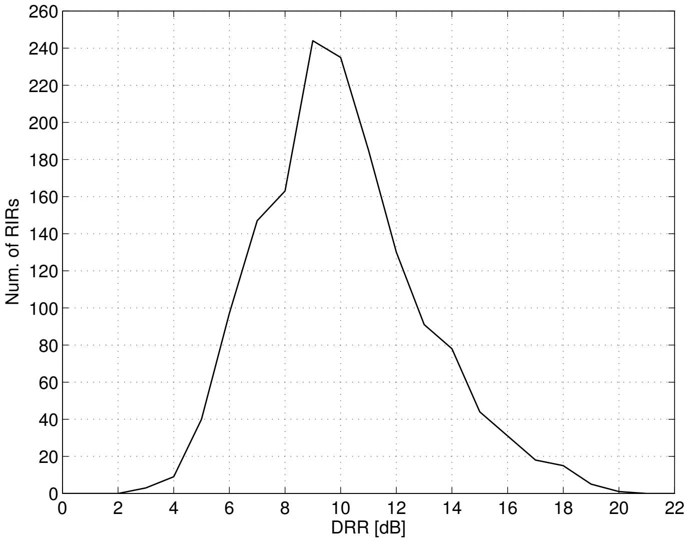


Figure A.4: Histogram of the DRR.

Table A.1: Coordinates of the microphones in meters.

Microphone	x	y	z
1	3.949	4.115	2.382
2	3.949	3.915	2.382
3	3.949	3.715	2.382
4	3.949	3.915	2.682
5	1.812	5.228	2.387
6	2.012	5.228	2.387
7	2.212	5.228	2.387
8	2.012	5.228	2.687
9	0.017	3.730	2.387
10	0.017	3.930	2.387
11	0.017	4.130	2.387
12	0.017	3.930	2.687
13	0.022	1.875	2.387
14	0.022	2.075	2.387
15	0.026	2.275	2.387
16	0.022	2.075	2.687
17	1.811	0.390	2.380
18	2.011	0.390	2.380
19	2.211	0.390	2.380
20	2.011	0.390	2.680
21	3.945	2.260	2.382
22	3.945	2.060	2.382
23	3.945	1.860	2.382
24	3.945	2.060	2.682

Table A.2: Coordinates of the measuring points in meters.

Position	x	y	z
1	3.5	4.5	1.4
2	2.4	4.5	1.4
3	1.2	4.5	1.4
4	0.5	3.6	1.4
5	0.5	2.6	1.4
6	0.5	1.6	1.4
7	1.2	1.9	1.4
8	2	1.7	1.1
9	2.6	2.7	1.1
10	2	3.7	1.1
11	1.4	2.7	1.1
12	2	1	1.4
13	3.5	1.6	1.4
14	3.5	2.6	1.4
15	3.5	3.6	1.4
16	2	3.8	1.4

Appendix B

Implementation of the Real-Time CS Client in the UPC Smart Room

Algorithms are usually first tested in controlled conditions. If the performance is satisfactory, the next (often challenging) step is to integrate those ideas into the practical applications. Here, we describe implementation of the real-time CS client in the UPC smart room. Our objective was to design it in a way, that it could be easily combined with existing ASR systems. The off-line experiments showed that the EV-based CS measure, developed in this thesis and described in Section 3.2.3, performs very well in reverberant environments, even if the amount of data to estimate the measure is small. Furthermore, since the CS method using this measure operates in the front-end and before the feature extraction, it can be easily combined with any ASR system. For these reasons it was selected for the real-time implementation. Nevertheless, there are few other practical aspects to be considered.

A typical on-line ASR system assumes a single continuous stream of speech samples. The ASR system-independent CS client should therefore be able to process several channels on the input, and output one uninterrupted stream that may be passed to the feature extraction part of the recognizer, as if only one microphone was recording the speech signal. In the off-line CS experiments we were working on the utterance level, so a microphone was selected for the whole file. In the on-line version, however, the signal stream is continuous and the utterance boundaries are not marked. The channel should not change if the speech is present, because this would introduce distortions into the signal due to different delays and power levels in each channel. To avoid this, the silence (or speech) detector may be used to indicate the non-speech intervals where the change is possible without harming the speech signal quality.

Also, the processing time of the whole recognition chain is an important factor to consider in real-time systems. If we simply used the off-line strategy, we would first have to buffer the signals from all channels between two silences, analyze them, select the best one, and then pass it to the output. The ASR system needs another time to extract the features and decode the utterance. The resulting accumulated delay may be unacceptable in case of long utterances. To avoid this problem we pass the speech samples from the selected channel immediately upon receiving without any delay. At the same time, the speech samples from all channels are buffered and used to extract the CS measure. The channel is selected just when the speech segment ends and the silence appears. The current channel is therefore selected using the measures extracted from the previous speech segment. The advantage of this approach is that no delay is introduced by the CS client. The disadvantage is that one adaptation speech segment is needed, and if the speaker's position or orientation changes significantly after each silence, the decision is not optimal.

The block diagram of CS system implemented in the UPC smart room is depicted in Figure B.1. There are three main functional blocks, a simple energy-based silence detector that marks the silence/non-silence intervals, the EV block where the speech samples from all channels are buffered and the CS measures are calculated for each segment between two silences, and finally the central part, where the decision is made about which channel should be passed to the output and when is the right moment to do the so. After the start, before the first speech that may be used to extract the CS measure arrives, the system outputs the default channel. The same channel is also always used by the silence detector.

The output of the CS client may be seen in Figure B.2. Notice the different signal levels in the silent portions in the upper figure. This is because the signals were recorded by different microphones. The exact moment when the channel was changed is also clearly visible in the spectral domain. Since it happens in the silent portion after the utterance, the speech distortion is minimized and the ASR performance should not be affected.

The experimental results, testing the real-time CS client implementation in the UPC smart room, may be found in Section 4.7.

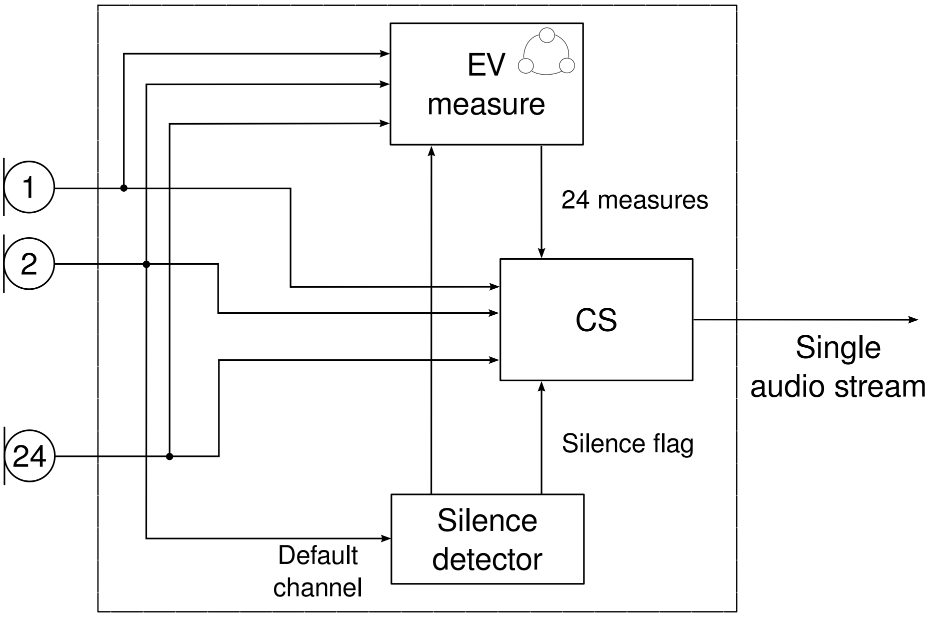


Figure B.1: Block diagram of the real-time CS client using the EV-based measure.

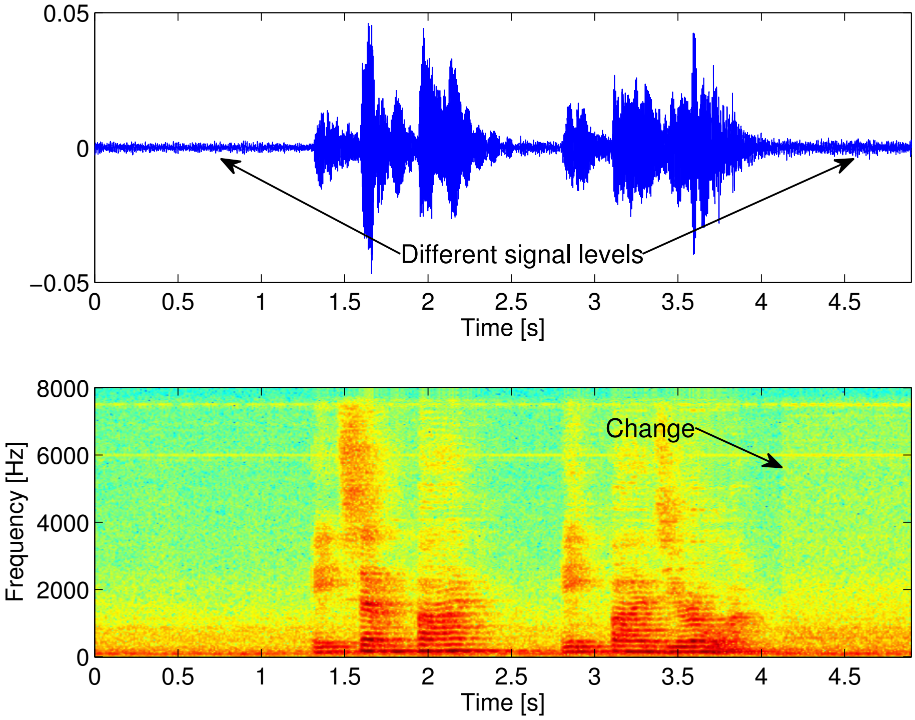


Figure B.2: Signal on the output of the real-time CS client with the channel change after the 4^{th} second. The depicted signal corresponds to the utterance "eight seven five eight five zero one".

References

[1] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] J. Benesty, S. Makino, and J. Chen, eds., *Speech Enhancement.* Springer, 1 ed., 2005.

[3] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: An overview," in *Progress in Nonlinear Speech Processing* (Y. Stylianou, M. Faundez-Zanuy, and A. Esposito, eds.), vol. 4391 of *Lecture Notes in Computer Science*, pp. 217–248, Springer Berlin / Heidelberg, 2007.

[4] P. Krishnamoorthy and S. Mahadeva Prasanna, "Temporal and spectral processing methods for processing of degraded speech: A review," *IETE Technical Review*, vol. 26, no. 2, pp. 137–148, 2009.

[5] J. Benesty, J. Chen, and E. A. Habets, "Multichannel speech enhancement with filters," in *Speech Enhancement in the STFT Domain*, SpringerBriefs in Electrical and Computer Engineering, pp. 77–92, Springer Berlin / Heidelberg, 2012.

[6] M. Brandstein and D. Ward, *Microphone Arrays.* Birkhäuser, 2001.

[7] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 127–140, 2012.

[8] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1676–1691, 2010.

[9] C. Nadeu, J. Hernando, and M. Gorricho, "On the decorrelation of filter-bank energies in speech recognition," *Proc. of EUROSPEECH*, pp. 1381–1384, 1995.

[10] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[11] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ, USA: Prentice Hall PTR, 1993.

[12] X. Huang, H.-W. Hon, and R. Reddy, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development.* Prentice Hall PTR, 2001.

[13] M. Gales and S. Young, “The application of hidden Markov models in speech recognition,” *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[14] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, P. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. Cretin, H. Bourlard, and M. Athineos, “Pushing the envelope - aside [speech recognition],” *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[15] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[16] C. Nadeu, D. Macho, and J. Hernando, “Time and frequency filtering of filter-bank energies for robust HMM speech recognition,” *Speech Communication*, vol. 34, pp. 93–114, Apr. 2001.

[17] M. Wölfel and J. McDonough, *Distant Speech Recognition.* Hoboken, NJ: Wiley, June 2009.

[18] S. Furui, “On the role of spectral transition for speech perception,” *The Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1986.

[19] L. E. Baum and J. A. Eagon, “An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology,” *Bulletin of the American Mathematical Society*, vol. 73, no. 3, pp. 360–363, 1967.

[20] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” *The Annals of Mathematical Statistics*, vol. 41, pp. 164–171, Feb. 1970.

[21] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[22] H. Kuttruff, *Acoustics.* Taylor & Francis, 2007.

[23] Y. H. Takanobu Nishiura, "Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria.," in *Proc. of INTERSPEECH*, pp. 1082–1085, 2007.

[24] T. Fukumori, M. Morise, and T. Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria RSR-dn with acoustic parameters.," in *Proc. of INTERSPEECH*, pp. 562–565, ISCA, 2010.

[25] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition," in *Proc. of INTERSPEECH*, pp. 1094–1097, 2007.

[26] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *Proc. of INTERSPEECH*, pp. 2111–2114, Sept. 2009.

[27] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," in *Proc. of INTERSPEECH*, vol. 2, pp. 1065–1068, 2002.

[28] R. Häb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," in *Proc. of ICASSP*, vol. 1, pp. 13–16, 1992.

[29] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, pp. 269–277, July 2008.

[30] E. A. Lehmann, "Fast simulation of acoustic room impulse responses." http://www.mathworks.com/matlabcentral/fileexchange/25965-fast-simulation-of-acoustic-room-impulse-responses-image-source-method. [Online; accessed 18-April-2013].

[31] R. Pertick, *Robuste Spracherkennung unter raumakustischen Umgebungsbedingungen.* PhD thesis, Technischen Universiät Dresden, Dresden, Germany, 2009.

[32] A. Sehr, *Reverberation Modeling for Robust Distant-Talking Speech Recognition.* PhD thesis, University of Erlangen-Nuremberg, Erlangen, Germany, Erlangen, Oct. 2009.

[33] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[34] A. M. Toh, R. Togneri, and S. Nordholm, "Combining MLLR adaptation and feature extraction for robust speech recognition in reverberant environments," in *Eleventh Australasian International Conference on Speech Science and Technology*, pp. 88–91, 2006.

[35] C. Segura, *Speaker Localization and Orientation in Multimodal Smart Environments.* PhD thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, Mar. 2011.

[36] I. Tashev, "Gain self-calibration procedure for microphone arrays," in *Proc. of IEEE International Conference on Multimedia and Expo*, vol. 2, pp. 983–986, 2004.

[37] I. Tashev, "Beamformer sensitivity to microphone manufacturing tolerances," *Proc. of Nineteenth International Conference Systems for Automation of Engineering and Research*, pp. 132–136, 2005.

[38] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. of ICASSP*, vol. 1, pp. 227–230, 1997.

[39] Y. Obuchi, "Noise robust speech recognition using deltacepstrum normalization and channel selection," *Electronics and Communications in Japan (Part II: Electronics)*, vol. 89, pp. 9–20, July 2006.

[40] K. Kumatani, J. Mcdonough, B. Rauch, P. N. Garner, W. Li, and J. Dines, "Maximum kurtosis beamforming with the generalized sidelobe canceller," in *Proc. of INTERSPEECH*, pp. 423–426, 2008.

[41] M. L. Seltzer, *Microphone array processing for robust speech recognition.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2003.

[42] Alberto Abad Gareta, *A multi-microphone approach to speech processing in a smart-room environment.* PhD thesis, Universitat Politècnica de Catalunya, Feb. 2007.

[43] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

[44] B. W. Gillespie and A. E. Atlas, "Strategies for improving audible quality and speech recognition accuracy of reverberant speech," in *Proc. of ICASSP*, vol. 1, pp. 676–679, IEEE, 2003.

[45] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. of ICASSP*, vol. 1, pp. I–92–5, 2003.

[46] T. Hikichi, M. Delcroix, and M. Miyoshi, "Inverse filtering for speech dereverberation less sensitive to noise and room transfer function fluctuations," *EURASIP J. Appl. Signal Process.*, vol. 2007, p. 6262, Jan. 2007.

[47] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[48] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[49] K. Kumar and R. Stern, "Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation," in *Proc. of ICASSP*, pp. 4282–4285, 2010.

[50] B. Yegnanarayana and P. Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 267 –281, May 2000.

[51] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. of ICASSP*, pp. 4604–4607, 2011.

[52] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[53] O. Ichikawa, T. Fukuda, and M. Nishimura, “Dynamic features in the Linear-Logarithmic hybrid domain for automatic speech recognition in a reverberant environment,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, pp. 816 –823, Oct. 2010.

[54] T. Houtgast and H. J. M. Steeneken, “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria,” *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.

[55] R. Drullman, J. M. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,” *The Journal of the Acoustical Society of America*, vol. 95, pp. 1053–1064, Feb. 1994.

[56] B. E. Kingsbury, N. Morgan, and S. Greenberg, “Robust speech recognition using the modulation spectrogram,” *Speech Communication*, vol. 25, pp. 117–132, Aug. 1998.

[57] R. Petrick, K. Lohde, M. Lorenz, and R. Hoffmann, “A new feature analysis method for robust ASR in reverberant environments based on the harmonic structure of speech,” in *Proc. of EUSIPCO*, 2008.

[58] T. Takiguchi and Y. Ariki, “Robust feature extraction using kernel PCA,” in *Proc. of ICASSP*, vol. 1, pp. 509–512, May 2006.

[59] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, “Training of HMM with filtered speech material for hands-free recognition,” in *Proc. of ICASSP*, vol. 1, pp. 449–452, 1999.

[60] V. Stahl, A. Fischer, and R. Bippus, “Acoustic synthesis of training data for speech recognition in living room environments,” in *Proc. of ICASSP*, vol. 1, pp. 21–24, 2001.

[61] L. Couvreur, C. Couvreur, and C. Ris, “A corpus-based approach for robust ASR in reverberant environments,” *Proc. of ICSLP*, vol. 1, pp. 397–400, 2000.

[62] L. Couvreur and C. Couvreur, “Blind model selection for automatic speech recognition in reverberant environments,” *Journal of Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 36, pp. 189–203, Feb. 2004.

[63] C. J. Leggetter and P. C. Woodland, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” *Computer Speech & Language*, vol. 9, pp. 171–185, 1995.

[64] J.-l. Gauvain and C.-h. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[65] M. Gales and S. Young, “Robust speech recognition in additive and convolutional noise using parallel model combination,” *Computer Speech & Language*, vol. 9, pp. 289–307, Oct. 1995.

[66] M. J. F. Gales and S. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, 1996.

[67] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, “HMM adaptation using vector taylor series for noisy speech recognition,” in *Proc. of International Conference on Spoken Language Processing*, vol. 3, pp. 869–872, 2000.

[68] H.-G. Hirsch and H. Finster, “A new approach for the adaptation of HMMs to reverberation and background noise,” *Speech Communication*, vol. 50, pp. 244–263, Mar. 2008.

[69] A. Sehr, M. Gardill, and W. Kellermann, “Adapting HMMs of distant-talking asr systems using feature-domain reverberation models,” in *Proc. of EUSIPCO*, pp. 540–543, 2009.

[70] C. Raut, T. Nishimoto, and S. Sagayama, “Model adaptation for long convolutional distortion by maximum likelihood based state filtering approach,” in *Proc. of ICASSP*, vol. 1, pp. 1133–1136, 2006.

[71] M. J. F. Gales and Y. Q. Wang, “Model-based approaches to handling additive noise in reverberant environments,” in *Proc. of HSCMA*, pp. 121–126, 2011.

[72] Y. Q. Wang and M. J. F. Gales, “Improving reverberant VTS for hands-free robust speech recognition,” in *Proc. of ASRU*, pp. 113–118, 2011.

[73] Y. Obuchi, “Multiple-microphone robust speech recognition using decoder-based channel selection,” in *Workshop on Statistical and Perceptual Audio Processing*, (Jeju, Korea), 2004.

[74] M. Wölfel, C. Fügen, S. Ikbal, and J. W. Mcdonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proc. of INTERSPEECH*, 2006.

[75] K. Kumatani, J. McDonough, J. Lehman, and B. Raj, "Channel selection based on multichannel cross-correlation coefficients for distant speech recognition," in *Proc. of HSCMA*, (Edinburgh, United Kingdom), pp. 1–6, June 2011.

[76] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing.* Springer, 2008.

[77] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura, "Speech recognition based on space diversity using distributed multi-microphone," in *Proc. of ICASSP*, vol. 3, pp. 1747–1750, 2000.

[78] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[79] J. Openshaw and J. Masan, "On the limitations of cepstral features in noise," in *, 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94*, vol. ii, pp. II/49 –II/52 vol.2, Apr. 1994.

[80] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic feature space," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 21–24, 2001.

[81] A. de la Torre, J. C. Segura, C. Benítez, A. M. Peinado, and A. J. Rubio, "Nonlinear transformations of the feature space for robust speech recognition," in *Proceedings of ICASSP 2002*, p. 401404, 2002.

[82] M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones," in *Proc. of INTERSPEECH*, pp. 582–585, 2007.

[83] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, p. 179188, 1936.

[84] M. Wolf and C. Nadeu, "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. of INTERSPEECH*, pp. 80–83, 2010.

[85] M. Wolf and C. Nadeu, “Towards microphone selection based on room impulse response energy-related measures,” in *Proc. of I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, (Porto Salvo, Portugal,), pp. 61–64, 2009.

[86] M. Jeub, C. Nelke, C. Beaugeant, and P. Vary, “Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals,” *Proc. of EUSIPCO*, 2011.

[87] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[88] J. Hui, “Confidence measures for speech recognition: A survey,” *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[89] M. Wolf and C. Nadeu, “Pairwise likelihood normalization-based channel selection for multi-microphone ASR,” in *Proc. of IberSPEECH*, (Madrid, Spain), pp. 513–522, Nov. 2012.

[90] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, “Confidence measures for large vocabulary continuous speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 288–298, 2001.

[91] M. Wolf and C. Nadeu, “Channel selection using N-best hypothesis for multi-microphone ASR,” in *Proc. of INTERSPEECH*, pp. 3507–3511, 2013.

[92] R. O. Duda, P. E. Hart, D. G. Stork, and D. G, *Pattern Classification.* Wiley, 2001.

[93] M. Weintraub, “LVCSR log-likelihood ratio scoring for keyword spotting,” in *Proc. ICASSP*, pp. 129–132, 1995.

[94] A. Stolcke, Y. König, and M. Weintraub, “Explicit word error minimization in n-best list rescoring,” *Proc. of EUROSPEECH*, vol. 1, pp. 163–166, 1997.

[95] R. Leonard, “A database for speaker-independent digit recognition,” in *Proc. of ICASSP*, vol. 3, pp. 111–114, 1984.

[96] ICSI, “ICSI Meeting Recorder Digits corpus.” `http://www1.icsi.berkeley.edu/Speech/mr/mrdigits.html`, 2003. [Online; accessed 24-January-2013].

[97] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macías-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proc. of ICASSP 2004 Meeting Recognition Workshop*, Prentice Hall, 2004.

[98] S. Young and et. al., *The HTK Book (for HTK Version 3.4).* Cambridge University Press, 2006.

[99] "SPeech quality assurance (SPQA) package."

[100] C. Segura and J. Hernando, "GCC-PHAT based head orientation estimation," in *Proc. of INTERSPEECH*, pp. 1—4, 2012.

[101] R. Maas, M. Wolf, A. Sehr, C. Nadeu, and W. Kellermann, "Extension of the REMOS concept to frequency-filtering-based features for reverberation-robust speech recognition," in *Proc. of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, (Edinburgh, UK), pp. 13–18, IEEE, May 2011.

[102] T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann, "Using artificially reverberated training data in distant-talking ASR," in *Text, Speech and Dialogue*, no. 3658 in Lecture Notes in Computer Science, pp. 226–233, Springer Berlin Heidelberg, Jan. 2005.

[103] P. Pujol, J. Padrell, C. Nadeu, and D. Macho, "Speech recognition experiments with the SPEECON database using several robust front-ends," in *Proc. of INTERSPEECH*, pp. 2105–2108, 2004.

[104] R. Maas, A. Sehr, M. Gugat, and W. Kellermann, "A highly efficient optimization scheme for REMOS-Based distant-talking speech recognition," in *Proc. of EUSIPCO*, pp. 1983–1987, 2010.

[105] R. Maas, "Evaluierung numerischer Optimierungsverfahren für die robuste Spracherkennung nach dem REMOS-Konzept," *M.Sc. thesis, University of Erlangen-Nuremberg, Erlangen, Germany*, 2009.

[106] M. Wolf and C. Nadeu, "Channel selection for multi-microphone speech recognition," *accepted in Speech Communication*, publication pending, October 2013.

[107] J. Neumann, J. R. Casas, D. Macho, and J. R. Hidalgo, "Integration of audiovisual sensors and technologies in a smart room," *Personal Ubiquitous Computing*, vol. 13, no. 1, pp. 15–23, 2007.

[108] R. Petrick, C. Rueckert, and R. Hoffmann, "Room acoustic conditions and limits in home and office environments," in *Proc. of SPECOM*, (St. Petersburg), 2009.

[109] M. R. Schroeder, "New method of measuring reverberation time," *The Journal of the Acoustical Society of America*, vol. 37, no. 3, 1965.