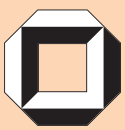
Matthias Wölfel

Robust Automatic Transcription of Lectures



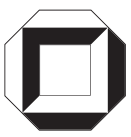
universitätsverlag karlsruhe

Matthias Wölfel

Robust Automatic Transcription of Lectures

Robust Automatic Transcription of Lectures

by
Matthias Wölfel



universitätsverlag karlsruhe

Dissertation, Universität Karlsruhe (TH)
Fakultät für Informatik
Tag der mündlichen Prüfung: 02.02.2009
Referenten: Prof. Dr. A. Waibel, Prof. Dr. S. Nakamura

Impressum

Universitätsverlag Karlsruhe
c/o Universitätsbibliothek
Straße am Forum 2
D-76131 Karlsruhe
www.uvka.de

Universitätsverlag Karlsruhe 2009
Print on Demand

ISBN: 978-3-86644-394-5

Summary

Automatic transcription of lectures and oral presentations is becoming an important task. Possible applications can be found in the fields of automatic translation, automatic summarization, information retrieval, digital libraries, education and communication research. Ideally those systems would operate on distant recordings, freeing the presenter from wearing body-mounted microphones. However, this task is surpassingly difficult, given that the speech signal is severely degraded—due to the lager distance between the mouth of the speaker and the microphone—by both, background noise and reverberation. Furthermore, the automatic transcription of lectures is challenging on other aspects: for example lecture speech varies in speaking style from freely presented to read, comprising spontaneous events as well as hyper articulation, and has a high pitch variation in comparison to private conversation.

The main goal of this thesis is to investigate, invent and present methods to improve—in comparison to state-of-the-art—automatic transcription of lectures and presentations in real environments. These improvements are established at different processing steps by various refinements and the introduction of novel techniques which will be briefly discussed next:

Feature Extraction: A critical component in feature extraction is the estimate of the speech spectrum. We have developed a spectral envelope, dubbed warped-twice minimum variance distortionless response, which is robust to noise and which enables adaptation by moving spectral resolution to lower or higher frequency regions. A change in the overall resolution is common to all spectral envelope techniques by the model order.

Model Driven Feature Adaptation: To improve the robustness of spectral envelope estimation to fundamental frequency changes we have proposed to vary the model order in dependence on the acoustic model of the speech recognition system for each individual speaker (we assume that the fundamental frequency is similar over time for the same speaker).

Signal Driven Feature Adaptation: To improve phoneme classification, it is important to emphasize the relevant characteristics while dropping the irrelevant characteristics. Traditionally all phonemes are treated equally which contradicts the observation that the important regions on the frequency axis vary for different phoneme types. Thus, to improve phoneme discrimination we have proposed to steer spectral resolution to lower or higher frequency regions according to the input signal.

Feature Enhancement: To estimate clean speech features, where a single input signal is contaminated, particle filters, a.k.a. sequential Monte Carlo methods, have been recently introduced. Unfortunately, in the "working" domain, a non-linear relationship between the noisy signal, the noise estimate and the clean signal estimate exists, which has been approximated by a vector Taylor series. We have noted that Monte Carlo already works with point observations (represented by particles) instead of distributions which allows to drop the vector Taylor series approximation. We have demonstrated that this is leading to better results while using less computational effort.

Another critical aspect in particle filter design for speech processing is the particle weight calculation which is traditionally based on a general, time independent speech model approximated by a Gaussian mixture model. We have replaced this general speech model by phoneme-specific models. The phoneme alignment is obtained by first pass text hypothesis of the speech recognition system. The proposed method, therefore, establishes a coupling between the two processing stages, enhancement and recognition, which have been treated as independent components in the past.

While previous particle filter methods, to predict the estimate of the next state, have relied either on random walk or on a predicted walk using a prior knowledge, we have proposed an integrated approach to estimate the predicted walk model within the particle filter.

A significant drawback of particle filter based enhancement methods is their limited capacity to compensate only for additive distortions. To overcome this drawback we have proposed a generalized particle filter framework which is capable to jointly track additive noise and reverberation on a frame-by-frame basis by extending the filter with an auxiliary model of late reflection.

Multi-Source Processing: In those cases, where microphone array or blind source separation techniques might not lead to improvements over "the best" single channel, selecting the channel which is leading to the lowest word error rate is an important task. We have suggested a novel channel selection method. Its advantages, compared to other selection methods, are that the evaluation of channel quality takes place on the actual features of the recognition system and that it overcomes the need for silence regions.

Combining the proposed robust feature extraction front-end with the proposed feature enhancement technique, which jointly compensates for additive and convolutive distortions can lead to further improvements. On realistic recordings in noisy and reverberant environments we have been able to demonstrate relative reductions in WER by up to 26.0% compared to the mel-frequency cepstral coefficient front-end without feature enhancement after unsupervised acoustic model adaptation.

Even though the focus of the presented work has been on lecture type of speech, the presented improvements carry over to other conditions such as speech transmitted over a telephone channel, in a meeting scenario or in human robot interaction.

Zusammenfassung

Die automatische Transkription von Vorträgen, Vorlesungen und Präsentationen wird immer wichtiger und ermöglicht erst die Anwendungen der automatischen Übersetzung von Sprache, der automatischen Zusammenfassung von Sprache, der gezielten Informationssuche in Audiodaten und somit die leichtere Zugänglichkeit in digitalen Bibliotheken. Im Idealfall arbeitet ein solches System mit einem Mikrofon das den Vortragenden vom Tragen eines Mikrofons befreit. Dies ist jedoch unvergleichlich schwer, da das Sprachsignal, durch die größere Entfernung zwischen Sprecher und Mikrofon, stärker durch Hall und Hintergrundgeräusche gestört ist. Daher müssen neue Verfahren entwickelt werden um die zusätzlichen Störungen im Signal zu kompensieren. Erschwerend kommt hinzu, dass die automatische Transkription von Vorträgen weitere zusätzliche Anforderungen an den Spracherkenner stellt: so ist z.B. sowohl die Varianz des Sprachsignals und der Sprachgeschwindigkeit als auch die Varianz der Fundamentalfrequenz im Vergleich zu einem privaten Gespräch wesentlich erhöht.

Das Hauptaugenmerk der hier vorliegenden Arbeit ist darauf gerichtet, die automatische Transkription von Vorträgen und Präsentationen in reeller Umgebung — im Vergleich zu „state-of-the-art" — zu analysieren und neue Methoden zu entwickeln. Dies wird durch gezielte Verfeinerungen und Weiterentwicklung von bekannten als auch Einführung von neuartigen Verfahren erreicht. Im Folgenden werden diese Verfahren kurz beschrieben:

Robuste Merkmalsextraktion: Eine kritische Komponente der Merkmalsextraktion ist die Schätzung des Sprachspektrums. Daher haben wir eine Einhüllende entwickelt, die besonderst robust gegenüber der Variation der Fundamentalfrequenz ist und die es weiterhin erlaubt, die spektrale

Auflösung in höhere oder niedrigere Frequenzregionen zu verschieben um bestimmte Adaptionsmethoden erst zu ermöglichen. Die Variation der Auflösung durch die Veränderung der Modellordnung liegt allen Einhüllenden zugrunde.

Modellbasierte Merkmalsadaption: Um weitere Robustheit gegenüber der Variation der Fundamentalfrequenz zu erreichen, haben wir vorgeschlagen, die Frequenzauflösung anhand des akustischen Modells des Spracherkennersystems für jeden Sprecher (wir nehmen an, dass sich die Fundamentalfrequenz je Sprecher nicht sehr verändert) individuell zu variieren.

Signalbasierte Merkmalsadaption: Um die Phonemklassifikation bei verrauschten Sprachsignalen zu verbessern, ist es wichtig die klassifikationsrelevanten Eigenschaften zu verstärken und die anderen Eigenschaften zu unterdrücken. In herkömmlichen Vorverarbeitungen werden alle Phoneme gleich behandelt. Dies widerspricht der Beobachtung, dass die wichtigen Regionen für verschiedene Phonemklassen an verschiedenen Stellen liegen. Daher haben wir vorgeschlagen, die spektrale Auflösung in Abhängigkeit des beobachteten Eingangssignals in höhere oder niedere Frequenzbereiche zu verschieben.

Merkmalsverbesserung: Um einkanalige, verunreinigte Eingangssignale zu säubern wurden vor kurzem Partikelfilter, auch bekannt als sequentielle Monte Carlo Methoden, eingeführt. Aufgrund von Nichtlinearitäten zwischen dem Sprach- und Störsignal im Repräsentationsraum wurde bisher eine Näherung durch eine Taylorreihenentwicklung verwendet. Wir haben angemerkt, dass Monte Carlo Methoden auf eine solche Näherung verzichten können, und gezeigt, dass dadurch bei verringertem Aufwand die Genauigkeit des Verfahrens verbessert werden kann.

Die bisher verwendeten Partikelfilteransätze verwenden entweder eine zufällige Vorhersage oder eine Vorhersage die auf a priori Wissen zurückgreift. Um eine zuverlässige Vorhersage zu ermöglichen, ohne dabei auf a priori Wissen zurückgreifen zu müssen, haben wir eine Methode entwickelt, die ein Vorhersagemodell innerhalb des Partikelfilters berechnet.

Ein weiterer kritischer Punkt ist die Propagierung der Partikel. Hierfür sind in der Literatur zwei Verfahren bekannt: Extended Kalman Filter und Lineare Prädiktion. Der Nachteil des Extended Kalman Filters ist der erhöhte rechnerische Aufwand. Der Nachteil der Linearen Prädiktion beruht auf der Notwendigkeit die Lineare Prädiktionsmatrix auf Geräuschregionen zu berechnen. Um die soeben genannten Nachteile zu überwinden, haben wir eine Methode entwickelt, die es ermöglicht die Lineare Prädiktionsmatrix direkt aus dem verrauschten Signal zu berechnen.

Ein großer Nachteil von partikelbasierten Methoden ist ihre Einschränkung nur additive Geräusche kompensieren zu können. Um diesen Nachteil zu überwinden schlagen wir eine Erweiterung des Partikelfilters vor, indem wir ein Hilfsmodell zur Berechnung von Reflexionen in den Filter integrieren. Dadurch ist es möglich sowohl additive Geräusche als auch Hall aus dem gestörten Eingangssignal herauszufiltern.

Mehrkanalaufnahmen: Bei Mehrkanalaufnahmen kann sich die Signalqualität der einzelnen Kanäle sehr stark unterscheiden. In solchen Fällen kann keine Verbesserung durch Array-Signalverarbeitung erreicht werden und eine zuverlässige Auswahl des ,,besten" Kanals, der zur niedrigsten Wortfehlerrate führt, ist wichtig. Basierend auf der Klassentrennung haben wir eine neue Methode entwickelt, die die Evaluation direkt auf den Merkmalen des Spracherkenners ausführt und auf Sprachpausen verzichten kann.

Durch Kombination der vorgeschlagenen robusten Merkmalsextraktion mit der vorgeschlagenen Merkmalsverbesserungstechnik, die sowohl additive als auch gefaltete Störungen kompensieren kann, sind weitere Verbesserungen möglich. Auf verrauschen und verhallten Aufnahmen konnten wir eine relatieve Reduzierung, im Vergleich zu Mel-Frequenz Kepstralkoeffizienten ohne Merkmalsverbesserungstechnik nach unüberwachter Modelladaption, der Wortfehlerrate von bis zu 26% erzielen.

Obwohl der Fokus der hier vorgestellten Arbeit auf der automatischen Transkription von Vorträgen liegt, lassen sich Teile der hier vorgestellten Verbesserungen auf andere Szenarien, z.B. auf Telefongespräche, Meetings oder Roboterinteraktionen, übertragen.

Acknowledgements

During my stay with interACT at Universität Karlsruhe (TH), which with Forschungszentrum Karlsruhe now jointly forms the Karlsruher Institut für Technologie (KIT), and at Carnegie Mellon University I had great opportunities to develop and foster my skills in research, teaching as well as cultural understanding through frequent exchange of international colleagues at our institute or our own visit at their institutes or while participating at conferences. First of all I would like to thank the director of interACT Prof. Dr. Alex Waibel for raising my interest in automatic speech processing, for his trust in my research and teaching skills and for financial support to make this work possible. I am also very thankful to Prof. Dr. Satoshi Nakamura from Advanced Telecommunication Research Institute International who agreed to be my second adviser, showed great interest in my work and gave me wonderful feedback concerning my research.

Several people, besides the ones I have already mentioned, have played a decisive role in my research career. I am indebted to many people for their long-lasting support and encouragement which were invaluable for the successful completion of this thesis. In the following lines some of them are gratefully acknowledged. However, I am aware of the fact that there are many more and these words cannot express the gratitude and respect I feel for all of them.

I would like to acknowledge the instructions and lecturs by Prof. Dr. Mari Ostendorf, University of Washington who was a visiting professor in our lab from 2005 until 2006. I had a wonderful time and got very interesting insights while being a guest at Tokyo Institute of Technology with Prof. Dr. Sadaoki Furui and at Hong Kong University of Science and Technology with Prof. Dr.

Pascale Fung and Prof. Dr. Dekai Wu. Dr. John McDonough, now with Universität des Saarlandes, was my Diplomathesis supervisor and a supporter of my ideas and research ever since. In addition we have published extensively including conference and journal publications and the book "Distant Speech Recognition" published by Wiley which was a tremendous amount of effort. I further thank Dr. Gerasimos Potamianos from IBM (now with the Foundation for Research and Technology — Hellas), Prof. Dr. Lori Lamel and Prof. Dr. Jean-Luc Gauvain both from Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur who were close project partners, and good tutors, over nearly the whole time I have been working at interACT.

Furthermore I like to express further greatest thanks for help and encouragement to my current and former colleagues and students (listed in alphabetic order) which I had the pleasure to work with—sometimes literally night and day:

Keni Bernardin, Ralf Biedert, Susanne Burger, Paisarn Charoenpornsawat, Michael Dambier, Maria Danninger, Matthias Denecke, Matthias Eck, Hazim Kemal Ekenel, Friedrich Faubel, Christian Fügen, Petra Gieselmann, Dominik Heger, Sanjika Hewavitharana, Silja Hildebrand, Hartwig Holzapfel, Roger Hsiao, Qin Jin, Szu-Chen Jou, Kenichi Kumatani, Thilo Köhler, Muntsin Kolss, Florian Kraft, Ian Lane, Kornel Laskowski, Rob Malkin, Florian Metze, ThuyLinh Nguyen, Kai Nickel, Jan Niehues, Mohamed Noamany, Sebastian Ochs, Matthias Paulik, Jürgen Reichert, Cedrick Rochet, Ivica Rogina, Kay Rottmann, Thomas Schaff, Tim Schlippe, Tanja Schultz, Hagen Soltau, Rainer Stiefelhagen, Sebastian Stüker, Yik-Cheung Tam, Ashish Venugopal, Stephan Vogel, Jie Yang, Ying (Joy) Zhang, Bing Zhao, and Andreas Zollmann.

Last but by no means least, I like to express my deepest gratitude to my parents Helmut and Doris Wölfel for their constant support in all circumstances. I am particularly indebted to my parents and my wife Irina Wölfel for their never-ending encouragement and ongoing support. In particular Irina had to spend endless evenings and weekends waiting for me—after finishing this thesis I hope to find more time which I can spend together with her and our baby. I would also like to thank Helmut and Irina for reading my thesis at earlier stages which included pouring through numerous typos, grammatical errors and unfinished ideas and for being my toughest critics.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

A *lecture* is an oral presentation intended to transfer information or teach people about a particular subject. Usually the lecturer stands at the front of the room and recites not only the lecture's content but might incorporate additional activities, e.g. writing on a chalk-board, making exercises and class questions or multimedia presentation.

A lecture is mainly a one-way communication that does not involve significant audience participation—although it might involve some questions or comments from the audience—while a *meeting* is mainly a two-way communication involving at least two active participants.

It has become a common trend at universities and elsewhere to record lectures. Those recordings are made available to students or the public for further access, usage and processing [170]. The number of lecture recordings available for download on the Internet is growing continuously. Platforms which provide video lectures in any language emerge, for example, *Research Channel* [78] which houses more than 2,700 talks and presentations, *Videolectures* [202] which offers more than 3,100 lectures or the *World Lecture Project* [205] which also includes various lectures provided by Universität Karlsruhe (TH). Other universities provide extensive courses online, probably the most prominent example is MIT's *Open Course Ware* (OCW) [165] which offers more than nine hundred of its courses for download in the Internet.

With the increasing amount of lecture recordings available worldwide, however, it becomes more and more important to offer extensive and comfortable *search* functionality, *summaries* of the lectures, *information retrieval*, *question and answer* procedures, as well as *translation* into other languages or the automatic *transformation* of lectures *into documents* [186]. The *Lecture Browser* [68], for example, enables to more effectively disseminate audio and video recordings of academic lecture material. Other platforms, such as *Lecturefinder* [134] help finding adequate lectures, however, they have to rely on text material additional to the audio or video stream; e.g. *Open Yale Courses* (OYC) [171] provide not only the audio and the video but also hand driven transcriptions.

In contrast to offline applications, *real time* applications of automatic speech recognition offer services to the audience while attending the lecture. Such services include the online translation of lectures which has been presented by our lab already in 2005 for the language pair English–Spanish [197, 76, 96] and is currently ported for other language pairs; e.g. German–English. Another online scenario where automatic speech recognition is highly desirable is *note-taking* for deaf students. At the time being, at least in Japan [125], such transcripts are provided by student volunteers. As those volunteers are untrained they can not write fast enough and thus words or even whole sentences are missing in the transcript. In addition they struggle with technical terms as they might not be experts in the required field.

In order to provide additional functionality as discussed in the previous paragraph, and thus to make the recorded lectures more useful, with limited amount of labor work and cost[1], it becomes elementary to provide transcriptions of sufficient accuracy by automatic speech recognition. To facilitate progress in automatic transcription of lecture speech, European funded projects such as *FAME* [19], *CHIL* [81] or *TC-STAR* [196] as well as US funded projects such as *STR-Dust* [193] or Japanese funded projects such as *Spontaneous Speech Corpus and Processing Technology* [99] have included automatic transcription of lectures and presentations in their project goals.

[1] 60 minutes transcription by hand ranges between 120 and 216 US$, source http://www.productiontranscripts.com

1.1 Specific Challenges in the Automatic Transcription of Lectures

Speech data of presentations and lectures differs from other sources of speech in

- the speaking style characteristics,
- the acoustic environment,
- the number and quality of acoustic channels, and
- the vocabulary and use of language.

In a wider sense, the first three items belong to feature extraction and acoustic modeling while the last item belongs to language modeling. While, in the past, research on the automatic transcription of lectures and presentations has been limited to language modeling, a brief overview is given in Section 1.6.1, we want to turn our attention on feature extraction and acoustic modeling.

In the remained of this chapter we highlight the differences in the acoustic signal and review work which has been presented in the literature on language modeling specific to the transcription of lecture speech.

1.2 Speaking Style Characteristics in Lecture Speech

Speaking style in a lecture or seminar differs from the speaking style in other scenarios. For example even though lectures follow a particular pattern (introduction, main body, conclusions), it has a higher degree of spontaneity than broadcast news which are carefully prepared. Thus lectures are quite similar to conversational speech which has been confirmed by Glass *et al.* [106] who have compared lecture speech with conversations and found that both kinds of speech contain similar amounts of spontaneous speech effects such as word contractions and reductions, extraneous filler words, non-lexical filled pauses, partial words and false starts.

In the next sections we discuss and highlight those variations and irregularities of the acoustic speech signal which can cause tremendous difficulties for automatic speech recognition. Differences in vocabulary and language structure will be presented and discussed in Section 1.5.

1.2.1 Clarity of Speech

In professional speech recordings, such as news shows or story telling, the speaker is usually trained and well-articulated, well-intonated and follows a particular structure. Lecture speech, on the other hand, has a more conversational character. The lecturer is usually not a trained speaker and thus the speech contains lots of breaks at positions that are not related to the content.

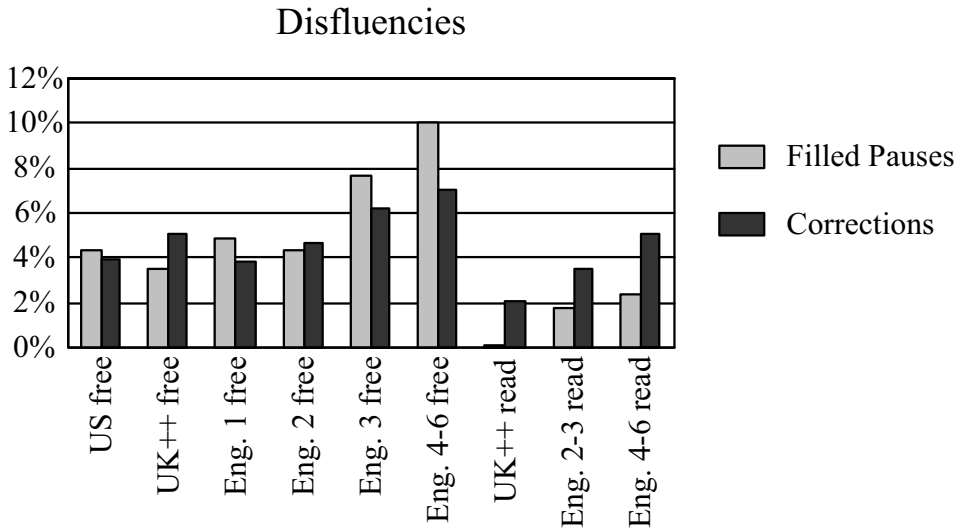


Figure 1.1: Disfluencies in percent per word count: native (US, UK++) vs. non-native (Eng.), free spoken vs. read presentations.

To investigate the percentage of *filled pauses* and *corrections*, including repetitions, corrections and false starts, in lecture speech we have investigated the TED corpus, described in Section 2.1.1, by first sorting the speakers into the categories US English speakers (*US*), UK and Australian English speakers (*UK++*) and non-native speakers (*Eng.*) labeled with the English skill on a scale from 1 to 6, where 1 was the best score. Second, we split the speech in freely spoken (*free*) or read (*read*), where read could also be prepared or memorized. The results in Figure 1.1 show a clear difference between freely spoken and read speeches: read speeches had a lower frequency of filled pauses (an average of 1.4% read vs. 5.8% free) and fewer corrections (an average of 3.5% read vs. 5.1% free). However, by looking only at the results of the 4-6 grade non-native read speeches we still see an average of 5% corrections despite the fact that these speakers were only reading prepared text. The freely produced speeches show similar rates for filled pauses and corrections for the groups *US free*, *UK++ free*, *Eng. 1 free* and *Eng. 2 free*. Speakers whose English skills were rated 4, 5 and 6 had more occurrences of filled pauses and corrections.

The occurrence of filled pauses and corrections shows an obvious correlation with the language skills of the speakers, as well as with speaking style in terms of prepared or freely spoken speech. However, two of the British native speakers seemed to stutter which increased their counts of corrections.

1.2.2 Speaking Rate

To investigate the speaking rate by counting how many words per minute a transcription contained, we have, similar as in the previous section, segmented the TED corpus by English skills and separated between freely spoken or read speech. The results in Figure 1.2 show that freely presented speeches given by speakers of US English were fastest with 168 words per minute, while speeches of non-native English speakers of level 4-6, who read their speeches were the slowest (93 words per minute). The number of words per minute decreases as the grade of English skills increases. This is consistent to Yuan *et al.* [211] who concluded that the speaking rate is dependent on the native tongue of the speaker. Taken the fact that the normal English speaking rate is between 130-200 words per minute, we observe that the average speaking rate of a native speaker given a lecture is in the same range as other speaking rates.

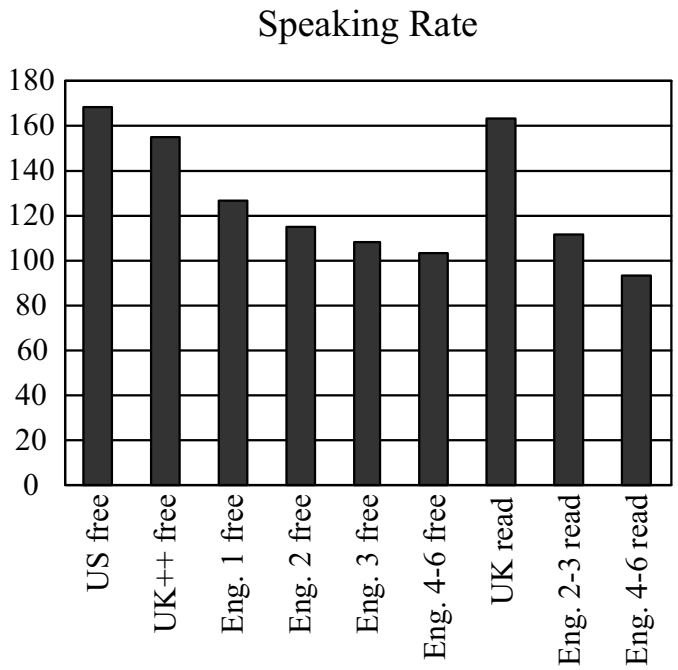


Figure 1.2: Speaking rate (words per minute): non-native vs. native, free spoken vs. read presentations.

1.2.3 Fundamental Frequency Variation

Speaking in public involves a greater variation in fundamental frequency (pitch) than speaking in private [122]. Manuals on public speaking advice speaking with a *lively voice* that varies in intonation. A lively voice is achieved by consciously modifying the prosodic dimensions of loudness, pitch and tempo. In [118] it was shown that pitch variation of the presenter's voice is indeed sensed as a lively voice.

1.3 Acoustic Environment Challenges in Lecture Speech

In a lecture scenario one will always be confronted with interfering signals or significant background noises. Thus, even if a close talk or lapel microphone is used, one will never be able to achieve a recording quality as good as in professional recordings. This might in particular happen if the close talk or lapel microphone is misplaced (this is actually quite common if untrained persons try to put on the microphone by themselves). In addition recordings of lectures have to be cheap, e.g., a laptop and an economically priced microphone might be used, and have to be made with the least human efforts as possible; e.g. automatic gain control. Furthermore the microphone might not be optimal placed. If not close enough to the mouth of the speaker severe distortions are introduced. For example consider two different microphone positions: close and in the distant to the speaker with two sound sources, the speaker and one noise source with a sound pressure level 5 dB below the sound pressure level of the speaker. As illustrated in Figure 1.3 the signals can take different paths from the sources to the microphone. The direct path (solid line) of the wanted speech signal follows a straight line starting at the mouth of the speaker. The ambient noise paths (dotted lines) follow a straight line starting at the noise source, while the reverberation paths (dashed lines) start at the wanted sound source or the ambient noise source being reflected at one object (note that the reflection is not limited to one object, but assumed here for simplicity). Furthermore, we assume a sound absorption of 5 dB at the reflecting wall.

If a sound pressure level L_1 is known at a particular distance d_1 from a point source, we can calculate the sound pressure level L_2 at another distance d_2, in the free-field, by

$$L_2 = L_1 - 20 \log \frac{d_2}{d_1} \qquad [dB] \tag{1.1}$$

With the interpretation of (1.1)—*each doubling of the distance brings down the*

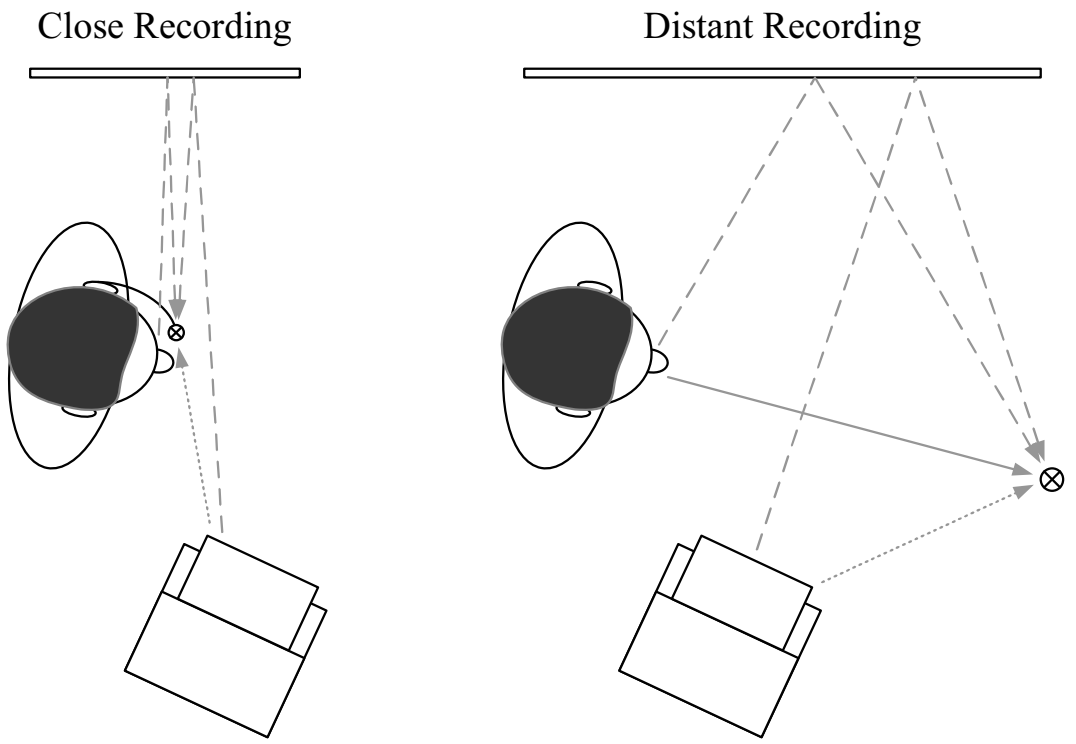


Figure 1.3: Close and distant recordings including the different paths the signals can take to the microphone.

sound pressure level by approximately 6 dB—we can plot the different sound pressure levels following the four paths of Figure 1.3. The paths start at the different distances from the sound source and sound pressure level. In addition, on the position of the reflection, we have to subtract 5 dB due to absorption. On the right hand side of the two images in Figure 1.4 we can read the differences of the direct speech signal to the distortion. From the two images it is obvious that the speech is heavily distorted on the distant microphone (2, 10 and 15 dB) while on the close microphone the distortion due to noise and reverberation is quite limited (21, 29 and 37 dB).

Hence, to automatically transcribe lectures the speech recognition system has to cope reasonably well with low-quality speech signals. The quality of the speech signals degrades by moving the microphone away from the speaker. To improve the quality of automatic transcription it might be advantageous to use additional microphones. This will be discussed in Section 1.4.

In the remained of this section we review different types of distortions.

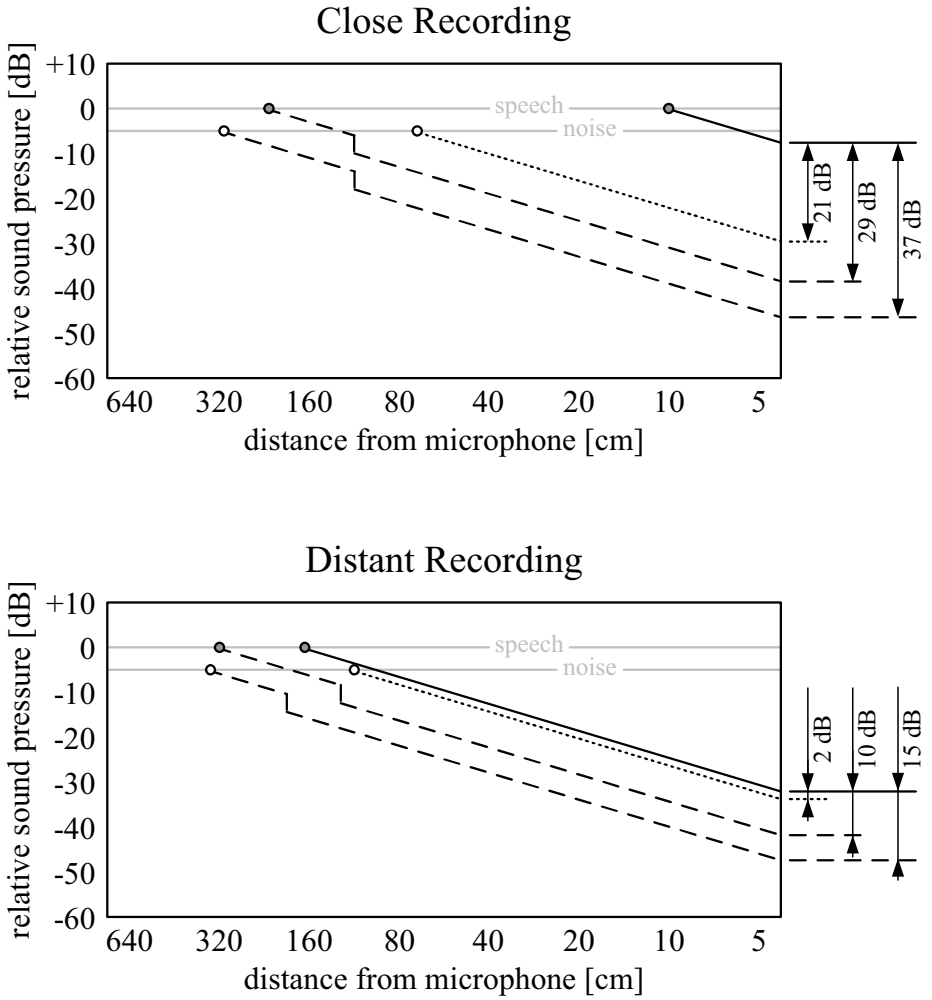


Figure 1.4: Relative sound pressure of close and distant recordings of the same sources.

1.3.1 Additive Distortions

Additive distortions, also referred to as ambient or background noise[2], is any additive sound other than the sound of interest. A broad variety of ambient noises exist, which can be classified into:

[2]We find the term background noise misleading as the "background" noise might be closer to the microphone as the "foreground" signal.

- *stationary*
  Stationary noises have statistics that do not change over long time spans. Some examples are computer fans, power transformers, and air conditioning.

- *non-stationary*
  Non-stationary noises have statistics that change significantly over relatively short periods. Some examples are interfering speakers, printers, hard drives, door slams, and music.

Note that most noises are neither entirely stationary, nor entirely non-stationary in that they can be treated as having constant statistical characteristics for the duration of the analysis window typically used for automatic speech recognition.

Additive distortions n are *uncorrelated* with the desired signal x and are mixed linear in the time domain

$$y(t) = x(t) + n(t).$$

1.3.1.1 Influence of ambient noise on speech

Let us consider a simple example. Figure 1.5 depicts the utterance of the word *cat* with an ambient noise level below 10 dB compared to the highest peak of the spoken word. Clearly the consonant /t/ is covered by the noise floor and therefore the uttered word is indistinguishable from words such as "cad", "cap", or "cab". The effect of additive noise is to fill in regions with low speech energy in the time-frequency plane.

1.3.2 Echo and Reverberation

An *echo* is a single reflection of a sound source, arriving some time after the direct sound. It can be explained as a wave that has been reflected by a discontinuity in the propagation medium, and returns with sufficient magnitude and delay to be perceived as distinct from the sound arriving on the direct path. The human ear cannot distinguish an echo from the original sound if the delay is less than 1/10 of a second [132]. This fact implies that a sound source must be more than 16.2 meters away from a reflecting wall in order for a human to perceive an audible echo. *Reverberation* occurs when, due to numerous reflections, a great many echoes arrive nearly simultaneously so that they are indistinguishable from one another. Large chambers—such as cathedrals, gymnasiums, indoor swimming pools, and large caves—are good examples of spaces having reverberation times

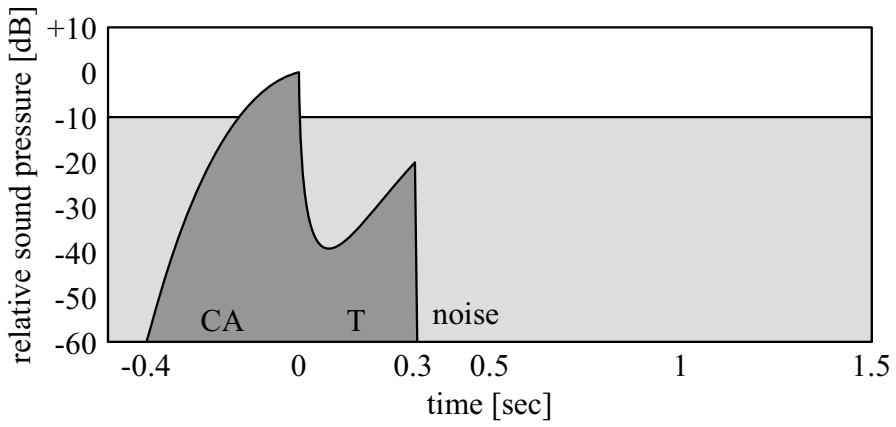


Figure 1.5: Simplified plot of relative sound pressure vs. time for an utterance of the word *cat* in additive noise.

of a second or more and wherein the reverberation is clearly audible. Those sound waves reach the ear or microphone by an infinite number of paths which can be separated into

- *direct wave*
  The direct wave is the wave which reaches the microphone on a direct path. The time delay can be calculated by the sound velocity whereas the frequency dependence can be neglected [62].
- *early reflections*
  Early reflections arrive at the microphone on an indirect path within approximately 50 to 100 ms after the direct wave and are relatively sparse. Frequency dependent attenuation is due to the reflecting surfaces.
- *late reflections*
  Late reflections are numerous reflections that follow one another so closely that they become indistinguishable and result in a diffuse field. The degradation becomes frequency dependent as the air attenuation [62] becomes more significant due to the longer sound traveling distance and frequency dependency of the reflecting surfaces.

In contrast to additive noise, the distortions introduced by echo or reverberation are *correlated* with the desired signal by the *impulse response* h of the surroundings through the convolution

$$y[k] = h[k] * x[k] = \sum_{m=0}^{M} h[k]x[k-m].$$

Note that the problem of dereverberation should not be confused with the echo cancellation problem of speaker phones where the first speaker's own voice is picked up by the microphone of the second speaker and transmitted back to the first speaker. This problem is significantly easier as the undistorted signal component which has to be suppressed in the back channel is available.

1.3.2.1 Influence of reverberation on speech

Now we consider the same simple example as before, but add reverberation with a reverberation time T_r of 1.5 s instead of ambient noise as the introduced distortion. In this case the effect is quite different, as it can be observed by comparing Figure 1.6 with Figure 1.5. It is clear that the consonant /t/ is covered again, yet however by the reverberation from the vowel /a/. Once more the word cat becomes indistinguishable from the words cad, cap or cab.

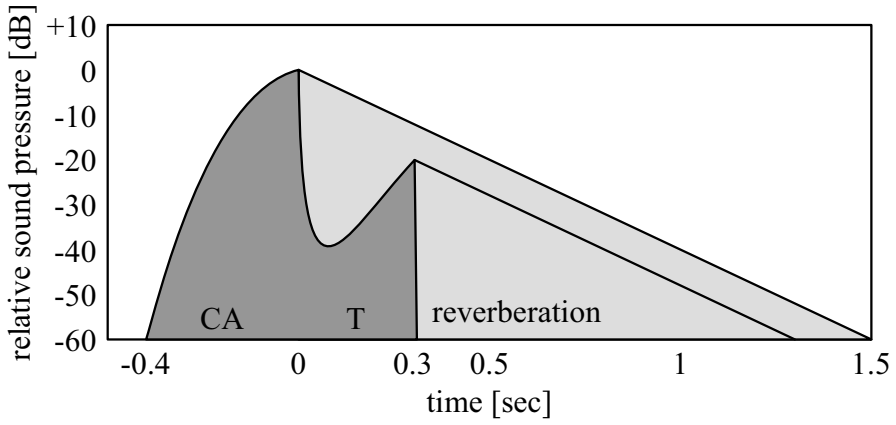


Figure 1.6: Simplified plot of relative sound pressure vs. time for an utterance of the word *cat* in a reverberant environment.

1.3.3 Room Modes

Any closed space will resonate at those frequencies where the excited waves are in phase with the reflected waves, building up a *standing wave*. The waves are in phase if the frequency of excitation between two parallel, reflective walls is such that the distance l corresponds to any integer multiplier of a half wavelength. Those frequencies at or near resonance are boosted and called *modal frequencies* or *room modes*. Therefore, the spacing of the modal frequencies—resulting in reinforcement and cancellation of acoustic energy—determines the amount of

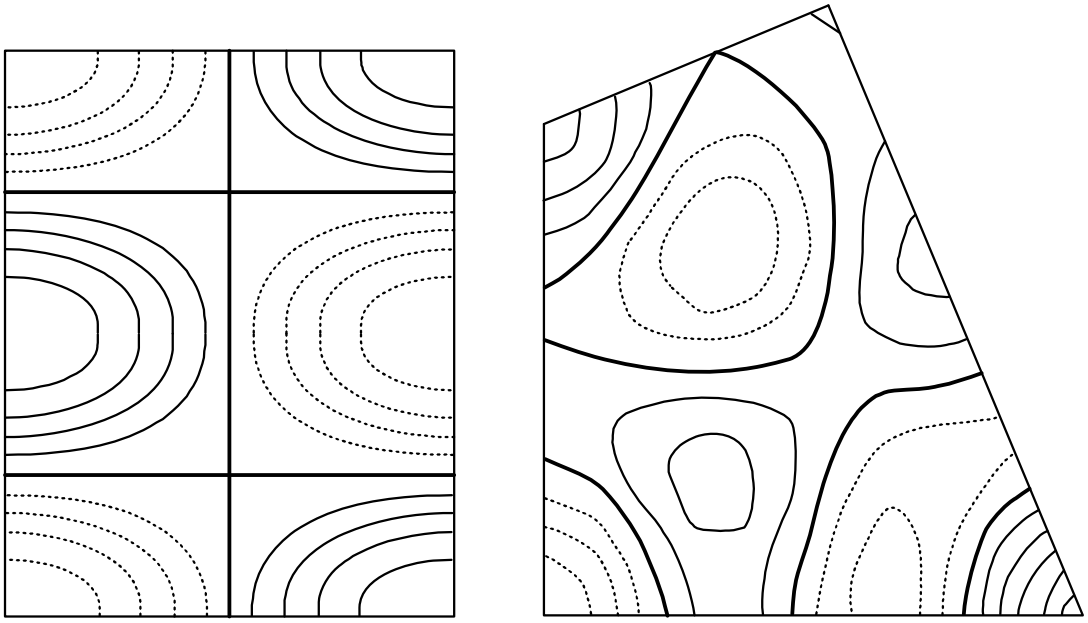


Figure 1.7: Two dimensional mode patterns of a rectangular and an irregular room shape. The bold lines indicate the knot of the modes, the thin lines positive amplitudes while the dashed lines indicate negative amplitudes.

coloration. Coloration is strongest for small rooms at bass frequencies between 20 and 200 Hz. At higher frequencies the room still has an influence, but resonances are less problematic because higher frequencies are better absorbed. The sharpness and height of the resonant peaks depend not only on the geometry of the room but also on its sound absorbing properties. A room filled with, for example, furniture, carpets and people will have high absorption and might have peaks and valleys that vary between 5 and 10 dB. A room with bare walls and floor, on the other hand, will have peaks and valleys that vary between 10 and 20 dB, sometimes even more. Note that additional coloration is introduced by the microphone transfer function.

For a rectangular room of dimensions (l_x, l_y, l_z) with its simple geometry and perfectly reflecting walls some basic conclusions can be drawn from wave theory. Figure 1.7 plots model patterns of a rectangular and an irregular room shape. The rectangular room has a very regular mode pattern while the irregular room has a complex mode pattern.

The boundary conditions require pressure extremes at all boundary surfaces, therefore we can express the sound pressure, for a rectangular room, in the form

$$p(x,y,z) = \sum_{i_x=0}^{\infty}\sum_{i_y=0}^{\infty}\sum_{i_z=0}^{\infty} p\cos\left(\frac{\pi i_x}{l_x}\right)\cos\left(\frac{\pi i_y}{l_y}\right)\cos\left(\frac{\pi i_z}{l_z}\right).$$

As stated by Rayleigh in 1869, solving the wave equation with the resonant frequency $\varrho = 2\pi i$, where i are integer values, the room modes are found to be

$$f_{\text{mode}}(x,y,z) = \frac{c}{2}\cdot\sqrt{\frac{\varrho_x^2}{l_x^2}+\frac{\varrho_y^2}{l_y^2}+\frac{\varrho_z^2}{l_z^2}}.$$

Room modes with $i = 1$ are called *first mode*, with $i = 2$ are called *second mode* and so forth. Modes with one dimension (e.g. $x \neq 0, y = 0, z = 0$) are called *axial modes*, modes with two dimensions are called *tangential modes*, and modes with three dimensions are called *oblique modes*.

An approximation of the number of resonant frequencies in a rectangular room which appear up to a certain frequency f is given by Kuttruff [131]

$$m \approx \frac{4\pi}{3}\left(\frac{f}{c}\right)^3 V + \frac{\pi}{4}\left(\frac{f}{c}\right)^2 S + \left(\frac{f}{c}\right)\frac{l}{8} \tag{1.2}$$

where V expresses the volume of the room, $S = 2(l_x l_y + l_x l_z + l_y l_z)$ its area of all walls and $l = 4(l_x + l_y + l_z)$ the summation of all lengths. Taking, for example, a room volume of 250 m^3, and neglecting S and l, there would be more than 720 resonances for frequencies below 300 Hz. This number demonstrates very well that only statistics can give a manageable overview of the sound field in an enclosed space. The situation becomes even more complicated if it comes to rooms with walls at odd angles or curved walls which can not be handled by simple calculations. One way to derive room modes in those cases is through finite-elements simulation.

The knowledge of room modes alone does not provide a great deal of information about the actual sound response, as it is additionally necessary to know the phase of each mode.

1.3.4 Head Orientation

It is common experience that people communicate more easily when facing each other. The reason behind this is that any sound source has propagation directivity characteristics that lead to a non-spherical radiation, mainly determined

by the size and the shape of the source and the frequency being analyzed. However, if the size of the object is small in comparison to the wavelength their directivity pattern becomes spherical. The *directivity* is defined by the different amount of output signal, generated by the direction to or from a point sound source.

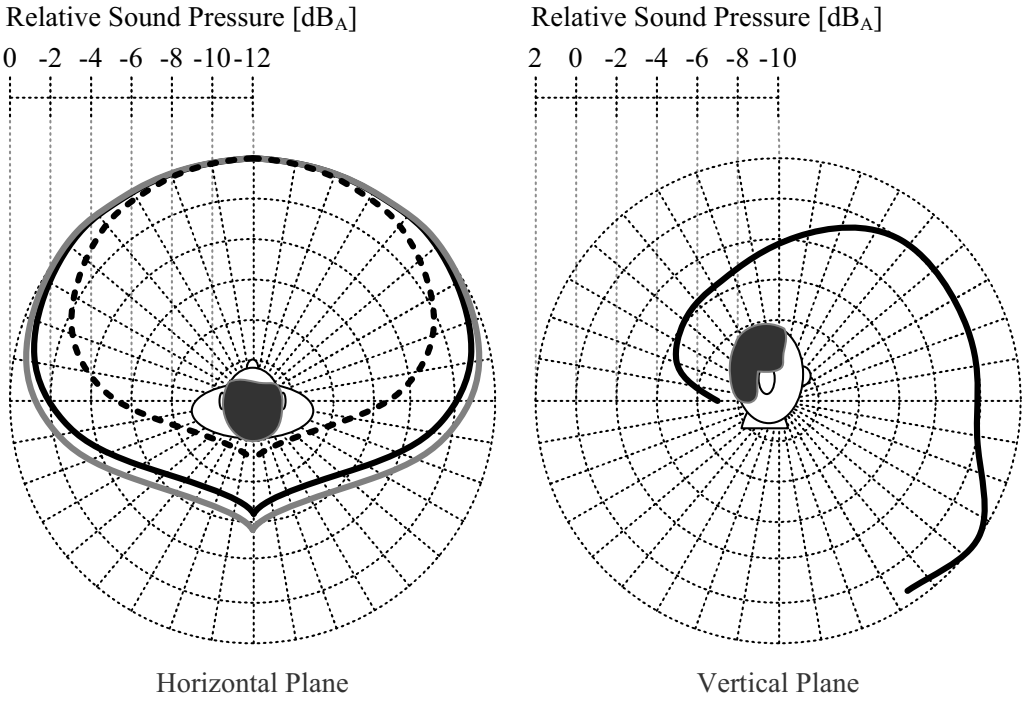


Figure 1.8: Relative pressure levels (A-weighted levels) around the head of an average human talker for three different voice levels (solid line - normal speech, gray line - loud speech, dotted line - quiet speech). The graphics follow measurements by Chu *et al.* [82].

Approximating the head as an oval object with a diameter slightly less than 20 cm and a single sound source, the mouth, one can expect a more directional radiation for frequencies above 500 Hz. Moreover, we expect to observe different properties in the horizontal and in the vertical planes [132]. Measurements of the sound field made by Chu *et al.* [82] in an anechoic chamber at one meter distance around actual human speakers, see Figure 1.8, confirm our expectations. Comparing their laboratory measurements with field measurements [83] they found the measurements in good agreement for male voice spectra. However, they observed some differences for female voice spectra. The directivity patterns between male and female speakers have no significant differences in the directivity pattern although they have different spectral patterns. For loud and normal voice levels similar directivities were observed, the directivity pattern

of quiet voice showed significant changes in the directivity pattern behind the head.

As a result of the made approximation and the measurements by Chu and Warnock as well as measurements by Moreno and Pfretzschner [146], which give similar results, we can conclude that for higher frequencies the human head influences the timbre and causes a radiation behind the head between 5 and 15 dB lower than measured at the sound source. Moreover, one can observe that the direct wavefront propagates just in the frontal hemisphere, and in a way that also depends on the vertical orientation of the head.

Head orientation in the content of speech quality and *word error rate* (WER) has, to our best knowledge, not been published or investigated. A rather new and challenging topic, which is related to the investigation of speech quality and WER is the automatic estimation of head orientation from acoustic signals. The methods used are based on acoustic energy information derived from the related array processing [182] or based on the use of coherences between microphone pair signals as fundamental information on which the head orientation is estimated [69].

1.4 Use of Multiple Microphones in Lecture Speech

In a lecture scenario different numbers and types of microphones are frequently used to record the lecturer as well as questions and comments from the audience. For example a close talk or lapel microphone is used to record the main speaker while at least one additional microphone is used to record questions from the audience. Microphones to record questions from the audience are usually either wireless hand held microphones which are passed on to the questioner or room microphones. The acoustic information provided by the speaker's microphone and those microphones used to record the audience are not well correlated, as they are, in general, far apart from each other and significantly differ in their signal-to-distortion ratios. In such cases, array processing techniques such as blind source separation or beamforming are not applicable. Whenever array processing techniques may not improve over a single channel, to provide optimal transcription of the lecture and questions or comments from the audience, an automatic selection of the channel which provides the highest accuracy is required.

1.5 Specific Vocabulary and Language Structure in Lecture Speech

Lectures contain topic specific vocabulary terms and language structures that are rarely used in general day-to-day conversations [106]. Lectures also follow a specific structure like the introduction, the main topic and the conclusion or summary. Zhang *et al.* [212] have demonstrated that the flow of the lecture can be distinguished by acoustic features.

A primary challenge is to obtain sufficiently relevant training material that can accurately predict the vocabulary and language usage of lectures representing a particular topic.

In order to support vocabulary selection and language adaptation one can rely on additional global or local lecture specific information:

- *global — additional text information*
  Additional text sources can be provided by accompanying slides or other text material such as a text book, articles or conference papers.
- *local — structured time flow*
  Besides the specific structure of a lecture one could use the temporal information of the slides.

The influence of the size and selection of words in the dictionary as well as different sources to train the *language model* (LM) for lectures has been investigated in [174].

1.5.1 Primary Challenge in Vocabulary Selection

To keep the WER and computation time low it is desirable to choose a compact set of words as the decoding vocabulary that includes most of the spoken words in a particular lecture. Since many lectures cover a particular topic and thus include words that are not typically seen in "daily life" speech, general corpora such as *Broadcast News* [108] or *Switchboard* [73] may not be used to adequately choose the vocabulary. Thus, good vocabulary coverage, for unseen lectures, requires an appropriate selection from various available data sources. In practice, topic specific vocabulary can easily be obtained from relevant text sources such as textbooks, journal articles, etc. while large databases such as the aforementioned Switchboard corpora provide backup vocabularies covering

conversational speech which can, in general, not be obtained from written material.

1.5.2 Primary Challenge in Language Modeling

In language modeling, in contrast to vocabulary selection, the problem of proper source material is compounded by the variance in word order and probability of word appearance between spoken language and written text. The lack of proper source material which might not be as easily overcome as in vocabulary selection where a mixture of different domains is appropriate. For example it has been observed by Glass *et al.* [106] that written materials can be a poor predictor of the spoken language used in lectures, even when the topic of these written materials is well matched to that of the lecture. This is due to the fact that many common words or phrases that are often used in informal conversational speech such as in spontaneous presentations are not present in more formal written text.

Although conversational speech training data is useful for modeling the type of spontaneous speech encountered in lectures, many specific word sequences are sparsely represented because they are off-topic. On the other hand, subject-specific text sources are in general only available from written text and thus can provide word sequences involving important content words, however, not patterns of spontaneous speech. Thus, to improve the LM of a particular lecture it must be considered how to effectively utilize multiple LM sources that are different in terms of content and usage patterns. Different solutions will be reviewed in Section 1.6.

1.6 Review of Prior Work

In this section we want to briefly discuss prior work on lecture speech. As already mentioned earlier most work regarding the automatic transcription of lectures have focused on LM adaptation. Thus the main focus of this section is to review language modeling for lecture speech. The last subsection will review speaking-rate dependent decoding as a way to improve the transcription of lectures in Japanese language.

1.6.1 Language Model Adaptation

The question "how to set up a useful special purpose LM from only a small special purpose text" is of major practical relevance to increase the accuracy of lecture speech recognition. The small special purpose text can be given by relevant publications, slides or both. It has been already more or less successfully addressed by different researchers [124, 163, 204, 155, 156, 77, 174].

In general the idea is to set up a LM that has the broad knowledge about word dependencies from the large corpus, but still favors those words and word combinations seen on the small, more relevant corpus. This is achieved by some processes of adaptation of the general purpose LM by the small relevant corpus or by previous seen context.

Different approaches have been suggested to adjust the LM in different context. In the following sections we briefly discuss the basic idea of some approaches and their performance with respect to the transcription of lectures.

1.6.1.1 Linear Interpolation

An adapted LM can be obtained by a linear combination of different LMs, such as a lecture specific LM which is assumed to contain better likelihood estimates for topic-specific words, a LM covering spontaneous speech effects, e.g. Switchboard, and a background LM which is assumed to contain better likelihood for infrequent words. The weights for the linear combination can be, for example, be determined by minimizing the perplexity of an adaptation or held-out text.

An extensive study into linear interpolation of LMs for the lecture domain has been conducted by Fügen [96]. Fügen has proposed an adaptation framework which considers different adaptation levels dependent on the amount and type of available information such as knowledge of the speaker, slides or accompanying proceedings. He has been able, given enough adaptation material, to demonstrate significant reductions in WER.

1.6.1.2 Minimum Discriminant Estimation

The idea behind *minimum discriminant estimation* (MDE) adaptation is the assumption that the uni-gram p_{uni} of the adaptation text is a rather good model for the real uni-gram of the text to be recognized, while for context-dependent

word likelihoods, it is better to rely on an N-gram LM trained on more, but less relevant data [204]. Hence, the resulting adapted LM p_{adapt} should have the uni-gram of the adaptation text as its marginal distribution according to

$$\sum_{h} p_{\text{adapt}}(h) \cdot p_{\text{adapt}}(w|h) = p_{\text{uni}}(w) \tag{1.3}$$

where w represents a word and h the arbitrary word history. Among those models fulfilling (1.3), MDE chooses the one which is the closest to the baseline LM $p_{\text{adapt}}(w|h)$.

1.6.1.3 Probabilistic Latent Semantic Analysis

To characterize topics in a corpus *probabilistic latent semantic analysis* (PLSA) [105] can be used. It can be interpreted as the problem of estimating a kernel of topic sub-spaces derived by topic-annotated training material. By projecting a document, such as slides or proceedings, into the PLSA subspace, the model should force semantically related words, e.g. words associated with a specific topic, to have meaningful probabilities concentrated in one or few basis distributions. A useful feature of PLSA is that a document/topic word distribution can be estimated from a relative small amount of adaptation material.

1.6.1.4 Web Based Language Model

The Internet provides a nearly unlimited text resource. In order to incorporate the knowledge from the Internet keywords or key-phrases have to be extracted from the slides, or even better the conference proceedings, to constitute a relevant search query. To filter out relevant information after the search within the text of the Internet collection it has been proposed to use a baseline LM [145]. Based on the filtered text collection a new LM has to be trained and interpolated with the baseline LM to generate a lecture specific LM. To further reduce the perplexity of the lecture specific LM, Fügen [96] has proposed to derive queries which have been extracted based on tf-idf N-grams heuristics.

1.6.1.5 Cache Language Model

A *cache language model* [126] assumes that used words are more likely to be reused. Thus it stores preceding words in a history C which is much longer than

the history of an N-gram model. The probability of the stored words is raised by a linear interpolation with the baseline N-gram model. The common application of the cache model uses knowledge which is provided by the hypotheses of the recognition system. In [125] it has been suggested to use the slide information by incorporating the text which is presented on a particular slide at a specific time into the context of the cache model C. Thus the LM is dynamic adapted according to the words on the corresponding slide.

1.6.1.6 Performance of the Different Approaches

Assuming that enough text material is available to extract content specific vocabulary and to estimate a content specific LM, best results might be reached by a simple linear interpolation of different LMs [77]. In our own experiments on language modeling for lectures we have observed additional gains by augmenting the different LMs by a web based LM which has been derived from appropriate queries [53]. This is consistent to Kawahara *et al.* who have reported in [125] that the web based as well as the PLSA LM reduces the perplexity and that the word accuracy is improved. They found that depending on the presented topic the web based LM performed better than the PLSA LM. Thus a clear conclusion which of the two models is the better strategy can not be taken. The improvements reported on the PLSA LM are not confirmed by Cettolo *et al.* [77] who have reported no improvement in word accuracy using a PLSA LM. Kawahara *et al.* also found that a cache LM, where the cache depends on the current slide, leads to improved word accuracies in the same order as the web or PLSA LM. The combination of a PLSA LM with a cache LM is able to lead to further improvements in word accuracy [125]. In [204] it has been reported that PLSA has been successfully applied to unsupervisely adapt the LM for individual Japanese lectures.

To conclude, even though various attempts have been made to adapt unsupervised the LM to individual lectures, only marginal improvements have been reported in the literature. On the other hand LMs which have been interpolated from different sources have demonstrated significant improvements in lowering the perplexity and word error. Here simple linear interpolation seems to be the best attempt. The experiments conducted in this thesis use a linear interpolated LM which is derived from different sources: transcripts of lectures, proceedings, web queries, and a background LM containing broadcast news and switchboard data.

1.6.2 Speaking-Rate Dependent Decoding

It has been observed, on Japanese lectures [156], that the speaking rate varies significantly. As a countermeasure different methods have been proposed and investigated:

1. shorter frame length and shift,
2. additional state-skipping in phoneme models,
3. use of syllable models, and
4. change insertion penalty dependent on the speaking rate.

For fast speaking rates methods 1 to 3 have shown improvements over the baseline while method 4 has showed improvements for slow speaking rates. In average a combination could lower the WER by 1.2% where 1.1% of the improvements is already reached by method 1. Therefore, we conclude that a shorter frame length and shift in comparison to a traditional window lenght and shift for read speech are useful (and can be confirmed by our own experiments on English lecture speech—this is, however, also true for other types of data) while the other proposed methods are not leading to statistically significant improvements.

1.7 Organization of this Work

This chapter has given a brief motivation why automatic transcription of lectures is an important and challenging task. We have analyzed and discussed frequently encountered acoustic distortions in the lecture scenario and have reviewed various proposed methods to adapt the LM to lecture speech either in general or for a particular lecture. Chapter 2 reviews different available English and German lecture corpora.

Chapter 3 starts the technical discussion by reviewing spectral estimation techniques based on the Fourier transformation, linear prediction and minimum variance distortionless response. It also includes a review of the properties of the human auditory system and the effects of the bilinear transformation if applied in the time- or frequency domain. The chapter concludes by proposing two refinements to the minimum variance distortionless response, namely warping of the frequency axis prior to spectral estimation and scaling of the spectral envelope.

In order to emphasize the different characteristics of speech signals, in noisy environments, we suggest to steer spectral resolution, on a frame-by-frame basis, to higher or lower frequencies. This becomes possible by the introduction of two warping stages into minimum variance distortionless response spectral estimation which is the topic of Chapter 4.

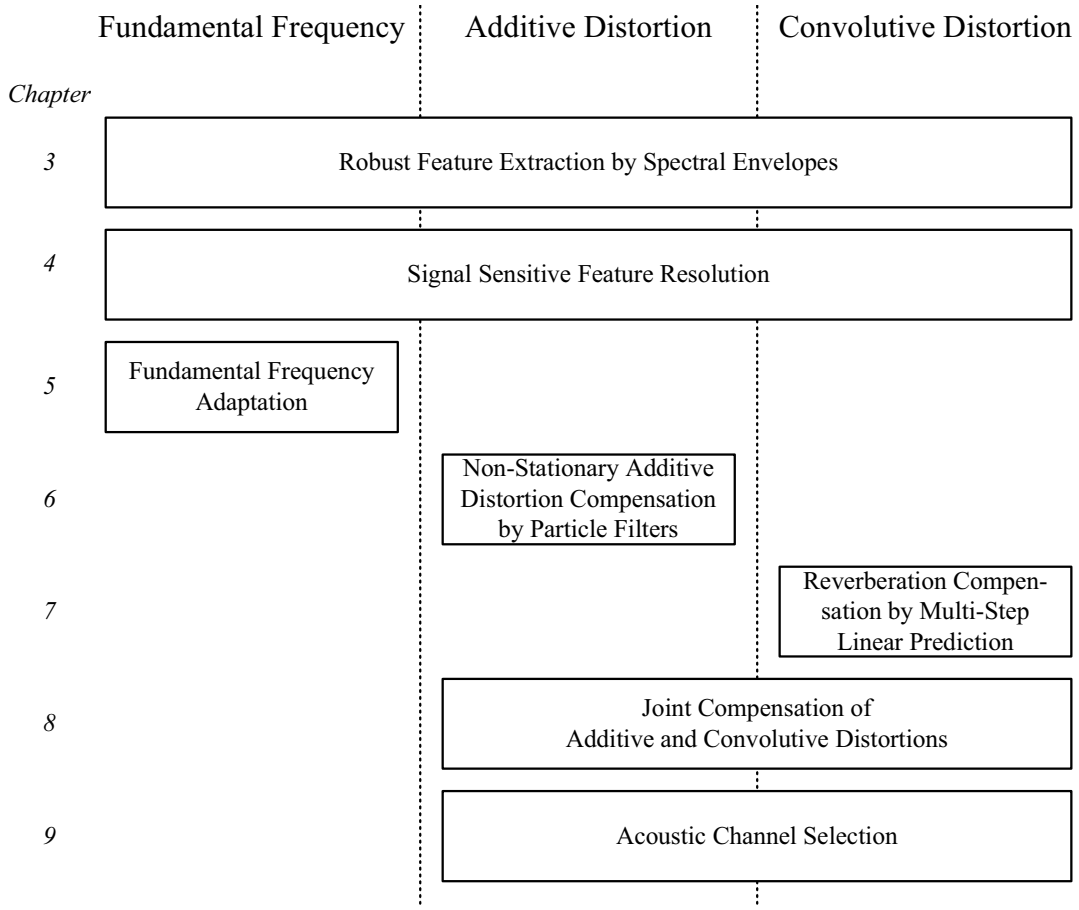


Figure 1.9: Overview of the different methods reviewed, refined and developed in this thesis against different kinds of distortions.

While the previous chapters have tackled all kind of distortions by robust and signal sensitive extraction of speech features, as also indicated in the overview Figure 1.9, Chapter 5 is exclusively focusing on the compensation of speaker dependent differences in the fundamental frequency. First we review maximum likelihood based feature compensation techniques and investigate the effect of fundamental frequency variation on the spectral envelope. Then we propose to reduce the effect on the features caused by variation in the fundamental frequency by adjusting the model order of the spectral envelope estimate according to the acoustic likelihood.

One major source of distortion which can not be compensated well by feature or model adaptation techniques is *non*-stationary additive distortion. In Chapter 6 we first very briefly review probabilistic model techniques to estimate additive distortions before we investigate how to track non-stationary additive distortions in the frequency domain by particle filters for their later removal. In order to overcome the disadvantages and to increase the accuracy of the original framework we propose a couple of refinements:

- augmenting a general acoustic model of clean speech by a phoneme dependent acoustic representation in order to cope for the non-stationarity of speech,
- introducing a dynamic estimate of the autoregressive processes therefore overcoming the need to estimate or update the autoregressive matrix on silence only regions, and
- replacing the vector Taylor series approximation by a deterministic representation which increases computational speed as well as accuracy.

Next we investigate the second source of distortion, *reverberation*, which also can not be compensated well by feature or model adaptation techniques. Chapter 7 starts by reviewing the harmful effect of reverberation on speech features and problems in dereverberation before it reviews multi-step linear prediction, an effective and computational less demanding possibility to estimate late reverberation energies. This method will be needed as an auxillary model in the following chapter.

In the real world non-stationary additive noise as well as reverberation are usually not observed independent from each other and thus should also not be compensated independently of each other. In Chapter 8 we propose a method to jointly estimate and remove non-stationary additive distortions as well as late reverberation.

In scenarios where array processing techniques are not effective, a reliable method to select the microphone-channel, among a couple of microphones, which provides the best speech recognition accuracy might be required. In Chapter 9 we suggest to replace the traditional signal-to-noise ratio by class separability in order to determine the best channel.

All theoretical developments are deemed to fail if they can not be applied on data captured with *real* speakers in *real* acoustic environments. In Chapter 10 we, therefore, present numerous experiments on actual recordings—not artificially distorted—to demonstrate the soundness of the theoretical development presented in the earlier chapters.

Last but not least, Chapter 11 presents the conclusion of this thesis, mention developments of others which might have been inspired by work of the present author and gives an outlook for further investigations and possible improvements.

1.8 Contributions of this Work

In this section we summarize the main contributions of this work in tabular form and give possible relative WER reductions after unsupervised acoustic model adaptation:

- *Robust feature extraction*
  The introduction of the warped MVDR spectral envelope front-end is leading to relative reductions in WER by 2.5% compared to the *mel-frequency cepstral coefficient* (MFCC) front-end.

- *Signal sensitive feature extraction*
  The introduction of the warped-twice MVDR spectral envelope front-end, which steers spectral resolution to lower or higher frequency regions according to the input signal, is leading to relative reductions in WER by 4.7% compared to the MFCC front-end.

- *Fundamental frequency adaptation*
  Adjusting the model order of the MVDR spectral envelope due to the acoustic likelihood of the speech recognition system is leading to relative reductions in WER by 3.4% compared to the MFCC front-end.

- *Refinements in non-stationary additive distortion compensation*
  Replacing the Gaussian mixture approach with the statistical interference approach and introducing a dynamic estimation of the autoregressive progress to predict the noise estimates are leading to relative reductions in WER by 5.6% compared to the particle filter baseline and 11.2% compared to no compensation. An additional signifficant advantage is that the dynamic estimation of the autoregressive progress overcomes the requirement to calculate the autoregressive matrix before the application of the particle filter.

- *Joint compensation of additive and convolutive distortions*
  Extending the dimensionality of the particle filter, in which the additional dimensions represent the scale of the reverberation estimate, is leading to relative reductions in WER by up to 12.6% compared to the particle

filter framework which compensates only for additive distortions and up to 22.4% compared to no compensation.

- *Acoustic channel selection*
  Selecting the "best channel" according to the acoustic quality of channels by class separability instead of the signal-to-noise ratio is leading to relative reductions in WER by up to 7.4%.

Combining the proposed robust feature extraction front-end with the proposed feature enhancement technique which jointly compensates for additive and convolutive distortions can lead to relative reductions in WER by 26.0% compared to the MFCC front-end without feature enhancement after unsupervised acoustic model adaptation.

Chapter 2

Available Lecture and Seminar Corpora

The common refrain in *automatic speech recognition* (ASR) systems when it comes to acoustic and language model training is, "there's no data like more data". At the same time, however, data should be closely related to the task of interest, in our case the recognition of lectures in English or German. Data similar to the data of interest is dubbed *in-domain* as it comes from the same kind of source while other data is called *out-of-domain* respectively. Thus any number of contributing factors, such as language, dialect, acoustic channel, sampling rate, domain or topic, speaking style, speaker age and education, characterizes the classification into in-domain or out-of-domain data.

For good recognition performance a relatively homogeneous collection of in-domain data has to be available or eventually collected, under somewhat controlled circumstances. The scope of this chapter is to describe different available corpora for acoustic and language model training as well as for testing which are related to lecture speech.

2.1 English Corpora

English is probably the most widely investigated language in ASR and indeed different English lecture corpora are available.

2.1.1 Description of the TED corpus

The *Translanguage English Database* (TED) is a corpus of recordings made of oral presentations at Eurospeech 1993 in Berlin. It provides speeches on average lasting 15 minutes, covering a specific topic in speech processing. The recorded speakers can be subdivided into two sets: Native speakers of English with their different dialects and non-native speakers, who speak English at different levels. Speaking styles vary from completely free, prepared, memorized to read speech. In addition to the spoken material, there are associated text materials including written versions of the proceedings and any oral preparations that were supplied by the presenters.

A part of the 188 recordings was released by ELRA/LDC in 2002. The release includes manual transcriptions of 39 of the speeches. The first publication on speech recognition using the TED corpus was published by the University of Twente and ITC-irst [136] which had separated the 39 speakers transcribed by LDS/ELDA into an adaptation set and a test set containing 8 speakers (ca. 2 hours).

2.1.2 Description of the CHIL corpus

The *Computers in the Human Interaction Loop* (CHIL) corpus was collected at various sights, e.g. data was collected during a series of seminars held by students and visitors at the Universität Karlsruhe (TH), Germany, since fall 2003. The students and visitors spoke English, but mainly with German or other European accents, and with varying degrees of fluency. This data collection was done in a very natural setting, as the students were far more concerned with the content of their seminars, their presentation in a foreign language and the questions from the audience than with the recordings themselves. Moreover, the seminar room is a common work space used by other students who are not seminar participants. Hence, there are many "real world" events heard in the recordings, such as door slams, printers, ventilation fans, typing, background chatter, and the like.

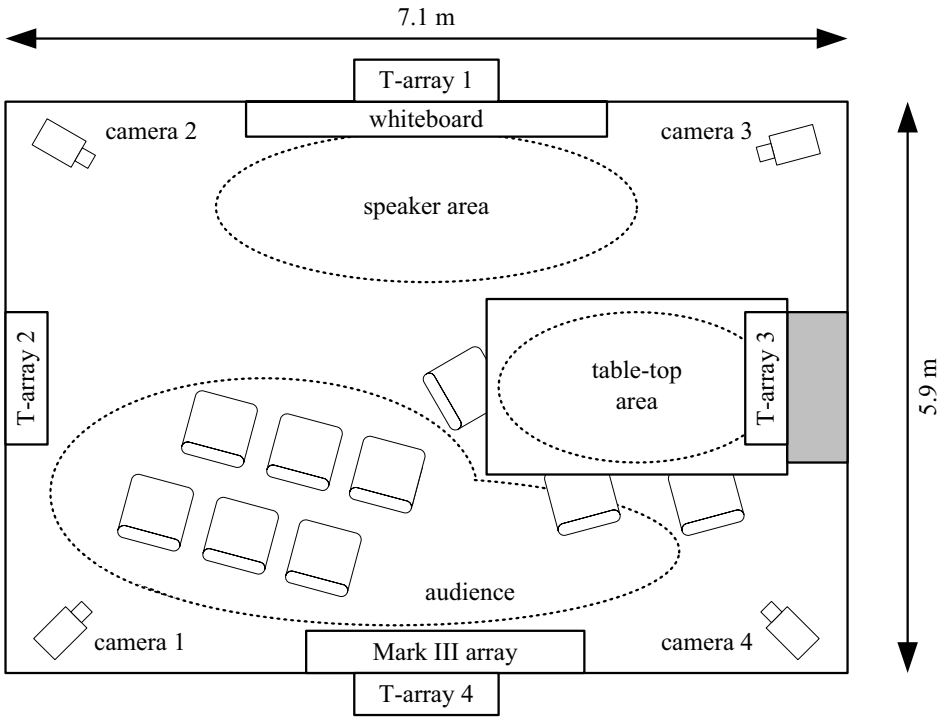


Figure 2.1: The CHIL seminar room layout at the Universität Karlsruhe (TH).

The seminar speakers were recorded with a Sennheiser *close-talking microphone* (CTM), a 64-channel Mark III *microphone array* (MA) developed at the NIST (National Institute of Standards and Technologies) mounted on the wall, four T-shaped MAs with four elements mounted on the four walls of the seminar room and three Shure Microflex table-top microphones located on the work table where the position was not fixed. A brief layout of the seminar room is given in Figure 2.1. All audio files have been recorded at 44.1 kHz with 24 bits per sample. The high sample rate is preferable to permit more accurate position estimations, while the higher bit depth is necessary to accommodate the large dynamic range of the far field speech data. For the recognition process the speech data was down-sampled to 16 kHz with 16 bits per sample. In addition to the audio data capture, the seminars were simultaneously recorded with four calibrated video cameras with a rate of 15 frames per second.

The data from the CTM was manually segmented and transcribed. The data from the far distance microphones was labeled with speech and non-speech regions. The location of the centroid of the speaker's head in the images from the four calibrated video cameras was manually marked every 0.7 second. Based on this marks the true position of the speaker's head in three dimensions could be calculated within an accuracy of approximately 10 cm [94]. The logarithmic

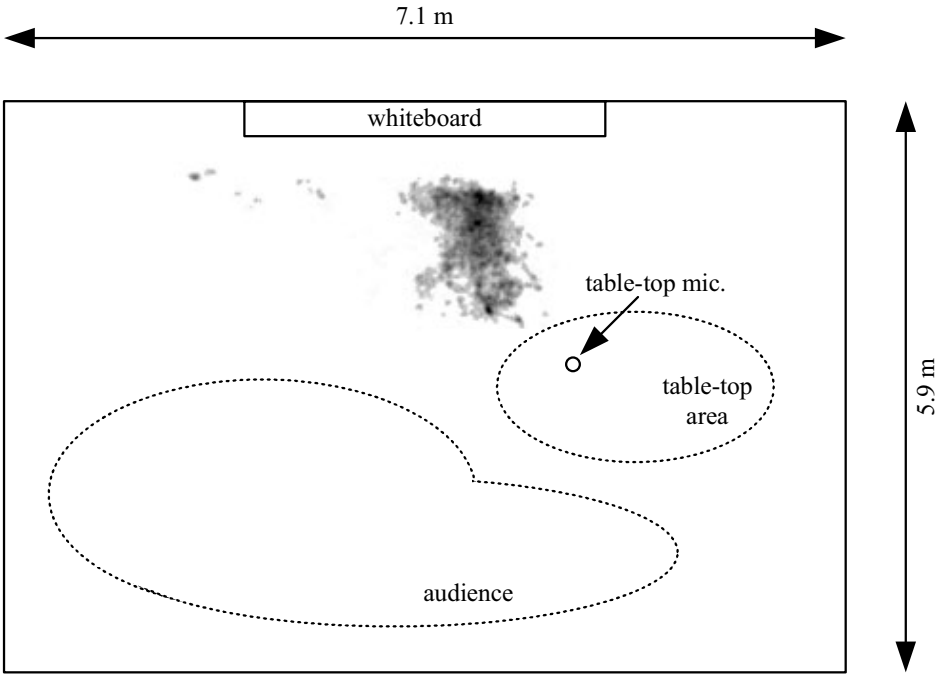


Figure 2.2: Logarithmic probability of speaker's position in Karlsruhe's CHIL room.

probability of speaker's position in Karlsruhe's CHIL room is given in Figure 2.2.

The probability distribution of the distance between the speaker and the average of the three table-top microphones is presented in Figure 2.3. The average distance is 2.37 meter. We observe that the speaker was never closer that one meter to the microphones and nearly not further away than three meters.

The probability distribution of the head orientation is plotted in the logarithmic scale in Figure 2.4. The numbers next to the diagram are the absolute counts. We observe that the speaker is mainly facing the audience. It seems that the speaker is sometimes turning to the whitebord over his right shoulder, but nearly never looked or turned into the direction of his left shoulder, which would be the direction of the table-top microphones. The speaker is nearly never facing the table-top microphones while the microphone array is faced most of the time.

As the CHIL recordings contain multiple distant microphones it enables the realistic evaluation of multi-source and single-source far-distant speech recognition technologies. The corpus presents significant challenges to both modeling components used in ASR, namely the language and acoustic models. Large portions of the data contain non-native, spontaneous, disfluent, and interrupted

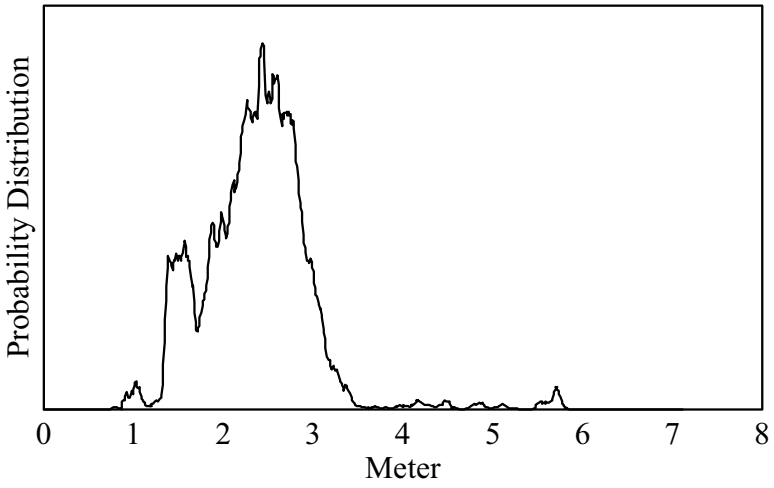


Figure 2.3: Probability distribution of distance between speaker and the average of three table-top microphones in Karlsruhe's CHIL room.

speeches, due to the interactive nature of seminars and the varying degree of the speakers' comfort with their topics. In addition the far-field data captured by table-top and wall mounted microphones (such as the T-shaped arrays and the Mark III) are exacerbated, in comparison to close talk recordings, by the much poorer acoustic signal quality caused by reverberation, background noise and overlapping speech. A drawback of this corpus is the lack of lapel microphones which are frequently used in real lectures.

2.1.3 Description of the MIT corpus

MIT has collected and analyzed a corpus of approximately 300 hours of audio lectures including 6 full MIT courses and 80 hours of seminars from the MIT website from which at least 169 hours have been manual transcribed [107]. Unfortunately, at the time being, the text transcripts are not public available. A release of some of the transcripts, however, might be released over the summer 2008 [1].

[1] Personal correspondence with Jim Glass, MIT Computer Science and Artificial Intelligence Laboratory.

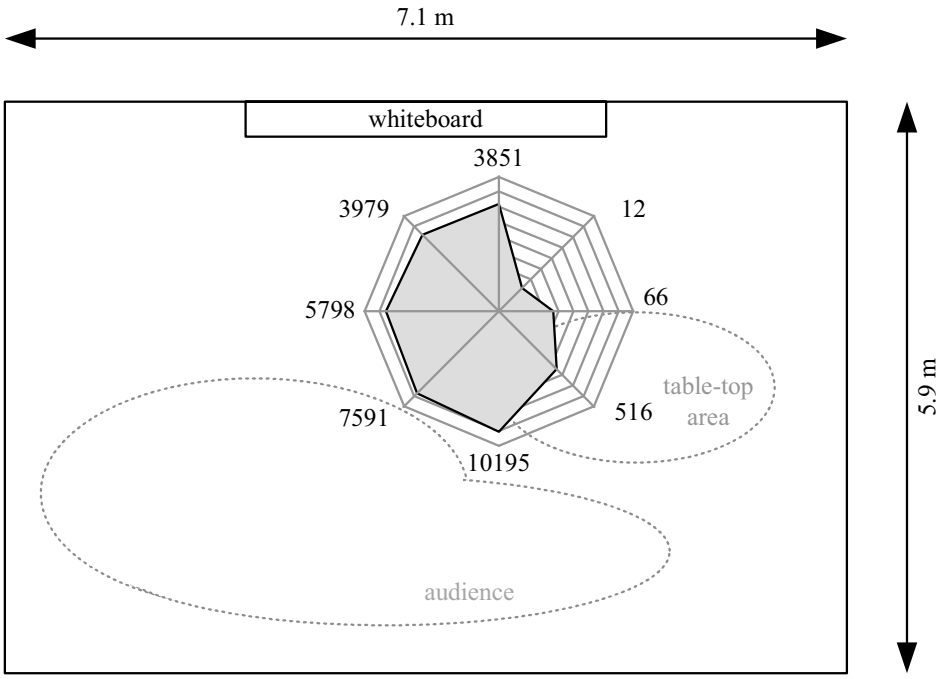


Figure 2.4: Probability distribution of head orientation in Karlsruhe's CHIL room.

2.1.4 Description of Meeting corpora

Due to the limited amount of in-domain lecture training data it might be useful to train the acoustic models of the speech recognition system using acoustic material with similar characteristics. Related corpora of such kind are meetings which are similar to lecture speech, except for overlapping speech which can be easily removed from the training material.

In [112] the use of various meeting corpora for the purpose of automatic speech recognition is explored. The authors conclude that each resource has distinctive features but the provided benefit by pooling the different data let them speak from a generic "conference meeting domain".

We augmented the available lecture training data by the following conference meeting training material: CMU (11 hours) [72], ICSI (72 hours) [65], NIST (13 hours) [190]. Far-field data is available for ICSI and NIST which has not been used for system training if not stated otherwise.

Additional meeting data is available from the AMI project [57]. In our experiments, however, we have not seen reduction in word error rate by adding this corpora and thus it is not used to train the acoustic models.

2.2 German Corpora

In contrast to the large amount of lecture corpora which are available in English, lecture corpora in German are not public available.

2.2.1 Description of the UKA Lecture corpus

We have started recording and transcribing our own German lecture corpus at *Universität Karlsruhe* (UKA). Therefore we have recorded and are still recording lectures given by various scholars at Universität Karlsruhe (TH) with close talk and lapel microphones. Some recordings are augmented with additional microphones. The recordings are transcribed following the transcription guidelines provided by Burger. At the time being more than 20 hours of data by 5 scholars have been transcribed.

2.3 Multilingual Corpora

A multilingual corpus provides transcriptions in different languages.

2.3.1 Description of the EPPS corpus

The *European Parliament Plenary Session* (EPPS) corpus includes recordings from the plenary sessions of the European Parliament. The major part of the sessions takes place in Strasbourg, France while the residual sessions are held in Brussels, Belgium. Today the European Parliament consists of members from 27 countries, and 22 official languages are spoken. The sessions are chaired by the President of the European Parliament. Typically when the president hands over to a member of the parliament, the speaker's microphone is activated. Interjections from the Parliament are therefore softened in the recording. Simultaneous translations of the original speech are provided by interpreters in all official languages of the EU.

It is possible to categorize speakers in two ways: Firstly there are native speakers as well as non-native speakers who have more or less pronounced accent. Secondly there are original speakers and interpreters. Although most of the speeches are planned, almost all speakers exhibit the usual effects known from spontaneous speech (hesitations, false starts, articulatory noises). The interpreters' speaking style is somewhat choppy: dense speech intervals (bursts) alternate with pauses especially when he/she is listening to the original speech.

Europe by Satellite broadcasts the EPPSs live in the original language and the simultaneous translations via satellite on different audio channels: one channel for each official language of the EU and an extra channel for the original untranslated speeches. These channels are additionally available as 30 minute long internet streams for one week after the session. The audio transmissions are monaural. The internet audio streams have a sample rate of 16 kHz and are encoded with the RealAudio Sipro codec at a bit rate of 16 kbit/s. The satellite audio streams have a sample rate of 48 kHz and are encoded with the MPEG 1 layer II codec at a bit rate of 64 kbit/s.

Within the TC-Star [196] project approximately 100 hours of English EPPS speech have been transcribed. Transcriptions in other languages such as Spanish and German are currently in progress. The large amount of acoustic material available is also suitable for unsupervised training.

2.4 Additional Text Sources

In contrast to the previous corpora, including audio and transcriptions, which are useful for acoustic and language modeling, additional text sources can be useful to train or augment the language model. Additional text sources include written text material such as proceedings, newspapers, books and various resources from the internet, and transcriptions of well prepared or read speech such as broadcast news. With this material it is not possible to train a background language model covering a broad number of topics and transcriptions of spontaneous speech, such as Switchboard, to cover spontaneous speech effects. Note that the audio data from broadcast news as well as Switchboard can not be used because the former has a style mismatch (read or well prepared speech) while the latter has a channel mismatch (telephone channel, 8 kHz sampling rate).

Chapter 3

Robust Feature Extraction by Spectral Envelopes

Acoustic modeling requires that the speech waveform $s(t)$ has to be processed in a sequence of feature vectors $O = o_1, o_2, \cdots, o_T$ of a relative small number of dimensions to not run into the problem known as *curse of dimensionality* [63]. This processing is called speech *feature extraction*, *acoustic pre-processing* or *front-end processing*. Feature extraction as applied in *automatic speech recognition* (ASR) systems aims to preserve the information needed to determine the phonetic class while being invariant to other factors including speaker differences such as accent, emotions or speaking rate, or other distortions such as background noise, channel distortion or reverberation. Dimensionality reduction of the feature stream also helps to overcome the curse of dimensionality. This term has been coined by Richard Bellman to describe the problem caused by the exponential increase in volume associated with adding extra dimensions to a space. For example, to capture 10% of the space in 3 dimensions we need to cover 46.42% of each coordinate while in 10 dimensions we need to cover 79.43% of each coordinate.

To conclude feature extraction is a critical step, because if useful information is lost in this step, it can not be recovered in later processing. Feature reduction is elementary to not overstress the amount of training data needed for acoustic modeling.

Since the 1940s short time spectral estimation [130] has been used to carry out speech analysis and became the fundamental approach underlying any speech processing front-end. The non-linear frequency resolution of the ear is implemented into the front-end by a non-linear scaling prior to spectral analysis, by the bilinear transformation, or posterior, by non-linear scaled filter banks. The application of the *cepstrum*[1] marks a milestone in speech feature extraction. Already introduced to speech processing by Noll [162] it took more than a decade to be widely accepted in speech recognition and adopted by the two most widely used front-ends, namely *mel frequency cepstral coefficients* [84] and *perceptual linear prediction* [117]. After the cepstral transformation both front-ends are traditionally augmented by either *dynamic features*, which were introduced into speech feature extraction by Furui [98], or a *stacking* of neighboring frames. The dimension of the augmented features might be reduced by linear discriminant analysis [110] or neural networks.

Over the years many different speech feature extraction methods have been proposed. The variety of methods are distinguished by the extent to which they incorporate information about the speech production (reviewed in Section 3.1) and the human auditory processing and perception (reviewed in Section 3.2), such as the non-linear frequency resolution (Section 3.3), robustness to distortions and length of the observation window as well as the methods used to extract the relevant frequency information (Section 3.4).

3.1 Speech Production Model

Knowledge of the human vocal system and the properties of the resulting speech waveform is essential in designing an approximate model of speech production.

Due to the inherent limitations of the vocal tract, speech signals are highly redundant and contain a variety of different, speaker dependent speech parameters, e.g., pitch, formants, spectra, phase and vocal tract area function. By removing the irrelevant information, contained in the waveform, a simple model of human speech production is obtained. In the case of ASR, for example, only the formants and the spectra are of interest.

The human speech production process reveals that the generation of each *phoneme*, the basic linguistic unit, is characterized by two basic factors:

[1] A transformation to separate the excitation signal and the transfer function by analyzing the output of the natural logarithm of the Fourier transformed signal by an inverse Fourier transformation.

- the excitation by either a random noise or an impulse train, or
- the vocal tract shape.

In order to model speech production, we must model these two factors. To understand the source characteristics, it is assumed that the source and the vocal tract model are independent [86].

As an aid to understand the spectral estimation process for speech signals, we adopt the *source filter* model of speech production [120], wherein speech is divided into two broad classes: *voiced* and *unvoiced*. Voiced speech is quasi-periodic, consisting of a *fundamental frequency* f_0, which can range from 60 Hz for a large man up to 300 Hz for a small woman or child, corresponding to the pitch of a speaker and its harmonics. Unvoiced speech is stochastic in nature and best modeled as white noise convolved with an infinite impulse response filter.

Speech consists of pressure waves created by the flow of air through the vocal tract. These pressure waves originate in the lungs when the speaker exhales. The vocal folds in the larynx can open and close quasi-periodically to interrupt this airflow. This results in *voiced speech*, which is characterized by its periodic and tends to have relatively high energy. Vowels are typical examples.

Some consonants like /f/, /s/ (here /·/ denotes a phoneme) on the other hand are examples of the so called *unvoiced speech*. These sounds are noisy in nature due to turbulence created by the airflow through a narrow constriction in the vocal tract. The positioning of the vocal tract articulators acts as a filter, amplifying certain sound frequencies while attenuating others.

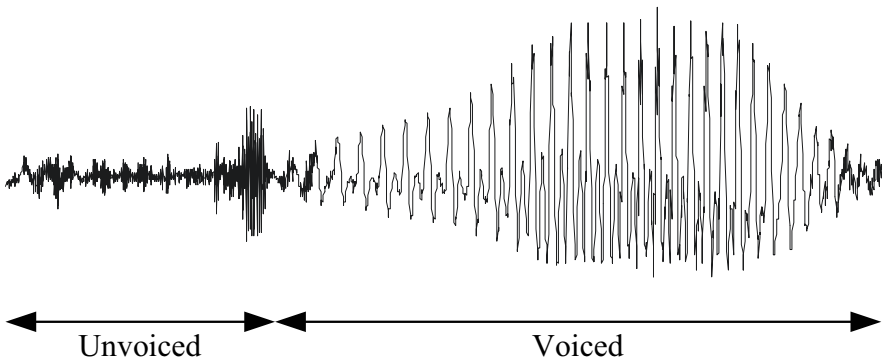


Figure 3.1: A speech segment (time domain) of unvoiced and voiced speech

A time-domain segment of unvoiced and voiced speech is shown in Figure 3.1. A general linear discrete-time system to model this speech production process is shown in Figure 3.2.

In this system, a vocal tract filter $V(z)$ and a lip radiation filter $R(z)$ are excited by a discrete-time excitation signal. The local resonances and anti-resonances are present in the vocal tract filter $V(z)$ which has an overall flat spectral trend. The lips behave as a first order high-pass filter and thus the lip radiation filter $R(z)$ grows at 6 dB/octave.

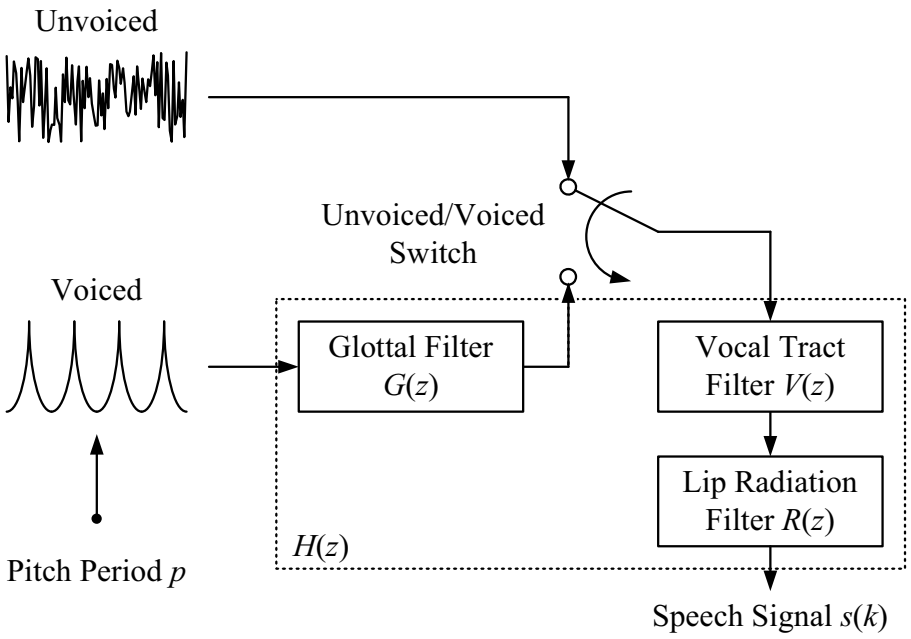


Figure 3.2: Block diagram of the simplified source filter model of speech production

To get the excitation signal for unvoiced speech, a random noise generator with a flat spectrum is typically used. In the case of voiced speech the spectrum is generated by an impulse train with pitch period p and an additional glottal filter $G(z)$. The glottal filter is usually represented by a second order low-pass filter, falling off at 12 dB/octave.

The periodicity of voiced speech gives rise to a spectrum containing harmonics of the fundamental frequency of the vocal fold vibration. A truly periodic sequence, observed over an infinite interval, will have a discrete-line spectrum but voiced sounds are only locally quasi-periodic. The resonances in the power spectrum of voiced speech, known as *formants*, are a product of the shape of the vocal tract. The spectrum for unvoiced speech ranges from flat spectra to those lacking low

frequency components. The variability is due to place of constriction in the vocal tract for different unvoiced sounds—the excitation energy is concentrated in different spectral regions. Due to the continuous evolution of the shape of the vocal tract, speech signals are non-stationary. The gradual movement of vocal tract articulators, however, results in speech that is quasi-stationary over short segments of 5-25 ms. This allows a splitting of the speech signal in short frame segments of 16-25 ms to perform frequency analysis which will be discussed in the following sections.

3.2 Aspects of the Human Auditory System

It is widely known that in speech recognition an adaptation of the aspects of the human auditory system can reduce calculation costs and increase word accuracy [117]. Therefore, in this section, we want to review several aspects of the human auditory system and discuss how the same can be applied to an ASR system.

- **Phase insensitivity**
  The phase components of a speech signal play a negligible role in speech perception, with weak constraints on the degree and type of allowable phase variations [85]. The human ear is fundamentally phase "deaf" and perceives speech primarily based on the magnitude spectrum.

  This can easily be applied in a speech recognition approach by using the absolute of the complex spectrum.

- **Perception of spectral shape**
  Spectral peaks (corresponding to poles in the system function) are more important to perception than spectral valleys (corresponding to zeros) [183].

  This can be applied by using all-pole models such as linear prediction or minimum variance distortionless response.

- **Frequency masking**
  Every short-time power spectrum has an associated masking threshold. The shape of this masking threshold is similar to the spectral envelope of the signal, and any noise inserted below this threshold is "masked" by the desired signal and thus inaudible.

  This feature may be applied by a spectral envelope.

- **Frequency dependent spectral resolution**
  Spectral information in the human auditory system is processed on a non-uniform frequency scale.

  This can be applied by frequency warped spectral features; e.g., by the mel filterbank or bilinear transformation.

- **Temporal masking**
  Sounds can mask noise up to 20 ms in the past (backward masking) and up to 200 ms in the future (forward masking) given that certain conditions are met regarding the spectral distribution of signal energy [169].

  As far as we know this principle has not been applied to speech recognition yet.

3.3 Warping — Time vs. Frequency Domain

In the speech recognition community it is well known that features based on a non-linear frequency mapping improve the recognition accuracy over features on a linear frequency scale [84]. Transforming the linear frequency axis ω to a non-linear frequency axis $\tilde{\omega}$ is called *frequency warping*. One way to achieve frequency warping is to apply non-linear scaled filterbanks, such as mel-filterbanks, to the linear frequency representation. An alternative possibility is to use a conformal mapping such as a first order all-pass filter, also known as a *bilinear transformation* [167][67], which preserves the unit circle. The bilinear transformation is defined in the z-domain as

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \; \forall -1 < \alpha < +1, \tag{3.1}$$

where α is the *warp factor*. The relationship between $\tilde{\omega}$ and ω is non-linear as indicated by the phase function of the all-pass filter [141]

$$\arg\left(e^{-j\tilde{\omega}}\right) = \tilde{\omega} = \omega + 2\arctan\left(\frac{\alpha \sin\omega}{1 - \alpha\cos\omega}\right). \tag{3.2}$$

The mel-scale, which, along with the Bark scale, is one of the most popular non-linear frequency mappings in speech processing, was proposed by Stevens *et al.* in 1937 [191]. It models the non-linear frequency resolution of the human ear and is widely applied in audio feature extraction. A good approximation of the mel-scale by the bilinear transformation is possible, if the warp factor is set accordingly. The optimal warp factor depends on the sampling frequency and can be found by different optimization methods [188]. Figure 3.3 compares the

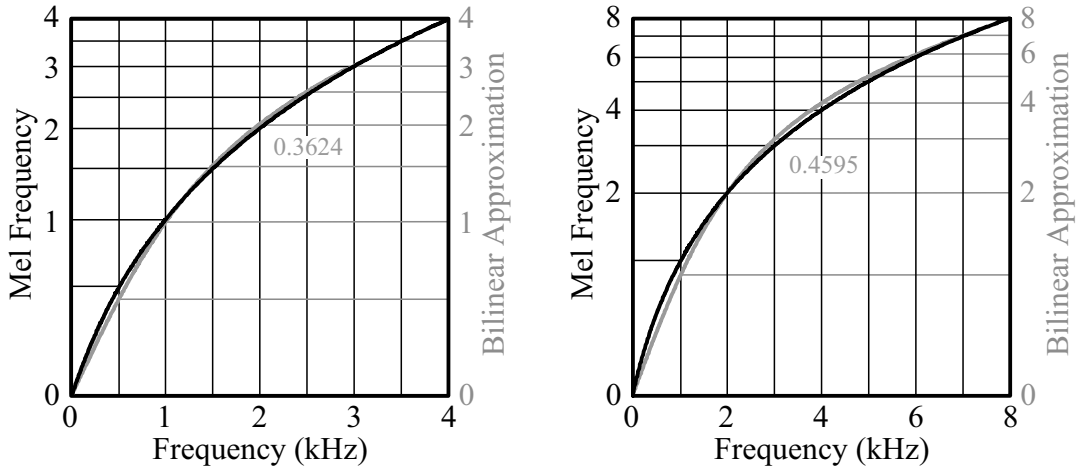


Figure 3.3: Mel-scale can be approximated by the bilinear transformation (gray lines including the warping factor in gray digits) as demonstrated for 8 and 16 kHz sampling rates.

mel-scale with the approximation of the bilinear transformation for a sampling frequency of 8 and 16 kHz.

Frequency warping by bilinear transformation can either be applied in the *time domain* or in the *frequency domain*. In both cases, the frequency axis is non-linearly scaled; however, the effect on the spectral resolution differs for the two domains. This effect can be explained as follows:

- *Warping in the time domain* modifies the values in the autocorrelation matrix and therefore, in the case of linear prediction, more linear prediction coefficients are used, for $\alpha > 0$, to describe lower frequencies and less coefficients to describe higher frequencies.

- *Warping in the frequency domain* does not change the spectral resolution as the transformation is applied after spectral analysis. As indicated by Nocerino *et al.* [161], a general warping transformation in the same domain, such as the bilinear transformation, is equivalent to a matrix multiplication

  $$f_{\text{warp}}[n] = \mathbf{L}_\alpha f[n],$$

  where the matrix $\mathbf{L}_\alpha$ depends on the warp factor. It follows that the values $f_{\text{warp}}[n]$ on the warped scale are a linear interpolation of the values $f[n]$ on the linear scale. In the case of linear prediction or minimum variance distortionless response, the prediction coefficients are not altered as they are calculated before the bilinear transformation is applied.

Figure 3.4 demonstrates the effect of warping on the spectral envelope applied either in the time or in the frequency domain and compares the warped spectral envelopes with the unwarped spectral envelope.

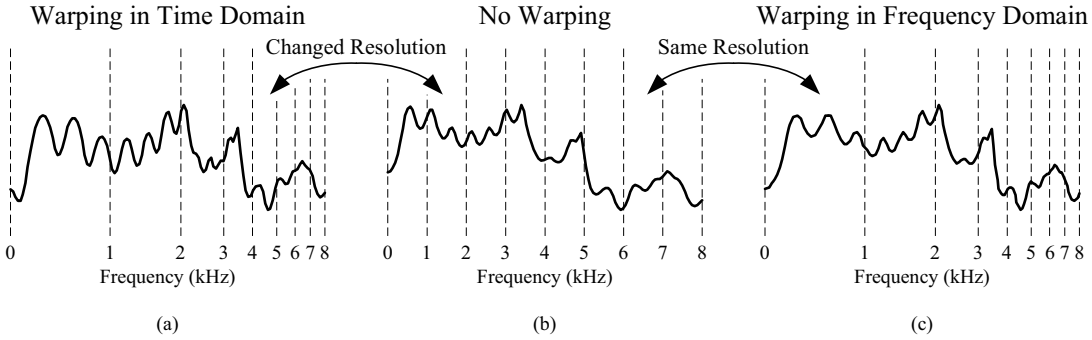


Figure 3.4: Warping in (a) time domain, (b) no warping and (c) warping in frequency domain. While warping in the time domain is changing the spectral resolution and frequency axis, warping in frequency domain does not alter the spectral resolution but still changes the frequency axis.

For clarity we briefly investigate the change of spectral resolution, for the most interesting case, where the bilinear transformation is applied in the time domain with warp factor $\alpha > 0$. In this case we observe that spectral resolution decreases as frequency increases. In comparison to the resolution provided by the linear frequency scale, $\alpha = 0$, the warped frequency resolution increases for low frequencies up to the *turning point frequency* [114]

$$f_{\mathrm{tp}}(\alpha) = \pm \frac{f_{\mathrm{s}}}{2\pi} \arccos(\alpha), \tag{3.3}$$

where f_{s} represents the sampling frequency. At the turning point frequency, the spectral resolution is not affected. Above the turning point frequency, the frequency resolution decreases in comparison to the resolution provided by the linear frequency scale. For $\alpha < 0$, spectral resolution increases as frequency increases.

In the case of spectral envelope estimation Strube [194] has observed that the prediction error minimization of the predictors $\tilde{c}_m$ in the warped domain is equivalent to the minimization of the output power of the warped inverse filter

$$\tilde{C}(z) = 1 + \sum_{m=1}^{M} \tilde{c}_m \tilde{z}^{-m}(z) \tag{3.4}$$

in the linear domain, where each unit delay element z^{-1} is replaced by a bilinear transformation $\tilde{z}^{-1}$. The prediction error is therefore given by

$$E(e^{j\omega}) = |\tilde{C}(e^{j\omega})|^2 S(e^{j\omega}), \tag{3.5}$$

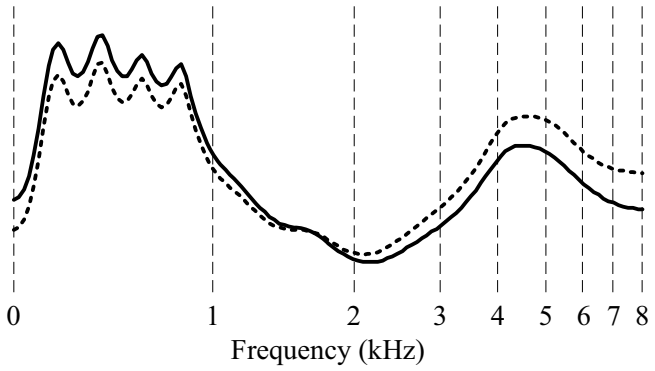


Figure 3.5: The plot of two spectral envelopes demonstrates the effect of spectral tilt. While the spectral tilt is not compensated for the dashed line, it is compensated for the solid line. It is clear to see that high frequencies are *emphasized* if no compensation is applied.

where $S(e^{j\omega})$ is the power spectrum of the signal. The total squared prediction error can be expressed as

$$e = \int_{-\pi}^{\pi} E(e^{j\tilde{\omega}}) d\tilde{\omega} = \int_{-\pi}^{\pi} E(e^{j\omega})\, W^2(e^{j\omega}) d\omega \tag{3.6}$$

where

$$W(z) = \frac{\sqrt{1-\alpha^2}}{1-\alpha z^{-1}}. \tag{3.7}$$

The minimization of the squared prediction error e, however, does *not* lead to minimization of the power, but the power of the error signal filtered by the weighting filter $W(z)$, which is apparent from the presence of this factor in (3.6). Thus, the bilinear transformation introduces an unwanted spectral tilt. To compensate for this negative effect, we apply the inverted weighting function

$$\left|\tilde{W}(\tilde{z}) \cdot \tilde{W}(\tilde{z}^{-1})\right|^{-1} = \frac{\left|1+\alpha\cdot\tilde{z}^{-1}\right|^2}{1-\alpha^2}. \tag{3.8}$$

The effect of the spectral tilt of the bilinear transformation and the remedy by (3.8) are depicted in Figure 3.5.

3.4 Spectral Analysis

Spectral analysis is a fundamental part of speech feature extraction for automatic recognition and many other speech processing algorithms, including compression, coding, and voice conversion. These applications present a variety of

requirements for spectral resolution, variance of the estimated spectra, and to model the frequency response function of the vocal tract during voiced speech. To satisfy these requirements, a broad variety of solutions has been proposed in the literature, all of which can be classified as either *parametric methods*, using a small number of parameters estimated from the data (e.g., linear prediction) or *non-parametric methods* based on periodograms (e.g., the power spectrum).

In this work, we will concentrate on spectral estimation techniques which are useful in extracting the features needed by an ASR system.

The extraction of cepstral features for ASR is traditionally based on one of mel-scaled frequency coefficients, *linear prediction* (LP) [140] or perceptual LP [117]. Though widely used, the basis of each of these feature extraction schemes, namely the Fourier transformation or LP, is ill-suited to reliably estimate spectral envelopes of speech signals, in particular for voiced speech. Therefore, following Murthi and Rao [151, 152], we propose to replace these traditional approaches by *minimum variance distortionless response* (MVDR) spectral estimation. Our investigations were inspired by the work of Dharanipragada and Rao [90], who originally used the MVDR in the front-end of an ASR system. The MVDR approach has been shown to overcome the problems in modeling voiced speech associated with LP spectral estimation techniques. Nevertheless, the basic MVDR approach is not without limitations. In this work, we seek to address these limitations by proposing MVDR estimation on a non-linear frequency scale in Section 3.4.6 and by rescaling the spectral envelope in Section 3.4.7.

3.4.1 Power Spectrum

A very simple approach to spectral analysis of a discrete signal $x[n]$ for $n = 0, \ldots, N$ begins with the calculation of the *discrete circular autocorrelation*

$$R[m] = \sum_{n=0}^{N-1-m} x[n]x[(n+m)\%N] \tag{3.9}$$

where % stands for the modulo of N. Thereafter, the discrete Fourier transformation of the autocorrelation coefficients is calculated, resulting in the *discrete power spectrum*:

$$S(k) = \sum_{n=0}^{N-1} R[n]e^{-j2\pi nk/N} \quad , \; 0 \leq k < N.$$

This power spectrum is widely used in speech processing because it can be

quickly calculated via the fast Fourier transformation [120]. Nonetheless, it is poorly suited to the estimation of speech spectra, because it models spectral peaks and valleys equally well. This characteristic is bad for two reasons:

- Noise in the logarithmic power spectral domain is most evident in spectral valleys; hence, an exact representation of these regions is less useful than an approximation of the spectral power. The spectral peaks, on the other hand, should be faithfully represented as they contain the most relevant information and, as we will see in Section 3.4.7, are less distorted by noise.
- Furthermore, the power spectrum cannot suppress the effect of the fundamental frequency and its harmonics in voiced speech, and therefore provides a poor estimate of the response function of the vocal tract, as it is needed for the recognition of all non-tonal languages.

Therefore, we investigate spectral estimation techniques providing the desired properties in the next sections.

3.4.2 Spectral Envelope

A *spectral envelope* is a curve in the amplitude-frequency plane of the signal energy with following desirable properties:

- **Envelope fit**
  The curve of the spectral envelope should wrap tightly around the power spectrum, linking the peaks. If it is not possible to link every peak, e.g., when the additive analysis finds a group of peaks close to each other with high energies, then it should find a reasonable intermediate path.
- **Robustness**
  The estimation method to derive the envelope has to be applicable to a wide range of signals with very different characteristics, from high pitched harmonic sounds with their wide spaced partials to noisy sounds or mixtures of harmonic and noisy sounds.
- **Smoothness**
  The spectral envelope should provide a certain smoothness. This means it must not oscillate too much, but it should give a general idea of the distribution of the signal energy over frequency.
- **Stability**
  The estimation method to derive the envelope should be stable.

- **Locality**
  The spectral envelope should be local which states that it should be possible to achieve a local change of the spectral envelope, i.e., without affecting the intensity of frequencies further away from the point of manipulation. Ideally, the representation would fulfill the requirement of orthogonality, where one component of the spectral envelope can be changed without affecting the others at all.
- **Speed of synthesis**
  The calculation cost to derive the spectral envelope should be as small as possible.
- **Insensitivity to noise**
  The requirement of insensitivity to noise mandates that the representation is resilient to small changes caused by noise, but must result in equally small or even smaller changes.
- **Minimum Variance**
  The variance of the envelope of the same phoneme should be as small as possible.

3.4.3 LP Envelope

A spectral envelope is commonly modeled by an all-pole model via LP. In LP, the signal $x[n]$ at time n is predicted from a linear combination of the previous M samples and some input $u[n]$ as

$$\hat{x}_M[n] = -\sum_{m=1}^{M} c_m x[n-m] + u[n].$$

Hence, it is necessary to determine the values of the LP coefficients $c_m \; \forall \; m = 1, \cdots, M$ for a given model order M. Given a block of speech data $\mathbf{x} = x[1], \cdots, x[N]$ and assuming that $u[n]$ is unknown and thus $x[n]$ must be predicted from a weighted combination of prior samples, the error between $x[n]$ and the prediction $\hat{x}_M[n]$ is given by

$$\epsilon_M[n] = x[n] + \sum_{m=1}^{M} c_m x[n-m].$$

The set of prediction coefficients $\mathbf{c}$ can then be estimated by minimizing the total squared prediction error

$$\hat{\mathbf{c}} = \underset{\mathbf{c}=[c_1,\cdots,c_M]}{\operatorname{argmin}} \sum_{n=-\infty}^{\infty} \left(x[n] + \sum_{m=1}^{M} c_m \, x[n-m] \right)^2. \tag{3.10}$$

$$c_{0,0} = 1; \qquad \epsilon_0 = R[0]$$

For $n = 1, \cdots, M$

$$k_n = \frac{-1}{\epsilon_{n-1}} \sum_{i=0}^{n-1} R[i-n] c_{i,n-1} \tag{3.11}$$

where

$$c_{i,n} = \begin{cases} 1 & , i = 0 \\ c_{i,n-1} + k_n c^*_{n-i,n-1} & , i = 1, \cdots, n-1 \\ k_n & , i = n \end{cases} \tag{3.12}$$

and

$$\epsilon_n = \epsilon_{n-1}(1 - |k_n|^2). \tag{3.13}$$

After solving (3.11) to (3.13) recursively the coefficients are given by $c_i = c_{i,M}$ for $i = 1, \cdots, M$.

Algorithm 3.1: Computation of linear prediction coefficients by the Levinson-Durbin recursion.

A variety of approaches exists for minimizing (3.10), all of which yield slightly different LP coefficients [140]; e.g., the widely used *Levinson-Durbin recursion* is summarized in Algorithm 3.1.

So far we have introduced the basic concept of LP from a time domain formulation. By applying the z-transformation to (3.10) we obtain the formulation in the frequency domain as

$$\hat{\mathbf{c}} = \underset{\mathbf{c}=[c_1,\cdots,c_M]}{\operatorname{argmin}} \sum_{n=-\infty}^{\infty} \left(\left(z^n + \sum_{m=1}^{M} c_m z^{n-m} \right) X(z) \right)^2.$$

Assuming that $x[n]$ is deterministic, we can set $z = e^{j\omega}$ and apply Parseval's theorem to replace the infinite summation by a finite integral, as

$$\hat{\mathbf{c}} = \underset{\mathbf{c}=[c_1,\cdots,c_M]}{\operatorname{argmin}} \frac{1}{2\pi} \int_{-\omega}^{\omega} \left[A(e^{j\omega}) \cdot X(e^{j\omega}) \right]^2 d\omega \tag{3.14}$$

where

$$A(e^{j\omega}) = 1 + \sum_{m=1}^{M} c_m e^{-jm\omega}.$$

Once the LP coefficients $\mathbf{c}$ and the squared prediction error $\epsilon_M = G^2$ have been obtained from the Levinson-Durbin recursion, the transfer function of the

discrete all-pole model can be expressed as

$$H(z) = \frac{G}{1 + \sum_{m=1}^{M} c_m z^{-m}}.$$

The all-pole spectral estimate $\hat{S}(e^{j\omega})$, henceforth known as the *LP envelope*, is then given by

$$\hat{S}(e^{j\omega}) = |H(e^{j\omega})|^2 = \frac{\epsilon_M}{\left|1 + \sum_{m=1}^{M} c_m e^{-jm\omega}\right|^2}.$$

To understand the limitation of LP envelopes for modeling voiced speech, we need only follow Murthi *et al.* [152] and represent the short-time spectrum of a segment of voiced speech as the overtone series

$$S_{\mathrm{e}}(e^{j\omega_0 \cdot l}) = \sum_{l=1}^{L} 2\pi \frac{|a_l|^2}{4} \Big[\delta(\omega + \omega_0 l) + \delta(\omega - \omega_0 l)\Big] \tag{3.15}$$

where $\omega_0 = 2\pi f_0$ for a fundamental frequency of f_0. In the above, a_l is the amplitude of the lth harmonic and $L = f_s/2f_0$ is the number of harmonics, where f_s is the sampling frequency. We can now substitute (3.15) into (3.14) and write

$$\hat{\mathbf{c}} = \underset{\mathbf{c}=[c_1,\cdots,c_M]}{\operatorname{argmin}} \frac{1}{2\pi} \int_{-\omega}^{\omega} \left|A(e^{j\omega})\right|^2 \cdot S(e^{j\omega})_{\mathrm{harmonic}} d\omega$$

or, equivalently,

$$\underset{\mathbf{c}=[c_1,\cdots,c_M]}{\operatorname{argmin}} \sum_{l=1}^{L} \frac{|a_l|^2}{2} \left|A(e^{jl\omega_0})\right|^2.$$

To achieve the desired minimization of the squared prediction error, the LP envelope attempts to null out the harmonics $l\omega_0$ present in the original spectrum. With increasing M, the ability of the LP envelope to attempt to null out these harmonics increases. But in the process, the zeros of the LP envelope move ever closer to the unit circle, thereby causing sharper contours in the spectral envelope and an overestimation of the spectral power at the harmonics [152]. Such effects are particularly problematic for medium- and high-pitched voices. As such, the LP method does not provide spectral envelopes which reliably estimate the power at the harmonic frequencies in voiced speech.

3.4.4 Warped LP Envelope

Parameterizing the perceptually relevant aspects of the short-time speech spectrum in the front-end of an automatic speech recognition system can improve

the recognition accuracy. The LP all-pole model, however, approximates speech spectra equally well at all frequency bands, which is decidedly *not* the way the human auditory system functions. Moreover, post-processing the LP envelope cannot improve this frequency resolution.

To eliminate this inconsistency between LP based spectral estimation and human auditory analysis, Strube [194] proposed to perform LP analysis on a *warped* frequency axis as previously introduced in Section 3.3.

The inverse filter on the warped frequency axis

$$\tilde{A}(e^{j\tilde{\omega}}) = 1 + \sum_{m=1}^{M} \tilde{c}_m \frac{e^{-jm\omega} - \alpha}{1 - \alpha \cdot e^{-jm\omega}}$$

can be estimated by the Levinson-Durbin recursion using the warped autocorrelation coefficients. Note that applying the bilinear transformation to the spectrum of a finite sequence produces a spectrum corresponding to an infinite sequence,

$$\tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} = X(z) = \sum_{n=0}^{N-1} x[n]z^{-n},$$

therefore, the direct calculation of the warped autocorrelation coefficients

$$\tilde{R}[m] = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{x}[n-m] \tag{3.16}$$

is not feasible. To overcome this problem, a variety of solutions has been proposed [194, 201, 92]. For our experiments, we used the algorithm of Matsumoto *et al.* [142, 141]. To obtain the warped predictors, we must solve the normal equations

$$\sum_{j=1}^{p} \tilde{\phi}[i,j]\tilde{c}_{w,j} = -\tilde{\phi}[i,0] \quad , \; i = 1, \cdots, p \tag{3.17}$$

where

$$\tilde{\phi}[i,j] = \sum_{n=0}^{\infty} y_i[n]y_j[n]$$

and $y_k[n]$ is the output of the ith order all-pass filter excited by $y_0[n] = x[n]$. The last line implies that $\tilde{\phi}[i,j]$ is a component of the warped autocorrelation function

$$\tilde{R}[|i-j|] = \tilde{\phi}[i,j].$$

Thus, (3.17) is revealed to be an autocorrelation equation, exactly like the autocorrelation equation found in standard LP analysis. Furthermore, since $\tilde{\phi}[i,j]$

depends only on the difference $|i-j|$, we can replace (3.16) by

$$\tilde{R}[|i-j|] = \sum_{n=0}^{N-1} x[n] y_{|i-j|}[n] \tag{3.18}$$

where $y_k[n]$ is the output sequence given by

$$y_k[n] = \alpha \cdot (y_k[n-1] - y_{k-1}[n]) - y_{k-1}[n-1].$$

Hence, the warped autocorrelation coefficients $\tilde{\phi}[i,j]$ can be calculated with a finite sum.

Given the warped LP coefficients, we can now obtain the transfer function $H_{\text{warped LP}}(z)$ of the discrete all-pole model in the warped-frequency domain. Thereby, we derive an all-pole spectral estimate, henceforth referred to as the *warped LP envelope*

$$S_{\text{warped LP}}(e^{j\omega}) = |H_{\text{warped LP}}(e^{j\omega})|^2 = \frac{\tilde{\epsilon}_M}{\left|1 + \sum_{m=1}^{M} \tilde{c}_m e^{-jm\omega}\right|^2}.$$

Note that this spectrum is already in the warped frequency domain. Hence, upon setting α to approximate the mel scale, the mel-filterbank in the front-end of an automatic speech recognizer has to be replaced with a filterbank of uniformly spaced half overlapping triangular filters.

If we are interested in a warped envelope in the linear frequency domain, we can calculate the spectral estimate as

$$\tilde{S}(e^{j\omega}) = \frac{\tilde{\epsilon}_M}{\left|1 + \sum_{m=1}^{M} \tilde{c}_m \frac{e^{-jm\omega} - \alpha}{1 - \alpha \cdot e^{-jm\omega}}\right|^2}$$

which differs from conventional LP envelope inasmuch as it uses more parameters to describe the lower frequencies and fewer parameters to describe the higher ones. The conventional LP envelope uses an equal number of parameters for both.

The proposed warping of the LP envelope addresses the inconsistency between LP spectral estimation and that performed by the human auditority system. Unfortunately, for high-pitched voiced speech the lower harmonics become so sparse that single harmonics appear as spectral poles, which are highly undesireable in all-pole modeling. One proposed approach to overcome this drawback is to weight the warped autocorrelation coefficient $\widetilde{R}[m]$ with a lag window [141].

3.4.5 MVDR Envelope

Here we briefly review the *minimum variance distortionless response*[2] as originally introduced by Capon [74]. In order to overcome the problems associated with LP, Murti *et al.* [151] proposed the MVDR for all-pole modeling of speech. A detailed discussion of speech spectral estimation using the MVDR can be found in [152].

MVDR spectral estimation can be posed as a problem in filterbank design, wherein the filterbank is subject to the *distortionless constraint* [115]:

The signal at the frequency of interest ω_{foi} must pass undistorted with unity gain:

$$H(e^{j\omega_{\text{foi}}}) = \sum_{m=0}^{M} h(m)e^{-jm\omega_{\text{foi}}} = 1$$

where $h(m)$ is the mth sample in the time signal associated with $H(e^{j\omega_{\text{foi}}})$. This constraint can be rewritten in vector form as

$$\mathbf{v}^H(e^{j\omega_{\text{foi}}}) \cdot \mathbf{h} = 1$$

where $(\bullet)^H$ is the Hermitian transpose operator and $\mathbf{v}(e^{j\omega_{\text{foi}}})$ is the *fixed frequency vector*

$$\mathbf{v}(e^{j\omega}) = [1, e^{-j\omega}, \ldots, e^{-jM\omega}]^T$$

and

$$\mathbf{h} = [h(0), h(1), \ldots, h(M)]^T.$$

The distortionless filter $\mathbf{h}$ can now be obtained by solving the constrained minimization problem:

$$\min_{\mathbf{h}} \mathbf{h}^H \boldsymbol{\phi} \mathbf{h} \text{ subject to } \mathbf{v}^H(e^{j\omega_{\text{foi}}})\mathbf{h} = 1 \tag{3.19}$$

where $\boldsymbol{\phi}$ is the $(M+1)\cdot(M+1)$ Toeplitz autocorrelation matrix with (l,k)th element $\phi_{l,k} = R[l-k]$ of the input signal x of length L. The autocorrelation $R[m]$ is defined as

$$R[m] = \sum_{n=0}^{L-n} x[n]x[n-m].$$

The solution of the constrained minimization problem is given by [115] as

$$\mathbf{h} = \frac{\boldsymbol{\phi}^{-1}\mathbf{v}(e^{j\omega_{\text{foi}}})}{\mathbf{v}^H(e^{j\omega_{\text{foi}}})\boldsymbol{\phi}^{-1}\mathbf{v}(e^{j\omega_{\text{foi}}})}.$$

[2] Also known as Capon's method or the maximum-likelihood method [153].

1. Compute the LPCs $c_{0\cdots M}^{(M)}$ of order M and the squared prediction error ϵ_M

2. Correlate the LPCs, as

$$\mu_m = \begin{cases} \dfrac{1}{\epsilon_M} \displaystyle\sum_{i=0}^{M-m} (M+1-m-2i) c_i^{(M)} c_{i+m}^{*(M)} & , m = 0, \cdots, M \\ \mu_{-m}^* & , m = -M, \cdots, -1 \end{cases}$$

3. Compute the *MVDR envelope*

$$S_{\text{MVDR}}(e^{j\omega}) = \frac{1}{\sum_{m=-M}^{M} \mu_m e^{-j\omega m}}$$

Algorithm 3.2: Fast computation of the MVDR spectral envelope.

This implies that $\mathbf{h}$ is the impulse response of the distortionless filter for the frequency ω_{foi}. The MVDR envelope of the spectrum $S(e^{-j\omega})$ at frequency ω_{foi} is then obtained as the output of the optimized constrained filter

$$S_{\text{MVDR}}(e^{j\omega_{\text{foi}}}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left|\mathbf{H}(e^{j\omega_{\text{foi}}})\right|^2 S(e^{-j\omega}) d\omega. \tag{3.20}$$

Although MVDR spectral estimation was posed as a problem of designing a distortionless filter for a given frequency ω_{foi}, this was only a conceptual device. The MVDR spectrum can in fact be represented in parametric form for all frequencies and computed very simply as

$$S_{\text{MVDR}}(e^{j\omega}) = \frac{1}{\mathbf{v}^H(e^{j\omega}) \boldsymbol{\phi}^{-1} \mathbf{v}(e^{j\omega})}.$$

Under the assumption that the $(M+1) \cdot (M+1)$ Hermitian Toeplitz correlation matrix $\boldsymbol{\phi}$ is positive definite and thus invertible, Musicus [153] has derived a fast algorithm to calculate the MVDR spectrum from a set of *linear prediction coefficients* (LPC)s, as summarized in Algorithm 3.2.

The proposed MVDR envelope copes well with the problem of power overestimation at the harmonics of voiced speech. To show this, we once more model voiced speech as the sum of harmonics (3.15). Using the frequency form of the MVDR envelope given by (3.20), the spectral estimate at $\omega_0 \cdot l \;\forall\; l = 1, 2, \ldots$ is given by

$$S_{\text{e}}(e^{j\omega_0 \cdot l}) = \sum_{l=1}^{L} \frac{|a_l|^2}{4} \left\{ |H(e^{j\omega_l})|^2 + |H(e^{-j\omega_l})|^2 \right\}.$$

The MVDR distortionless filter $\mathbf{h}$ faithfully preserves the input power at $\omega_0 \cdot l$ while treating the other $(2L-1)$ exponentials as interference and attempting to minimize their influence on the output of the filter. Hence, the MVDR envelope models the perceptually important speech harmonics very well. Unlike warped LP, however, it does not mimic the human auditory system and model the different frequency bands with varying accuracy.

3.4.6 Warped MVDR Envelope

Our goal in this section is to adapt the warping approach to the MVDR envelope to overcome the problems inherent in LP while emphasizing the perceptually relevant portions of the spectrum. Hence, we replace the unit delay elements $e^{-jm\omega}$ of the fixed frequency vector $\mathbf{v}(e^{-j\omega})$ with the bilinear transformation (3.1), to obtain the *warped frequency vector*

$$\tilde{\mathbf{v}}(e^{j\omega}) = \left[1, \frac{e^{-j\omega}-\alpha}{1-\alpha\cdot e^{-j\omega}}, \ldots, \frac{e^{-jM\omega}-\alpha}{1-\alpha\cdot e^{-jM\omega}}\right]^T. \tag{3.21}$$

The distortionless filter $\tilde{\mathbf{h}}$ can now be obtained by solving the constrained minimization problem, wherein the constraint is applied in the warped frequency domain

$$\min_{\tilde{\mathbf{h}}} \tilde{\mathbf{h}}^H \tilde{\phi} \tilde{\mathbf{h}} \text{ subject to } \tilde{\mathbf{v}}^H(e^{j\omega_{\text{foi}}})\tilde{\mathbf{h}} = 1 \tag{3.22}$$

where $\tilde{\phi}$ is a Toeplitz autocorrelation matrix with (l,k)th element $\tilde{\phi}_{l,k} = \tilde{R}[l-k]$ of (3.16).

The solution of the warped constrained minimization problem is very similar to its unwarped counterpart. The warped MVDR envelope of the spectrum $S(e^{-j\omega})$ at frequency ω_{foi} can be obtained as the output of the optimized constrained filter:

$$S_{\text{warped MVDR}}(e^{j\omega_{\text{foi}}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\tilde{\mathbf{H}}(e^{j\omega_{\text{foi}}})\right|^2 S(e^{-j\omega})d\omega$$

with

$$\tilde{\mathbf{H}}(e^{j\omega_{\text{foi}}}) = \sum_{m=0}^{M} \tilde{h}(m)\frac{e^{-jm\omega_{\text{foi}}}-\alpha}{1-\alpha\cdot e^{-jm\omega_{\text{foi}}}} = 1.$$

Under the assumption that the Hermitian Toeplitz correlation matrix $\tilde{\boldsymbol{\phi}}$ is positive definite and thus invertible, Musicus' algorithm [153] can be readily extended to compute the warped MVDR spectrum as summarized in Algorithm 3.3.

1. Compute the warped LPCs $c^{(M)}_{0\cdots M}$ of order M and the squared prediction error $\tilde{\epsilon}_M$

2. Correlate the warped LPCs, according to

$$\tilde{\mu}_k = \begin{cases} \frac{1}{\tilde{\epsilon}_M} \sum_{i=0}^{M-m} (M+1-m-2i)\tilde{c}^{(M)}_i \tilde{c}^{*(M)}_{i+m} & , m = 0, \cdots, M \\ \tilde{\mu}^*_{-m} & , m = -M, \cdots, -1 \end{cases}$$

3. Compute the *warped MVDR envelope*

$$S_{\text{warped MVDR}}(e^{j\omega}) = \frac{1}{\sum_{m=-M}^{M} \tilde{\mu}_m e^{-j\omega m}} \tag{3.23}$$

Algorithm 3.3: Fast computation of the warped MVDR spectral envelope.

Note that the spectrum (3.23) is in the warped frequency domain and therefore we need to either

1. drop the mel spaced triangular filterbank traditionally used in the extraction of mel-frequency cepstral coefficients or

2. replace it by a filterbank of uniform half-overlapping triangular filters for spectral smoothing and feature reduction.

In case we are interested in a warped envelope in the linear frequency domain, we can replace (3.23) by

$$\tilde{S}_{\text{MVDR}}(e^{j\omega}) = \frac{\tilde{\epsilon}_M}{\sum_{m=-M}^{M} \tilde{\mu}_m \frac{e^{-jm\omega}-\alpha}{1-\alpha\cdot e^{-jm\omega}}}.$$

This envelope is different from the conventional MVDR envelope as it, much like the warped LP envelope, uses more parameters to describe the lower frequencies and fewer to describe the higher ones. This is illustrated in Figure 3.6 where the warp factor for the warped MVDR was set to 0.4595 to simulate the mel-frequency for a signal sampled at 16 kHz. While the MVDR exhibits frequency-independent spectral resolution, the mel warped MVDR envelope provides higher resolution for frequencies below 2 kHz and decreasing resolution for higher frequencies. Therefore, warping the MVDR provides properties similar to mel warped LP [123], which cannot be achieved if the MVDR is *followed* by

frequency warping. The warped MVDR envelope does not, however, exhibit the unwanted overestimation of the harmonic peaks in medium- and high-pitched voiced speech witnessed in the warped LP envelope.

As formerly done by Burg [70] for the LP and MVDR envelopes, we can express the relationship between the warped MVDR and the warped LP envelopes as

$$\frac{1}{S^{(M)}_{\text{warped MVDR}}(e^{j\omega})} = \sum_{m=0}^{M} \frac{1}{S^{(m)}_{\text{warped LP}}(e^{j\omega})}. \tag{3.24}$$

Equation (3.24) implies that the warped MVDR spectrum $S^{(M)}_{\text{warped MVDR}}(e^{j\omega})$ of order M is the harmonic mean of the LP spectra $S^{(m)}_{\text{warped LP}}(e^{j\omega})$ of orders 0 through M, and explains why the (warped) MVDR spectrum generally exhibits a smoother frequency response with decreased variance than the corresponding (warped) LP spectrum [152]. This characteristic makes the (warped) MVDR envelope also more interesting for our considerations in Section 5.2, because the "spectral resolution" can be changed by the model order in finer increments.

3.4.7 Scaled MVDR Envelope

In this section we investigate the influence of additive noise on the spectral peaks of the MVDR envelope. The peaks in the logarithmic domain are known to be particularly robust to additive noise, as $\log(a+b) \approx \log(\max\{a,b\})$ [61]. A more general analysis addressing LP envelopes corrupted by additive white noise can be found in [210]. We will show that spectral peaks of the logarithmic (warped) MVDR envelope are not as robust to noise as the spectral peaks of the logarithmic power spectrum. Therefore, we propose to match the MVDR spectrum to the highest spectral peak of the logarithmic power spectrum.

Deeper insight into this phenomenon can be obtained by plotting the energies of the logarithmic power spectrum before and after the addition of noise on the x- and y-axis, respectively. The gray line in Figure 3.7 shows the ideal case of a noise free speech signal; here all points fall on the line $y = x$. In the case of additive noise (black), the lower values of the power spectrum are lifted to higher energies; i.e., the low-energy components are masked by noise and their information is lost or "missing". The missing feature method determines these unreliable parts and either ignore them from subsequent processing, or they are filled by an estimation of their estimated values [180].

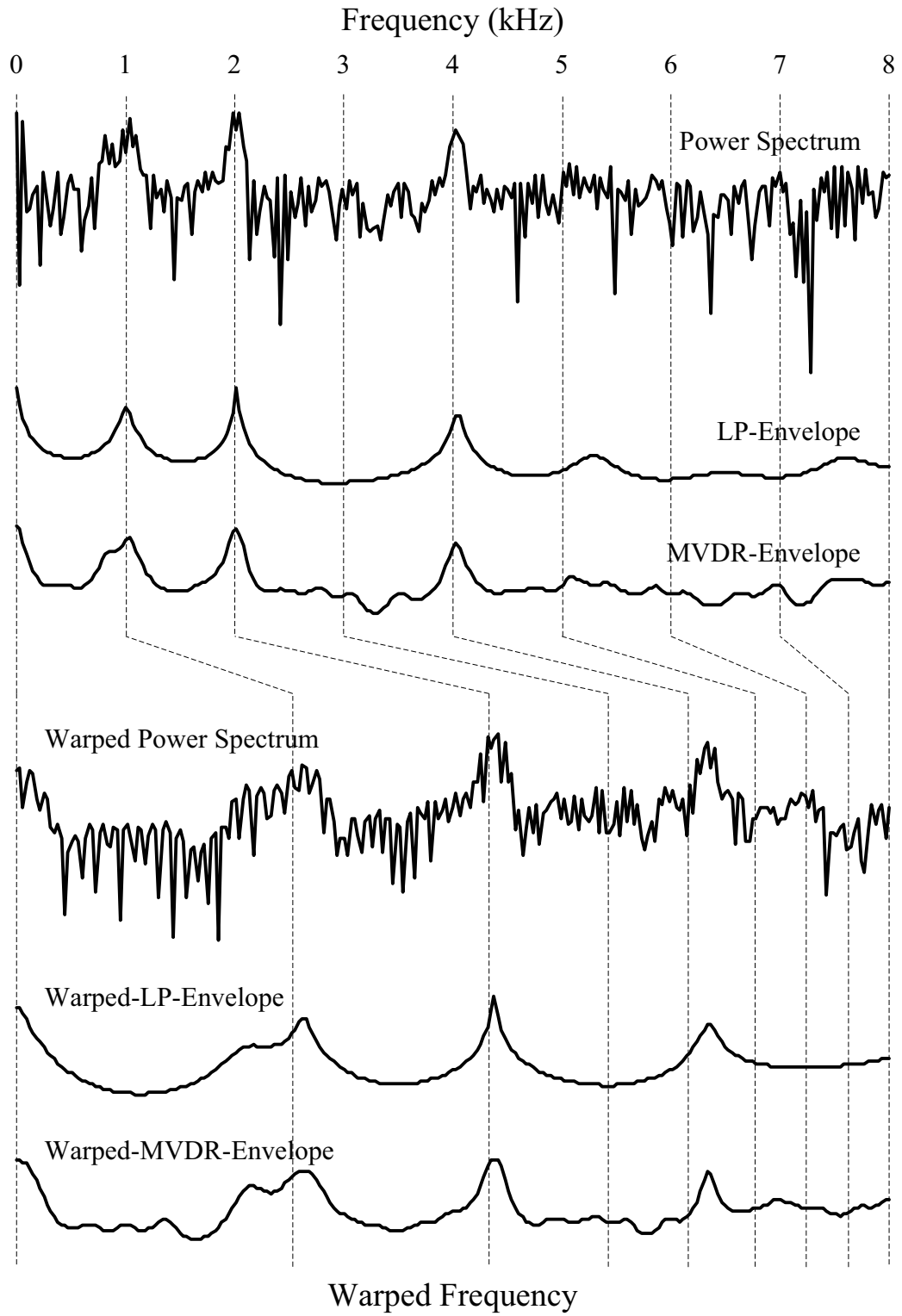


Figure 3.6: Different spectral estimations of voiced speech. LP and mel warped LP of model order 16, MVDR and mel warped MVDR of model order 80.

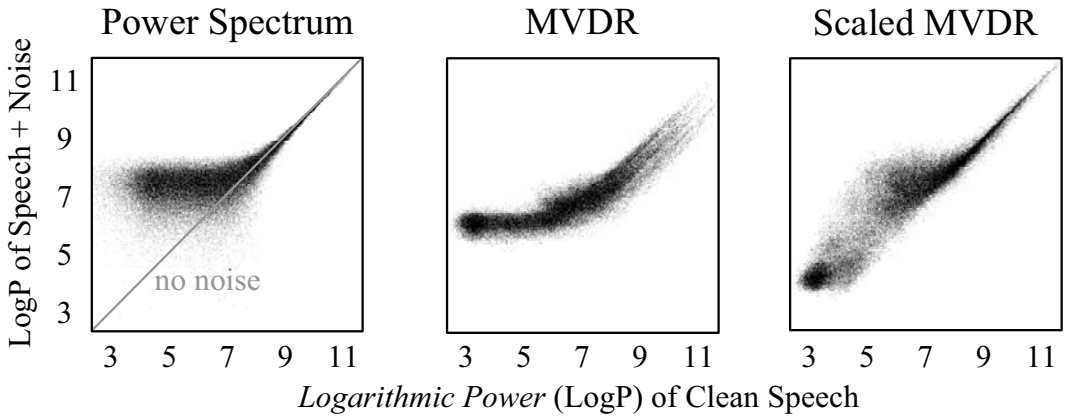


Figure 3.7: Influence of noise (signal to noise ratio = 8 dB) on the logarithmic power of the spectral features for different spectral estimation methods in dependence of their signal energies.

Comparing the influence of noise on the logarithmic scale, derived from the power spectrum with the MVDR envelope of Figure 3.7, clearly illustrates the problem which occurs if additive noise is present: Due to the high variance of the maximum amplitude in the MVDR approach, there is a broad band instead of a narrow ribbon even in the high energy regions. The use of the proposed scaling provides more robust features than both the conventional MVDR envelope and the power spectrum, as it can be seen by comparing the features with each other in Figure 3.7.

Our proposed scaling technique can overcome the drawback of the high variance due to additive noise of the high energy regions, and thereby provides an estimate that is more robust to noise.

3.5 Conclusion

This chapter has investigated different spectral estimation techniques used to efficiently extract acoustic features. We have learned that efficient feature extraction techniques are based on characteristics of the human auditory system. We have reviewed the different properties of the bilinear-transformation if applied in the time or frequency domain. Based on these findings we have proposed time-domain warping of the MVDR spectral envelope estimation by the bilinear-transformation in order to overcome the limitations apparent in linear prediction, warped linear prediction and MVDR spectral envelope estimation

techniques. In addition we have investigated the robustness of the spectral estimates to additive noise and proposed a rescaling of the warped MVDR.

Chapter 4

Signal Sensitive Feature Resolution

Many existing front-end designs are uniform in the sense that they extract the same features regardless of the signal to analyse. This is not a desired quality, as the information needed to discriminate between the phonemes /f/ and /s/ is quite different from the information needed for the discrimination between /aa/ and /ae/. An uniform feature extraction has to compromise to get good coverage over the entire range of distinct phonemes. To overcome the drawback of an uniform feature extraction, it was proposed by Nakatoh *et al.* [154] to adapt the resolution of the spectral envelope in such a way that discrimination between similar phonemes is emphasized. In order to steer the resolution they have suggested to use the knowledge of the signal to move the resolution to lower or higher frequency bands. This is in contrast to model based approaches which will be presented in Section 5.

While Nakatoh *et al.* have demonstrated the soundness of their approach using linear prediction spectral envelopes as a baseline, we adopt their approach to the *minimum variance distortionless response* (MVDR) spectral envelope.

4.1 Warped-Twice MVDR Spectral Envelope

As already mentioned in the introduction, our aim is to change the spectral resolution while keeping the frequency axis fixed. This becomes possible by compensating for the unwanted bending of the frequency axis, introduced by the first warping stage in the time domain, by a second warping stage in the frequency domain.

The use of two bilinear transformations introduces a second free parameters into the MVDR approach [35]. The first free parameter, the *model order*, is already determined by the underlying linear prediction model. Due to the application of two bilinear transformations which apply two warping stages into MVDR spectral estimation, we have proposed to dub this approach *warped-twice MVDR*. While the model order varies the overall spectral resolution of the estimate, compare the different envelopes for model order 30, 60 and 90 in Figure 4.1, the second free parameter, the *warp factor*, bends the frequency axis as already seen in Section 3.3. Bending the frequency axis can be used to apply the mel-scale or, when done on a speaker-dependent basis, to implement *vocal tract length normalization* (VTLN). Although the latter played no role in the experiments described here, as piece-wise linear warping leads to better results. Experiments comparing vocal track length normalization by piece-wise linear and bilinear warping are published in [29].

Fast computation of the warped-twice MVDR envelope

A fast computation of the warped-twice MVDR envelope of model order M is possible by extending Musicus' algorithm. A flowchart diagram of the individual processing steps is given in Figure 4.2.

1. **Computation of the warped autocorrelation coefficients**
   To compute warped autocorrelation coefficients $\tilde{R}[0] \cdots \tilde{R}[M+1]$, the linear frequency axis ω has to be transformed to a warped frequency axis $\tilde{\omega}$ by replacing the unit delay element z^{-1} with a bilinear transformation (3.1). This leads to the warped autocorrelation coefficients

$$\tilde{R}[n] = \sum_{m=0}^{\infty} \tilde{x}[m]\tilde{x}[m-n], \tag{4.1}$$

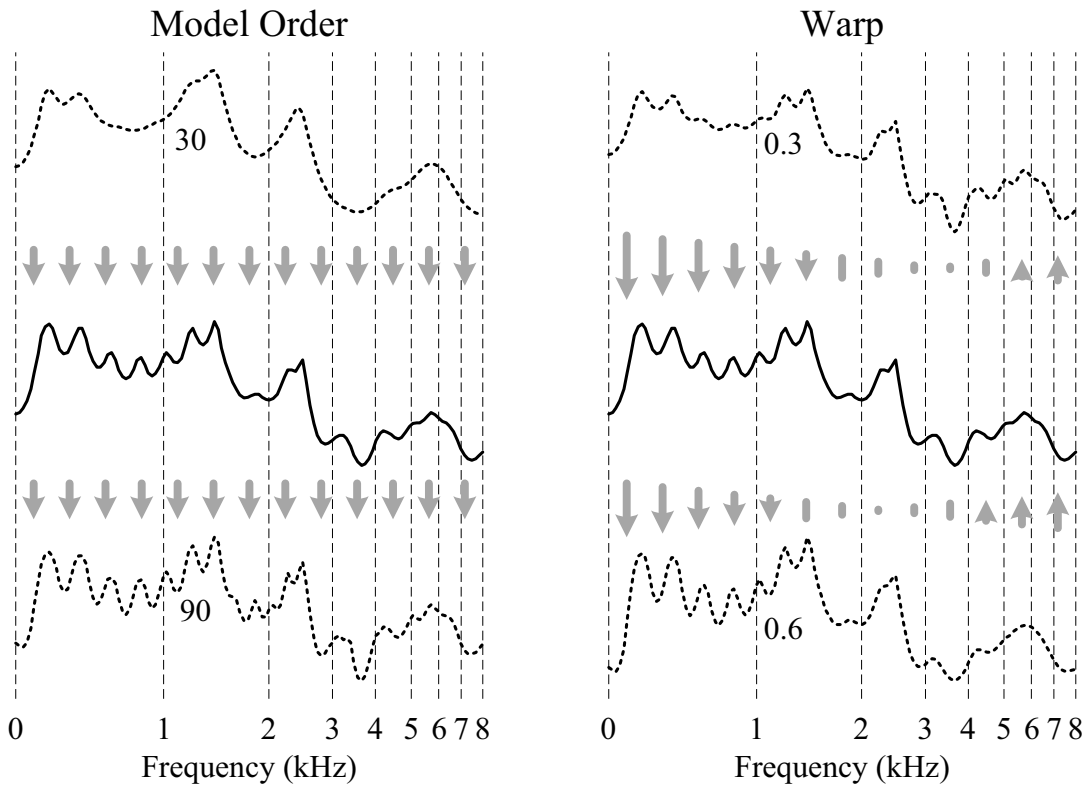


Figure 4.1: The solid lines show warped-twice MVDR spectral envelopes with model order 60, $\alpha = 0.4595$ and $\alpha_{\text{mel}} = 0.4595$. Its counterparts with lower and higher model order or warp factor α are given by dashed lines. The arrows point in the direction of higher resolution. While the model order changes the overall spectral resolution at all frequencies, the warp factor moves spectral resolution to lower or higher frequencies. At the turning point frequency, the resolution is not affected and the direction of the arrows changes.

where the samples $\tilde{x}$ of the warped speech signal are fully defined by [113]

$$\sum_{m=0}^{\infty} \tilde{x}[m]\tilde{z}^{-m} = \sum_{m=0}^{\infty} x[m]z^{-m}. \tag{4.2}$$

Note that we need to calculate $M+1$ warped autocorrelation coefficients (the additional coefficient is used in the compensation step).

2. **Calculation of the compensation warp factor**
   To fit the final frequency axis to the mel-scale, we need to compensate for the first warping stage with value α in a second warping stage with the warp factor

$$\beta = \frac{\alpha - \alpha_{\text{mel}}}{1 - \alpha \cdot \alpha_{\text{mel}}}. \tag{4.3}$$

3. **Compensation for the spectral tilt**
   To compensate for the distortion introduced by the concatenated bilinear transformations with warp factors α and β, we first concatenate the cascade of warping stages into a single warping stage with the warp factor

$$\chi = \frac{\alpha + \beta}{1 + \alpha \cdot \beta}. \tag{4.4}$$

   A derivation of (4.4) is provided in [54]. To get a flat transfer function, we now apply the inverted weighting function

$$\left| \tilde{W}(\tilde{z}) \cdot \tilde{W}(\tilde{z}^{-1}) \right|^{-1} \tag{4.5}$$

   to the warped autocorrelation coefficients, which can be realized as a second order finite impulse response filter

$$\widehat{R}[m] = \frac{1 + \chi^2 + \chi \cdot \tilde{R}[m-1] + \chi \cdot \tilde{R}[m+1]}{1 - \chi^2}. \tag{4.6}$$

4. **Computation of the warped *linear prediction coefficients* (LPC)s**
   The warped LPCs $\widehat{a}^{(M)}_{0\cdots M}$ can now be estimated using the Levinson-Durbin recursion [168], by replacing the linear autocorrelation coefficients R with their warped and spectral tilt compensated counterparts $\widehat{R}$.

5. **Correlation of the warped LPCs**
   The MVDR parameters $\widehat{\mu}_{-m}$ can be related to the LPCs by

$$\widehat{\mu}_m = \begin{cases} \frac{1}{\widehat{\epsilon}} \sum_{i=0}^{M-m} (M+1-m-2i) c_i^{(M)} c_{i+m}^{*(M)} & , m = 0, \cdots, M \\ \\ \widehat{\mu}^*_{-m} & , m = -M, \cdots, -1 \end{cases}$$

6. **Computation of the warped-twice MVDR envelope**
   The spectral estimate can now be obtained by

$$S_{\mathrm{W2MVDR}}(e^{j\omega}) = \frac{1}{\sum_{m=-M}^{M} \widehat{\mu}_m \frac{e^{j\omega} - \beta}{1 - \beta \cdot e^{j\omega}}}, \tag{4.7}$$

   where $\widehat{\epsilon}$ is the prediction error variance.

   Note that (4.7) is already in the mel-warped frequency domain and therefore we need to either

   (a) drop the mel spaced triangular filterbank traditionally used in the extraction of mel-frequency cepstral coefficients or

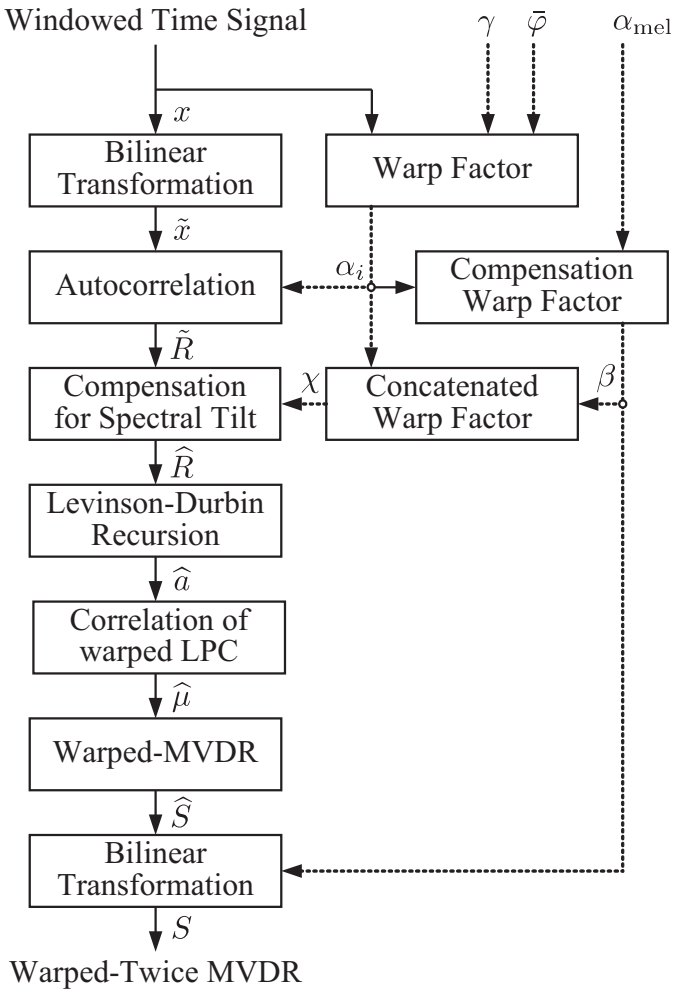


Figure 4.2: Flowchart of warped-twice minimum variance distortionless response. Symbols are defined as in the text.

(b) replace it by a filterbank of uniform half-overlapping triangular filters for spectral smoothing and feature reduction.

7. **Scaling of the warped-twice MVDR envelope**
   Similar to the warped MVDR, we match the warped-twice MVDR envelope to the highest spectral peak of the power spectrum.

Implementation Issues

Frequency warping including linear or non-linear VTLN can be realized using filterbanks. Such filterbanks have to be adjusted for each individual frame according to the compensation warp factor β and the VTLN parameter. In practice, however, it is sufficient to use a limited number of pre-calculated filterbanks; in this way, warped-twice MVDR spectral estimation can be implemented with only a very small overhead when compared to warped MVDR spectral estimation.

4.2 Steering Function

To support automatic speech recognition, the free parameters of the warped-twice MVDR envelope have to be adapted in such a way that classification relevant characteristics are emphasized while less relevant information is suppressed. Nakatoh *et al.* [154] proposed a method for steering the spectral resolution to lower or higher frequencies whereby, for every frame k, the first two autocorrelation coefficients were used to define the *steering function*

$$\varphi_k = \frac{R_k[1]}{R_k[0]}. \tag{4.8}$$

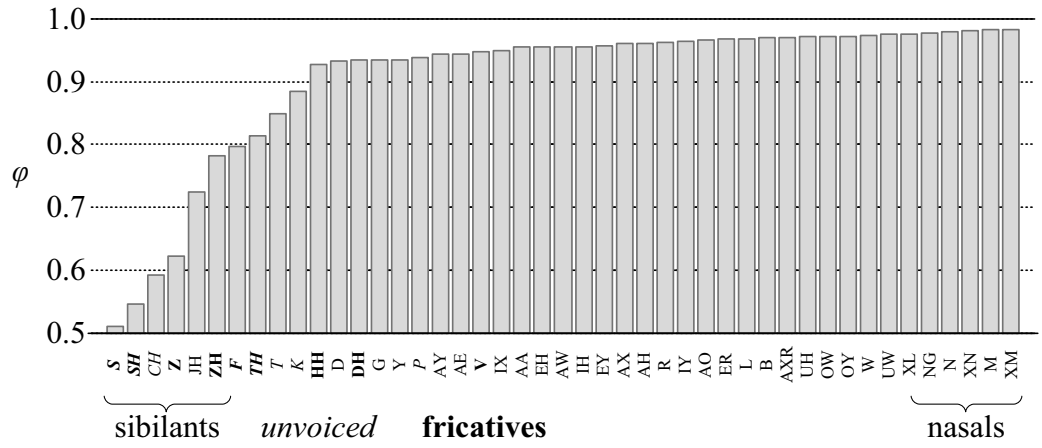


Figure 4.3: Values of the normalized first autocorrelation coefficient by phonemes. Different phoneme classes group either for small values, e.g. sibilants, unvoiced (italic) and fricatives (bold) or for high values, e.g. nasals.

To adjust the sensitivity of the steering function the factor γ is introduced, and the subtraction of the bias $\bar{\varphi} = \sum_k \varphi_k$ (i.e., the average over all values in the training set) keeps the average of α close to α_{mel}. This leads to

$$\alpha_k = \gamma \cdot (\varphi_k - \bar{\varphi}) + \alpha_{\text{mel}}. \tag{4.9}$$

The last equation is a slight modification of that originally proposed by Nakatoh *et al.* For our experiments, we kept γ fixed at 0.1; different values may lead to slightly different results. The influence of γ has been investigated in [154].

Figure 4.3 gives the different values of the normalized first autocorrelation coefficient φ averaged over all samples for each individual phoneme. A clear separation between the fricatives and non-fricatives can be observed. *Fricatives* are consonants produced by forcing air through a narrow channel made by placing two articulators close together. A particular subset of fricatives are the *sibilants* made by directing a jet of air through a narrow channel in the vocal tract towards the sharp edge of the teeth. Sibilants are louder than their non-sibilant counterparts, and most of their acoustic energy occurs at higher frequencies than for non-sibilant fricatives. A detailed discussion about the properties of different phoneme classes can be found in [166].

4.3 Conclusion

Based on the different behavior of the bilinear-transformation applied in the time or in the frequency domain as outlines in Section 3.3 we have proposed to use two warping stages within the MVDR estimation, one in the time and the other in the frequency domain. Following Nakatoh *et al.* we have noted that it is then possible to steer feature resolution to lower or higher frequency regions according to the input signal.

Chapter 5

Fundamental Frequency Adaptation

It is well known that inter-speaker acoustic variability is one of the major sources of error in automatic speech recognition. Typical sources of acoustic variations among speakers are the anatomical characteristics (e.g. vocal tract length, dimension of mouth and nasal cavities) and the speaking habits (e.g. accent, dialect and speaking rate). To reduce the inter-speaker acoustic variability, speech feature adaptation maps acoustic observations into a normalized acoustic domain. For each speaker, a transformation, or a series of transformations, is estimated with the goal to reduce the mismatch between the acoustic data of the speaker and the acoustic model of the recognition system. This transformation can be estimated by different mapping functions.

We first review the probably most widely used mapping function in automatic speech recognition, namely the *vocal tract length normalization* (VTLN). This task is commonly performed in the linear frequency domain. To allow the application within the warped and warped-twice *minimum variance distortionless response* (MVDR) front-ends, the mapping of the vocal tract has to be adapted into the non-linear mel-frequency domain, the way it works in the traditional mel-frequency cepstral coefficients front-end.

Besides VTLN, *constrained maximum likelihood linear regression* proposed by

Gales [100] is the second popular feature transformation technique. Here the goal is to transform the acoustic observation mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ by an affine transformation: The constrained model space transformation

$$\hat{\boldsymbol{\mu}} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^*$$

where $\mathbf{A}$ and $\mathbf{b}$ represent the matrix and the offset vector. The term *constrained* reflects that the same transformation matrix $\mathbf{A}$ is applied to transform the mean vector and the covariance matrix.

In contrast to these two widely used feature space methods which aim to compensate for speaker variation, such as vocal tract length, we propose to compensate for the variances in fundamental frequency. This becomes possible by the introduction of the model order as a mapping function. For a particular speaker we assume that the fundamental frequency changes to be small and therefore we propose to change the model order on the basis of individual speakers.

A frame based selection of the model order, which in concept is similar to Section 4, however changes the overall resolution, has been investigated by Wölfel [33]. There two frame based objective functions, namely autocorrelation and spectral entropy, have been compared to maximum likelihood per speaker as presented in Section 5.2. While the word error rate for maximum likelihood and autocorrelation based methods have been nearly alike, spectral entropy based methods do not show similar reduction in word error rate. The word error rate, however, is still lower than using a fixed model order.

5.1 Vocal Tract Length Normalization

As the name implies, *vocal tract length normalization* [58, 135] tries to normalize the length of a speaker's vocal tract. Much like a longer pipe in an organ produces a lower tone than a short pipe, the resonances or *formants* produced by a longer vocal tract will be lower than those of a shorter vocal tract.

For speaker-independent speech recognition a spectrum must be estimated that provides features that are well-matched to the speaker-independent acoustic models of the recognizer. Acero [54] has proposed to apply the bilinear transformation as a means of achieving a frequency warping effect. McDounough [143] used the property of the bilinear transformation that warping can be achieved through a linear transformation of the cepstral coefficients. In order to choose an optimal vocal tract normalization we can calculate the likelihood of the adaptation data C given the corresponding word string W and choose the warp factor

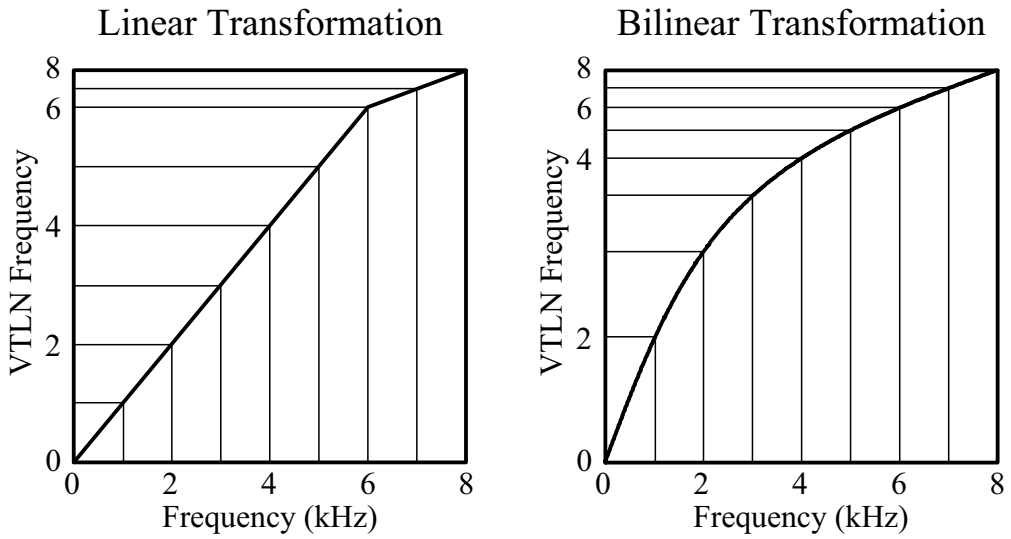


Figure 5.1: Mapping the vocal tract length to a normalized length by a piece wise linear and a bilinear transformation.

with the best likelihood

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \, \mathrm{P}(C|\lambda_l, W). \tag{5.1}$$

VTLN can be applied in different ways, e.g. using a piece wise linear or a bilinear transformation see Figure 5.1. It can also be applied in the linear or mel scale as shown in Figure 5.2. As already mentioned in Section 4.1 a piece wise linear transformation is superior to the bilinear transformation. Thus, in order to allow for the best possible performance, we can not just readily apply the bilinear transformation in the mel-domain, but have to find a mapping function which maps the piece wise linear transformation into the mel-domain. To allow for an efficient normalization in the warped-MVDR domain, the mapped linear transformation can directly be applied within the linear filterbanks.

5.2 Speaker-Dependent Model Order Selection

In general, the goal of all-pole modeling in speech processing is to define an envelope that provides the best possible estimate of the transfer function of the vocal tract, while suppressing the fundamental frequency and its harmonics. In this section, we will investigate the influence of the fundamental frequency on the optimal *model order* (MO) of an all-pole model. Moreover, we propose a

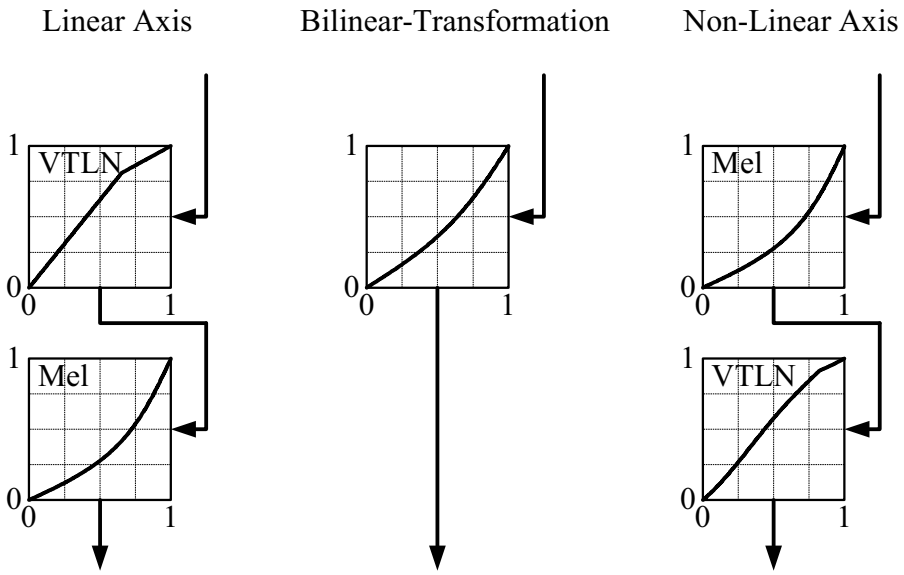


Figure 5.2: Implementation of the VTLN on the linear (left image) and non-linear (right image) frequency scale by a piece wise linear mapping. Center image shows the non-linear mapping and VTLN by a bilinear transformation.

speaker-dependent MO selection to improve the performance of an *automatic speech recognition* (ASR) system.

The selection of the MO is an important, but often difficult, aspect of using all-pole models for a particular application. Intuitively, the optimal MO depends on the length of data over which the MO will be applied. On the one hand, larger MOs can capture the dynamics of a richer class of signals. On the other hand, larger MOs also require proportionally larger data sets for the parameters to be robustly estimated.

In speech recognition, the MO of the spectral envelope estimate is usually set to reach the best recognition results on a development set and then kept constant for all speakers. This might not lead to the best possible recognition performance. Therefore, we investigate how the *fundamental frequency* f_0 influences the estimate of envelopes as a function of the MO.

As a first step we have generated speech-like signals, obtained by convolving an impulse train with a given f_0 and a vocal tract response function $H(z)$ with three formants at 1000, 2000, and 4000 Hz. With these signals we generated spectral envelopes using various MOs. Comparing these envelopes to the vocal tract response function $H(z)$, Figure 5.3 we observe that different MOs approximate the reference transfer function more or less precisely. Furthermore, we

realize that high MOs in combination with a high f_0 emphasize the excitation frequency and its harmonics. Therefore, reasoning as follows, we expect that voiced speech with a high f_0 will be modeled better by a low MO and vice versa: The fundamental frequency f_0 defines the interval between successive harmonics in the frequency domain. As sparse harmonics result in a lower resolution than dense harmonics, the MO should be reduced for sparse harmonics to obtain an optimal estimate.

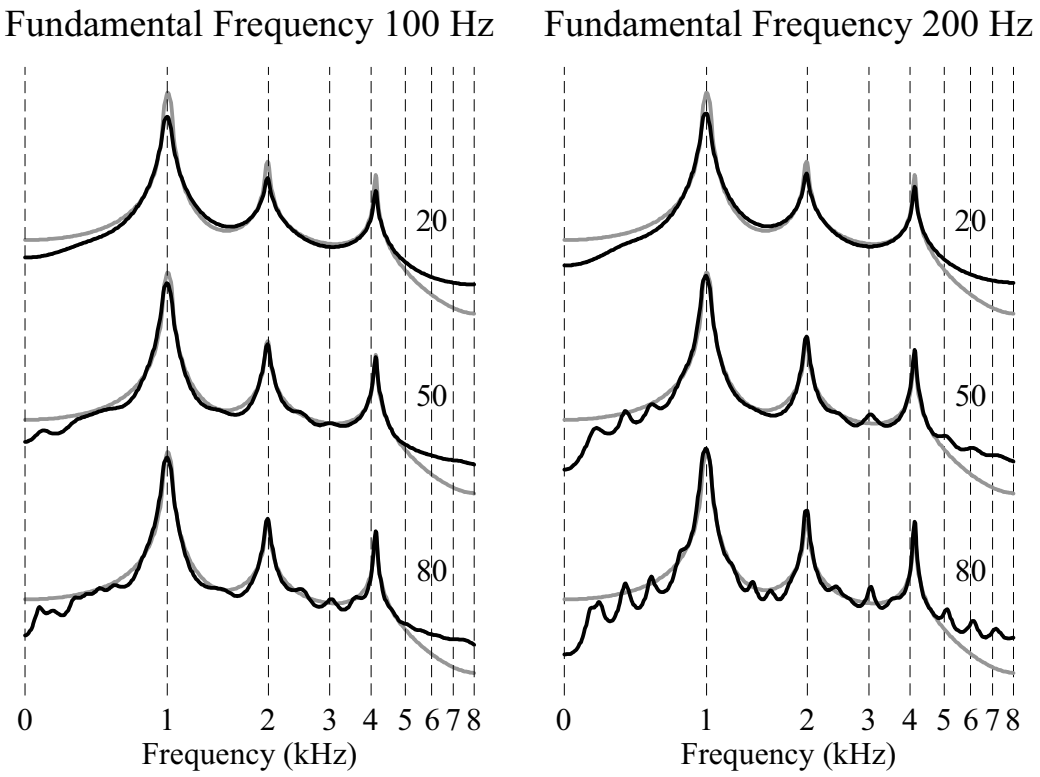


Figure 5.3: Warped MVDR envelopes (black lines) for different model orders (20, 50 and 80) and different fundamental frequencies (100 and 200Hz) in comparison to the spectral envelope (warped MVDR, order 200) of the transfer function $H(z)$ (gray lines).

Due to the previous investigation it becomes obvious that the MO of the spectral envelope has to be adapted to provide optimal features in an ASR front-end. Possible objective functions are:

- **maximum likelihood**
  In order to choose an optimal MO to best fit the acoustic models, we can first calculate cepstral features $\mathbf{c}_m$ corresponding to various MOs m. Let $\mathbf{C}_m$ denote a sequence of cepstral features which have been derived from

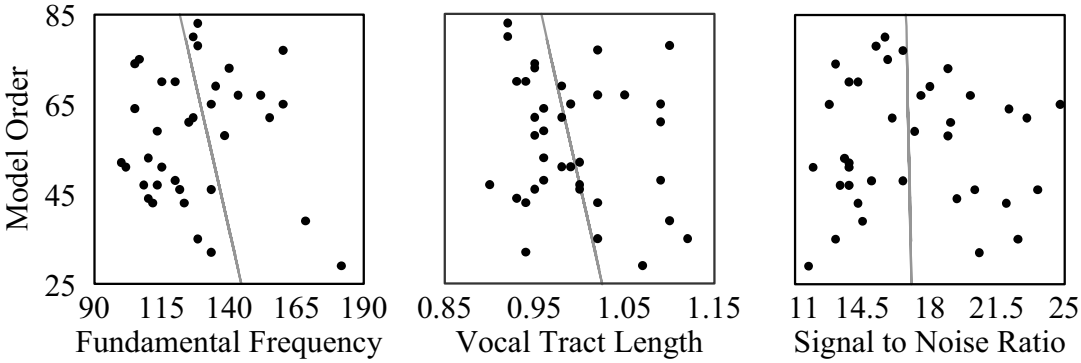


Figure 5.4: The relationship between the model order and the fundamental frequency (left), the vocal tract length (center) and the signal to noise ratio (right) for the 39 speakers of the Translanguage English Database. Each point represents a single speaker and the regression line is plotted in grey.

spectral envelopes of different MOs and let λ denote a set of given hidden Markov models trained on a broad variety of speakers with a *fixed* MO. The optimal MO $\hat{m}$ for the given speaker is then obtained by maximising the likelihood of the adaptation data $\mathbf{C}$ given the corresponding word string W:

$$\hat{m} = \underset{m}{\operatorname{argmax}} \, \mathrm{P}(\mathbf{C}_m | \lambda, W). \tag{5.2}$$

The estimated MO can then be used to train a new acoustic model.

- **class separability**
  In order to optimize the MO in term of class separability we can use the measure of class separability which compares the relationship between the within-class scatter matrix $\tilde{\mathbf{S}}_w$ as defined in (9.1) and between-class scatter matrix $\tilde{\mathbf{S}}_b$ as defined in (9.2)

$$\hat{m} = \underset{\tilde{m}}{\operatorname{argmax}} \operatorname{trace}_d \left\{ \left(\tilde{\mathbf{W}}^T \tilde{\mathbf{S}}_w \tilde{\mathbf{W}} \right)^{-1} \cdot \left(\tilde{\mathbf{W}}^T \tilde{\mathbf{S}}_b \tilde{\mathbf{W}} \right) \right\}$$

  where $\tilde{\mathbf{W}}$ defines the linear discriminant matrix optimized for $\tilde{\mathbf{S}}_w$ and $\tilde{\mathbf{S}}_b$.

To investigate the relationship between the optimal MO and fundamental frequency, vocal tract length, and signal to noise ratio of speech, we ploted scatter matrices and calculated regression lines, Figure 5.4. The fundamental frequency was calculated by the average magnitude difference function [178]. Comparing all 39 speakers in the Translanguage English Database corpus as described in Section 2.1.1, we can see on the left side of Figure 5.4 that the MO shows some functional dependence on the fundamental frequency. This relation also exists for the vocal tract length value shown in the center of Figure 5.4, which does not surprise as the fundamental frequency is also correlated to the vocal tract length. This implies that on average a male speaker with a lower f_0 and a warp factor lower than 1 should have a higher MO than a female speaker with a higher f_0 and a warp factor larger than 1. We could not find any statistically relevant correlation between the MO and the signal to noise ratio as shown on the right side of Figure 5.4. This seems to contradict Tierney [200] who has claimed that corrupted speech has to be modeled using a higher MO of the all-pole model, to model both speech and noise. But as we are only interested in the best prediction of the physical excitation of the vocal tract, we have no interest in modeling the noise and therefore we should not expect an increase in MO.

5.3 Conclusion

In this chapter we have investigated the influence of the fundamental frequency on spectral envelope estimation. We have found that a spectral envelope with a lower model order provides a better estimate for higher fundamental frequency and vice versa. Instead of directly using this property to adjust the model order we have suggested to optimize the model order according to the acoustic likelihood of the speech recognition system as our main interest focuses on a decrease in recognition error.

Chapter 6

Compensation of Non-Stationary Additive Distortion by Particle Filters

In the previous sections we have discussed speech feature extraction as well as feature adaptation. In this and the following two sections we want to draw our attention to speech feature enhancement. While the term *speech enhancement* includes various topics such as background noise reduction, dereverberation, blind source separation, beamforming, reconstruction of lost speech packets in digital networks or bandwith extention of narrowband speech, we want to distinguish between speech enhancement and *speech feature enhancement*. The latter term is used to describe algorithms or devices to improve the speech features, where a single contaminated waveform or single contaminated feature stream is available, with the goal to get higher classification accuracy. Note that an increase in classification accuracy might not necessary result in an improved or pleasing sound quality (if the reconstruction is at all possible).

We have learned in Section 1.3 that non-stationary additive noise and reverberation are the most severe and frequently encountered distortions in hands-free speech recordings. Thus additive noise and reverberation reduction are the most important methods to decrease the word error rate in distant speech recognition. Thus we limit our investigations to these problems and leave out other

sources of degradation in the speech signal such as coloration caused by head orientation or room modes.

If only the noisy waveform or noisy speech feature alone is available, enhancement to improve audio perception or speech recognition performance, has been and still is an outstanding and difficult problem in speech processing. In the case of speech recognition, modification can be applied either in the time domain, on the spectral features in the magnitude, power or logarithmic domain or on the cepstral features without the need to recreate the time signal. However, usually the enhancement takes place in the linear or logarithmic spectral domain. The main drawback of such methods, e.g. spectral subtraction, is that part of the noise remaining after processing has a very unnatural quality [66, 75]. This can be explained by the fact that the magnitude of the short-time power spectrum exhibits strong fluctuations in noisy areas. After spectral attenuation the frequency bands, which originally contained the noise, consist of randomly spaced spectral peaks corresponding to the maxima of the short-time power spectrum. Between these peaks, the short-time power spectrum values are close to or below the estimated averaged noise spectrum, which results in strong attenuations. As a result, the residual noise is composed of sinusoidal components with random frequencies that come and go in each short-time frame [66]. These artifacts are known as *musical*[1] *tones/noise* phenomenon. One way to reduce this unwanted effect is to median smooth the signal after spectral subtraction. Unfortunately, this leads to audible signal distortions [139]. To overcome this problem, we have proposed the use of spectral envelopes instead of smoothing [30].

To cope well with the non-stationary behavior of additive distortions various approaches have been suggested such as the interacting multiple model [127]. In the last couple of years various *particle filter* (PF) approaches have been proposed to track non-stationary additive distortions on speech features in the logarithmic power frequency domain [209, 187, 97]. The ability to compensate for non-stationary noise is, for example, highlighted in [187] where the PF approach, which serves as a baseline in our investigations, is compared with the vector Taylor series approach [148]. For different noise types, artificially added with different signal to noise ratios, the PF approach leads to significant lower word error rates.

To our knowledge Yao and Nakamura [209] were the first who proposed speech feature enhancement by particle filtering for speech recognition. Additional interesting work in this context has been published by Singh and Raj [187]. In their approach, they use a PF to track the noise sequence corrupting the speech signal. This estimated noise sequence is then used to clean or enhance the speech features. The two critical aspects in PF design are the choice of the

[1] This term is a reference to the presence of pure tones in the residual noise.

importance or proposal density and the particle weight calculation. A number of PF variants have been evaluated for the enhancement of speech features: auxillary and likelihood PFs [111] as well as PFs with an extended Kalman filter proposal density [97].

In Section 6.1 we give a brief overview of model based speech feature enhancement techniques. In Section 6.2 we review Bayesian speech feature enhancement, introduce the noise as a hidden variable, avoid intractable integration by Monte Carlo methods and give a general overview of speech feature enhancement by PFs. Section 6.3 reviews how to evaluate for the weights of the different noise samples. It also introduces phoneme dependent speech models as the speech dynamics can not be correctly represented by a single Gaussian mixture model. In the case of a phoneme dependent model, the phoneme sequence can be derived by a forced alignment from a previous speech recognition pass. Another important step in Bayesian filtering will be covered in Section 6.4. Previous PF methods have relied either on a random walk or on a predicted walk using a prior knowledge. To overcome the usage of a prior knowledge we propose to integrate the estimation of the predicted walk model within the PF framework. The presented feature enhancement framework tracks the noise instead of the clean speech signal. Thus an additional processing step is required which maps the noisy observation, given the noise estimate, to a clean speech estimate. Section 6.5 reviews a method which approximates the non-linearity by a vector Taylor series. With the observation that the probability density function is modeled by point observations in the applied framework, it becomes obvious that an approximation by the vector Taylor series is not needed.

6.1 Speech Feature Enhancement Techniques Based on Probabilistic Models

Speech feature enhancement methods attempt to map[2]

$$\hat{\mathbf{x}}_k = f(\mathbf{y}_k)$$

the noisy feature $\mathbf{y}_k$ to a clean feature estimate $\hat{\mathbf{x}}_k$. A broad family of mapping approaches apply a transformation based on a probabilistic model of the distortion between clean speech and noisy speech which has to be learned from a set of *stereo data*[3]. One prominent method of this kind is *stereo-based piecewise linear compensation for environments* (SPLICE) which is an extension to the

[2] In the work by Westphal and Waibel this mapping is referred to as acoustic transformation.

[3] We refer to stereo data as two time aligned channels, one providing distortion free observations while the other is a distorted observation of exactly the same source.

fixed codeword-dependent cepstral normalization (FCDCN) algorithm [56] which itself is a successor of *codeword-dependent cepstral normalization* (CDCN) [55]. The original version of SPLICE as proposed by Deng *et al.* [87] assumes that the noisy speech vector $\mathbf{y}_k$ lies in one of several partitions of the acoustic space. These partitions are determined from a mixture of M Gaussians. The mean and variances of the correction $\mathbf{r}$ are trained by vectors which have been classified into corresponding codewords. Furthermore, the SPLICE algorithm assumes that the relation between $\mathbf{x}_k$ and $\mathbf{y}_k$ is piecewise linear, according to

$$\mathbf{x}_k = \mathbf{y}_k + \mathbf{r}\left(\mathbf{y}_k\right) \approx \mathbf{y}_k + \mathbf{r}_{m(\mathbf{y}_k)},$$

where $m(\mathbf{y}_k)$ determines which part of the local linear approximation is used. Under these assumptions the clean speech features $\hat{\mathbf{x}}_k$, for frame k, can be calculated under the *minimum mean square error* (MMSE) criterion, which consists in finding the conditional mean, as follows:

$$\begin{aligned}
\hat{\mathbf{x}}_k &= \mathrm{E}\{\mathbf{x}_k|\mathbf{y}_k\} = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k \approx \int \left(\mathbf{y}_k + \mathbf{r}_{m(\mathbf{y}_k)}\right) p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k \\
&= \mathbf{y}_k + \int \mathbf{r}_{m(\mathbf{y}_k)} p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k = \mathbf{y}_k + \int \sum_{m=1}^{M} \mathbf{r}_m p(\mathbf{x}_k, m|\mathbf{y}_k) d\mathbf{x}_k \\
&= \mathbf{y}_k + \sum_{m=1}^{M} \int \mathbf{r}_m p(\mathbf{x}_k, m|\mathbf{y}_k) d\mathbf{x}_k = \mathbf{y}_k + \sum_{m=1}^{M} p(m|\mathbf{y}_k) \mathbf{r}_m.
\end{aligned}$$

The posterior probabilities $p(m|\mathbf{y}_k)$ are computed by Bayes' rule using the clustered parameters in the *Gaussian mixture model* (GMM) approximation of $p(\mathbf{y})$.

The major drawback of the earliest versions of SPLICE was their dependency on stereo data in order to calculate the estimate. Two extensions to the original approach have been proposed to overcome those limitations, one using a maximum likelihood criterion [206] and one using discriminative training by minimum classification error [207]. Deng *et al.*[89] report that the latter method is very similar to the feature space minimum phone error algorithm [176].

In order to overcome the need for stereo data Westphal and Waibel [203] have suggested to simulate the distortion caused by additive noise by the combination of a cleaned speech model with a noise model, represented as a single Gaussian derived on noise only frames. Therefore the distribution of the clean speech model and the distribution of the noisy speech model have a one-to-one correspondence for each Gaussian which can be expressed in the difference $\Delta_{\boldsymbol{\mu}_m} = \boldsymbol{\mu}_{y_m} - \boldsymbol{\mu}_{x_m}$ where the index m determines the Gaussian within the GMMs. Their approach, dubbed *model-combination-based acoustic mapping* (MAM), then follows the steps as suggested by Moreno *et al.* [147]. The MMSE

solution is then given, similar to SPLICE, by

$$\hat{\mathbf{x}}_k = \mathrm{E}\{\mathbf{x}_k|\mathbf{y}_k\} = \mathbf{y}_k - \int \Delta_{\mathbf{x}} p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k \approx \mathbf{y}_k - \sum_{m=1}^{M} p(m|\mathbf{y}_k)\Delta_{\boldsymbol{\mu}_m}.$$

To account for the non-linear relationship between y, x and n Moreno *et al.* [148] have suggested to use a *vector Taylor series* (VTS) around the mean values $\boldsymbol{\mu}_{x_m}$ of each Gaussian m within the GMM.

$$\begin{aligned}\hat{\mathbf{x}}_k &= \mathrm{E}\{\mathbf{x}_k|\mathbf{y}_k\} = \int (\mathbf{y}_k - f(\mathbf{x}_k, \mathbf{n}_k)) p(\mathbf{x}_k|\mathbf{y}_k) d\mathbf{x}_k \\ &\approx \mathbf{y}_k - \sum_{m=1}^{M} p(m|\mathbf{y}_k) f(\boldsymbol{\mu}_{x_m}, \boldsymbol{\mu}_{n_k}).\end{aligned}$$

The VTS can now readily applied, to account for the non-linearity, within the MAM framework as

$$\hat{\mathbf{x}}_k = \mathrm{E}\{\mathbf{x}_k|\mathbf{y}_k\} \approx \mathbf{y}_k + \sum_{m=1}^{M} p(m|\mathbf{y}_k) \log(1 - e^{(\Delta_{\boldsymbol{\mu}_m} - \mathbf{y}_k)}).$$

6.2 A Bayesian Approach to Compensate for Non-Stationay Additive Distortion

Speech feature enhancement which compensates for non-stationary distortions can be formulated as a tracking problem where the clean speech features $\mathbf{x}_k$ have to be estimated for each frame k, given the current observation and its history of the noisy features $\mathbf{y}_{1:k}$. A general description of such a system that relates two stochastic processes, namely the state $(X_k)_{k\in\mathbb{N}}$ representing the evolution of a hidden, inner system and the corresponding observation or measurement $(Y_k)_{k\in\mathbb{N}}$, is given by a statespace model consisting of two equations. In their most general (discrete) form these are

- the *state equation*
$$\mathbf{x}_k = f(\mathbf{x}_{k-1}, \mathbf{u}_{k-1}) \tag{6.1}$$
- and the *observation equation*
$$\mathbf{y}_k = g(\mathbf{x}_k, \mathbf{w}_k) \tag{6.2}$$

where f represents the non-linear transition function, g the non-linear observation function, $\mathbf{x}_k$ the state vector, $\mathbf{y}_k$ the observation vector, $\mathbf{u}_k$ the process noise and $\mathbf{w}_k$ the measurement noise. The state equation characterizes the *state transition probability* $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ while the observation equation describes the probability $p(\mathbf{y}_k|\mathbf{x}_k)$ which is coupled to the measurement noise model.

The MMSE solution to a tracking problem, which relates $\mathbf{x}$ and $\mathbf{y}$ by the probabilistic relationship $p(\mathbf{x}_k|\mathbf{y}_{1:k})$, consists in finding the conditional mean $\mathrm{E}\{\mathbf{x}_{1:k}|\mathbf{y}_{1:k}\}$. Assuming that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a Markov process and that the current observation is only dependent on the current state facilitates sequential calculation of the conditional mean. With the previous assumptions the MMSE solution is given by

$$\mathrm{E}\{\mathbf{x}_k|\mathbf{y}_{1:k}\} = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}) d\mathbf{x}. \tag{6.3}$$

6.2.1 Tracking Additive Distortion

To track the additive distortions instead of the speech signal directly we have to introduce the noise $\mathbf{n}_k$ as a hidden variable

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \int p(\mathbf{x}_k, \mathbf{n}_k|\mathbf{y}_{1:k}) d\mathbf{n}_k. \tag{6.4}$$

Given the relation $p(\mathbf{x}_k, \mathbf{n}_k|\mathbf{y}_{1:k}) = p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k)p(\mathbf{n}_k|\mathbf{y}_{1:k})$ and changing the integration order in (6.4) we obtain

$$\mathrm{E}\{\mathbf{x}_k|\mathbf{y}_{1:k}\} = \int v_k(\mathbf{y}_{1:k}, \mathbf{n}_k) p(\mathbf{n}_k|\mathbf{y}_{1:k}) d\mathbf{n}_k \tag{6.5}$$

where the function

$$v_k(\mathbf{y}_{1:k}, \mathbf{n}_k) = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) d\mathbf{x}_k \tag{6.6}$$

maps the noisy observation sequence $\mathbf{y}_{1:k}$ and noise estimate $\hat{\mathbf{n}}_k$ to the clean speech estimate $\hat{\mathbf{x}}_k$. Note that due to the non-linear relationship between $\mathbf{n}$, $\mathbf{y}$ and $\mathbf{x}$ in the chosen working domain, $v_k(\mathbf{y}_{1:k}, \mathbf{n}_k)$ is also non-linear. How to solve for $v_k(\mathbf{y}_{1:k}, \mathbf{n}_k)$ is described in Section 6.5 where a common method is reviewed and a novel method, which employs the fact that the noise probability density function is represented as a bunch of point estimates, is proposed.

The *filtering density* $p(\mathbf{n}_k|\mathbf{y}_{1:k})$ in (6.5) keeps track of the probability throughout time. It can be realized by sequential updating

$$\begin{aligned} p(\mathbf{n}_k|\mathbf{y}_{1:k}) &= p(\mathbf{n}_k|\mathbf{y}_k, \mathbf{y}_{1:k-1}) \\ &= \frac{p(\mathbf{y}_k|\mathbf{n}_k, \mathbf{y}_{1:k-1})p(\mathbf{n}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})} \\ &= \frac{p(\mathbf{y}_k|\mathbf{n}_k)p(\mathbf{n}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})} \quad (6.7) \\ &= \frac{p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{1:k-1})}. \quad (6.8) \end{aligned}$$

The nominator in (6.7) is composed of the likelihood function $p(\mathbf{y}_k|\mathbf{n}_k)$ and $p(\mathbf{n}_k|\mathbf{y}_{1:k-1})$ which can be rewritten by the Chapman-Kolmogorov equation [173] as

$$p(\mathbf{n}_k|\mathbf{y}_{1:k-1}) = \int p(\mathbf{n}_k|\mathbf{n}_{k-1})p(\mathbf{n}_{k-1}|\mathbf{y}_{1:k-1})d\mathbf{n}_{k-1}. \quad (6.9)$$

The normalization term can be solved by

$$\begin{aligned} p(\mathbf{y}_k|\mathbf{y}_{1:k-1}) &= \int p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1})d\mathbf{n}_k \quad (6.10) \\ &= \int p(\mathbf{n}_k|\mathbf{y}_{1:k-1})p(\mathbf{y}_k|\mathbf{n}_k)d\mathbf{n}_k. \quad (6.11) \end{aligned}$$

To solve for (6.9) requires the prediction of the current noise estimate $\mathbf{n}_k$ given the previous estimate $\mathbf{n}_{k-1}$ by the noise transition probability $p(\mathbf{n}_k|\mathbf{n}_{k-1})$. Section 6.4 reviews the random walk and the predicted walk by static autoregressive processes. It also proposes a method to integrate the estimation of the prediction matrix within the PF framework.

6.2.2 Monte Carlo Sampling

To avoid intractable integration, which can only be solved for some special cases of linearity and Gaussianity, we aim to approximate the *posterior filtering density* by a weighted approximation as

$$p(\mathbf{n}_k|\mathbf{y}_{1:k-1}) \approx \sum_{s=1}^{S} w_k^{(s)} \delta(\mathbf{n}_k - \mathbf{n}_k^{(s)})$$

where w represents the weights and s the samples.

As drawing the samples directly from the posterior density $p(\mathbf{n}_k|\mathbf{y}_{1:k-1})$ is often infeasible, a suboptimal importance density is frequently chosen [181].

Decomposing $p(\mathbf{n}_k|\mathbf{y}_{1:k})$ as in (6.8) let us now express the *empirical density* by the two approximations

$$p(\mathbf{n}_k, \mathbf{y}_k|\mathbf{y}_{1:k-1}) \approx \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{n}_k|\mathbf{n}_{k-1}^{(s)})p(\mathbf{y}_k|\mathbf{n}_k^{(s)}) \tag{6.12}$$

and

$$p(\mathbf{y}_k|\mathbf{y}_{1:k-1}) \approx \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}_k|\mathbf{n}_k^{(s)}) \tag{6.13}$$

where $p(\mathbf{y}_k|\mathbf{n}_k^{(s)})$ represents the corresponding likelihood for each sample s out of S samples. How to solve for $p(\mathbf{y}_k|\mathbf{n}_k^{(s)})$ will be explained in Section 6.3.

Those samples are probably better known as *particles* and the filter process is called particle filtering respectively. For a detailed introduction into particle filtering see for example [181]. After evaluation each particle represents a weighted distortion estimate, where each dimension of the particle may be associated for example with a distortion energy at a particular frequency bin or a scale term of a given distortion estimate. More details follow in later sections.

6.2.3 A General Particle Filter Framework to Compensate for Non-Stationary Additive Distortions

A variety of different PF variants have been proposed and evaluated for the enhancement of speech features: auxiliary and likelihood PFs [111] as well as PFs with an extended Kalman filter proposal density [97] or the use of static [187] or dynamic [37] autoregressive matrices. All approaches, however, are similar in structure and can be decomposed into a number of successive steps as depicted in Figure 6.1 with corresponding description in Algorithm 6.1.

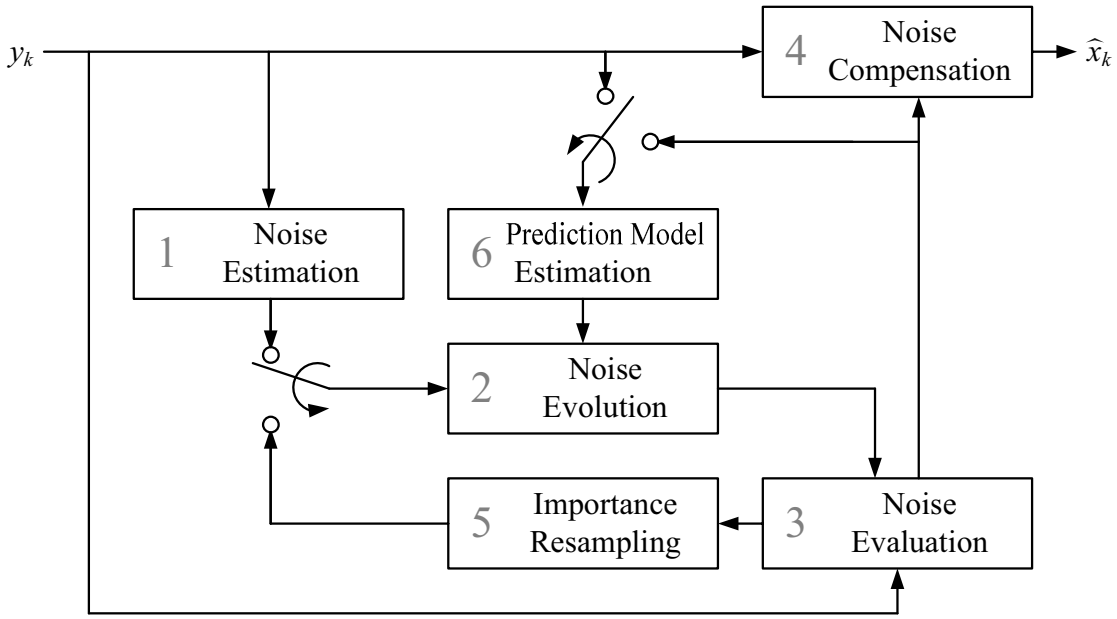


Figure 6.1: General flowchart of frame based speech feature enhancement for non-stationary additive distortion using a particle filter with importance resampling. The individual steps, gray numbers, are summarized in Algorithm 6.1.

1. *Noise Estimation & Particle Initialization*
   Estimate the prior noise density

$$p(\mathbf{n}_0) \approx \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}_m, \mathbf{\Sigma}_m)$$

   as a Gaussian mixture model trained on silence frames detected by voice activity detection.
   Noise samples $\mathbf{n}_0^{(s)},\ s = 1, ..., S$ are drawn from the prior noise density $p(\mathbf{n}_0)$.
2. *Noise Evolution (Sampling)*
   New particles $\mathbf{n}_k^{(s)}$ are sampled from the noise transition probability $p(\mathbf{n}_k|\bar{\mathbf{n}}_{k-1}^{(s)})$. The details of sampling will be laid out in Section 6.4.
3. *Noise Evaluation*
   The different noise hypotheses are evaluated and importance weights are assigned. The details of noise evaluation will be laid out in Section 6.3.
4. *Noise Compensation*
   The cleaned speech feature is calculated given the weighted noise samples. Two alternative approaches are described in Section 6.5.
5. *Importance Resampling*
   Possibly the normalized weights are used to resample among the noise particles $\mathbf{n}_k^{(s)},\ s = 1, ..., S$.
6. *Prediction Model Estimation*
   Possibly the noise transition probability model has to be updated or estimated; e.g. for dynamic autoregressive models.

Steps 2 until 6 are repeated with $k \mapsto (k+1)$ until all time-frames are processed or until the particle filter has to be reinitialized with step 1.

Algorithm 6.1: General framework of frame based speech feature enhancement using a particle filter with importance resampling.

6.3 Evaluation of Noise Samples

To solve for (6.12) and (6.13) requires the evaluation of the samples in the PF by the likelihood function which depends on the task and working domain. PFs for speech feature enhancements traditionally work in a dimension reduced, mel-scaled logarithmic power frequency domain. The dimension reduction is required as PFs are not capable to work in a high dimensional space or, at least, are very slow. The non-linear frequency scale, representing the non-linear frequency resolution of human hearing, is chosen because it has been profen advantageously over a linear frequency scale in various speech applications. In order to provide increased robustness already in the feature extraction process we decided to replace the Fourier transformation followed by a mel-filterbank with the warped *minimum variance distortionless response* (MVDR) spectral estimate, see Section 3.4.6. As the warped MVDR already provides non-linear frequency mapping and smoothing, no filterbank—which commonly reduces the number of bins—is used in the front-end. In order to reduce the dimension of the logarithmic spectral domain (in our case 129 bins) we first truncated the cepstral sequence to 20 dimensions and applied an inverse discrete cosine transformation, established by a simple 20×20 matrix multiplication, to derive 20 logarithmic spectral coefficients. The relation between the noisy observation $\mathbf{y}$, the clean feature $\mathbf{x}$ and noise $\mathbf{n}$ can be approximated by

$$\mathbf{x} = \mathbf{y} + \ln(\mathbf{1} - e^{\mathbf{n}-\mathbf{y}}) + \mathbf{e}_\theta + \mathbf{e}_{\text{envelope}} \approx \ln(e^{\mathbf{y}} - e^{\mathbf{n}}). \tag{6.14}$$

Note that the dimension of $\mathbf{n} = u(\mathbf{n}) = u(\mathbf{a}, \mathbf{s})$ has to be identical to $\mathbf{y}$ and $\mathbf{x}$ while $\mathbf{n}$ can have an arbitrary dimension which is mapped by the function u to $\mathbf{n}$.

The first error term

$$\mathbf{e}_\theta = \ln\left(1 + \frac{2\cos\theta(\Omega)}{\cosh\left\{\ln|N(\Omega)| - \ln|X(\Omega)|\right\}}\right)$$

introduced by the approximation in (6.14) is complicated to evaluate. Deng *et al.* [88] have, however, empirically verified that the average value of $\mathbf{e}$ is close to zero and that $\theta(\Omega)$ is Gaussian distributed. Note that this is true in particular for higher frequencies (mel-scale) as the central limit theorem can only be applied in those regions (low regions are combined of only a very little number of bins).

In the case of spectral or cepstral envelope techniques a second error term $\mathbf{e}_{\text{envelope}}$ is introduced and thus the relation of (6.14) is further weakened. The approximation in (6.14) however is still sufficient for our investigations.

6.3.1 Weight Calculation for Each Sample

With the approximation in (6.14) it is now possible to evaluate each sample $\mathbf{n}_k^{(s)} = u(\mathbf{n}_k^{(s)})$ according to the likelihood function

$$p(\mathbf{y}_k|\mathbf{n}_k^{(s)}) = \frac{p_{\text{speech}}(\mathbf{y}_k + \ln(\mathbf{1} - e^{\mathbf{n}_k^{(s)} - \mathbf{y}_k}))}{\prod_{b=1}^{B} 1 - e^{n_{k,b}^{(s)} - y_{k,b}}} \tag{6.15}$$

where $p_{\text{speech}}(\cdot)$ denotes the prior speech density represented by a GMM which has been trained on clean speech. The model is represented in a B dimensional space where each dimension b represents a frequency bin (non-equally scaled according to the mel-frequency). Thereafter, to get the normalized weights $\tilde{w}_k^{(s)}$, the likelihoods have to be divided by the sum over all likelihoods

$$\tilde{w}_k^{(s)} = \frac{p(\mathbf{y}_k|\mathbf{n}_k^{(s)})}{\sum_{m=1}^{S} p(\mathbf{y}_k|\mathbf{n}_k^{(m)})}. \tag{6.16}$$

Substantial overestimations of the actual noise lead to severe problems with the likelihood computations as the likelihood can only be evaluated if

$$n_{k,b}^{(s)} < y_{k,b} \; \forall \; b \; \in \; B.$$

This is an artifact of treating speech and noise as strictly additive. The physical reason behind this is that energy must always be a positive quantity. If this constraint is not satisfied, $p(\mathbf{y}_k|\mathbf{n}_k^{(s)})$ can not be evaluated and thus has to be rejected by setting the particle weight to zero. This causes a decimation of the particle population which results in a complete annihilation if all particle samples are rejected. Häb-Umbach *et al.* [111] have reported, that noise overestimation might lead to a severe decimation of the particle population, or even to its complete annihilation.

To overcome the problem associated with setting the particle weight to zero we have proposed the *fast acceptance test* [5] that virtually boosts the number of particles by sampling a new noise hypothesis from $p(\mathbf{n}_k|\mathbf{n}_{k-1}^{(i)})$ (i is randomly drawn), if $\mathbf{n}_k^{(s)}$ could not be evaluated (rejected). This can be repeated until $\mathbf{n}_k^{(s)}$ is accepted or a certain number B of iterations has passed. Thus dropouts can still occur, however less often.

6.3.2 Coupling Distortion Evaluation with Automatic Speech Recognition

So far, as well as in previous works by other authors, a general and thus static speech model, p_{speech}, has been used. It systematically ignores the dynamic properties of speech. To overcome this deficit we propose to use a phoneme-specific speech model where each phoneme is represented by a GMM. Since the phoneme sequence is not known in advance, we propose to use a *two-pass* PF as depicted in Figure 6.2. In the first pass the PF uses the general, static speech model to clean the speech spectra, and a first transcription is obtained by processing those features with the speech recognition system. In the second (and following) pass(es), the phoneme sequence is estimated and the phoneme-specific speech model can be used. The beauty of this approach is that the sophisticated acoustic and language models of the speech recognizer are incorporated into the particle filtering stage. Unfortunately, the phoneme-specific scoring function introduces two new problems into the PF:

1. Switching between the phonemes causes a very sudden change of the particles' (noise hypotheses') likelihoods which can destabilize the PF.
2. By correcting all corrupted speech spectra toward the hypothesis from the previous recognition pass we might tie ourselves to that hypothesis.

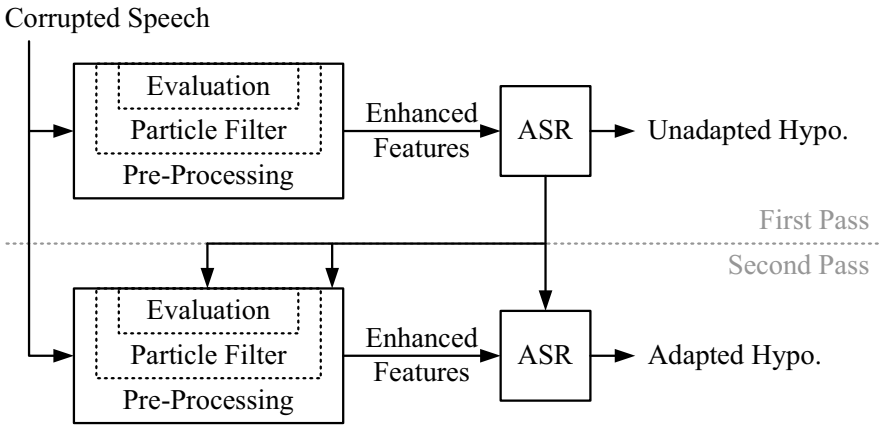


Figure 6.2: Flowchart of the coupling between the distortion evaluation process within the particle filter and the *automatic speech recognition* (ASR) engine.

To overcome those problems we loosen the strength of a phoneme specific model by interpolation with the general model to form the *mixture model*

$$\hat{p}_{mix(t)}(x) = \gamma \cdot \hat{p}_{\text{phon}(t)}(x) + (1-\gamma) \cdot p(x)$$

where γ denotes the mixture weight.

6.4 Prediction of Samples

From (6.9) it directly follows that tracking requires the prediction of the noise $\mathbf{n}_k$ given the previous noise estimate $\mathbf{n}_{k-1}$ and a model representing the distortion transition probability $p(\mathbf{n}_k|\mathbf{n}_{k-1})$ which can be represented in various ways.

In this section we review the random walk and the predicted walk calculated by a static autoregressive process. The *static* autoregressive process has the drawback that the prediction matrix has to be calculated on noise only regions either before the application of the PF or on silence regions within the utterances. To overcome this drawback we have proposed a *dynamic* autoregressive process which is able to calculate the prediction matrix within the PF framework as described in Section 6.4.3.

6.4.1 Random Walk

The simplest way to model the evolution of noise features is a random walk

$$\mathbf{n}_k = \mathbf{n}_{k-1} + \boldsymbol{\epsilon}_k \tag{6.17}$$

where $\mathbf{n}_k$ denotes the noise spectrum at time k while the random term $\boldsymbol{\epsilon}_k \sim \mathcal{N}(0, \boldsymbol{\Sigma}^{\mathbf{random}})$ is considered to be i.i.d. zero mean Gaussian with diagonal covariance matrix.

6.4.2 Predicted Walk by Static Autoregressive Processes

To consider information about the evolution of the noise, Raj *et al.* [179] proposed to use a higher-order autoregressive process $\mathbf{A}^{(1:L)}$, where L denotes the order, to predict the evolution of the noise

$$\begin{aligned} \mathbf{a}_k &= \mathbf{A}^{(1)}\mathbf{a}_{k-1} + \mathbf{A}^{(2)}\mathbf{a}_{k-2} + \cdots + \mathbf{A}^{(L)}\mathbf{a}_{k-m} + \boldsymbol{\epsilon}_k \\ &= \mathbf{A}^{(1:L)}\mathbf{a}_{k-1:k-L} + \boldsymbol{\epsilon}_k. \end{aligned} \tag{6.18}$$

The autoregressive noise model consists of two components that have to be learned:

- the *linear prediction matrix* $\mathbf{A}^{(1:L)}$ and
- the *covariance matrix* $\mathbf{\Sigma}^{\mathrm{AR}}$.

Minimization of the squared prediction error results in the following estimate of the linear prediction matrix

$$\mathbf{A}^{(1:L)} = \mathcal{E}\{\mathbf{n}_k \mathbf{n}_{k-1:k-L}^T\} \mathcal{E}\{\mathbf{n}_{k-1:k-L} \mathbf{n}_{k-1:k-L}^T\}^{-1}. \tag{6.19}$$

Those matrices can be derived from the noise data $1, 2, \ldots, K$ as

$$\mathcal{E}\{\mathbf{n}_k \mathbf{n}_{k-1:k-L}^T\} \;=\; \frac{1}{K} \sum_{k=l}^{K} \mathbf{n}_k \mathbf{n}_{k-1:k-L}^T$$

and

$$\mathcal{E}\{\mathbf{n}_{k-1:k-L} \mathbf{n}_{k-1:k-L}^T\} \;=\; \frac{1}{K} \sum_{k=l}^{K} \mathbf{n}_{k-1:k-L} \mathbf{n}_{k-1:k-L}^T.$$

To learn a linear prediction matrix of model order length L on B spectral bins requires $B^2 L$ coefficients. A reliable estimate is only possible on a huge amount of training data which, fortunately, can be composed of noise pieces as long as they contain enough history. For higher model orders, however, only a small reduction in the mean square error of the prediction is possible which is apparent from Figure 6.3. Thus, a first model order is sufficient for our investigations.

The static sample covariance matrix can be calculated by

$$\mathbf{\Sigma}^{\mathrm{AR}} = \frac{1}{K} \sum_{k=1}^{K} (\mathbf{a}_k - \mathbf{A}^{(1:L)} \mathbf{a}_{k-1:k-L})(\mathbf{a}_k - \mathbf{A}^{(1:L)} \mathbf{a}_{k-1:k-L})^T$$

where K denotes the number of frames.

6.4.3 Predicted Walk by Dynamic Autoregressive Processes

In the previous section we have used a linear prediction matrix which has been derived previous to its application. This approach has two obvious drawbacks:

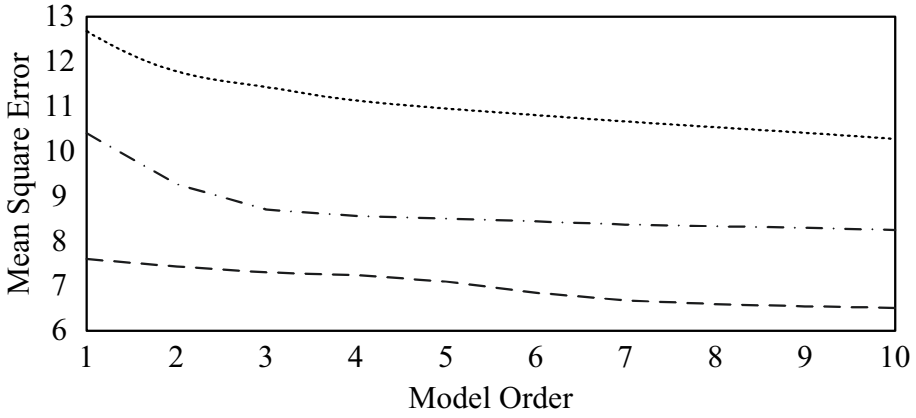


Figure 6.3: Mean square error of the predicted noise evolution for different noise types (static dashed line, semi-static dashed with points and dynamic pointed line) and model order.

- One has to know the noise a priori, or rely on voice activity detection.
- The prediction matrix can not adjust to different types of distortion in those regions where speech is present.

To overcome the drawbacks apparent in static autoregressive processes a *dynamic*, and thus instantaneous and integrated estimate of the linear prediction matrix

$$\mathbf{A}_k = \mathbf{A}_k^{(1)} = \mathcal{E}\{\mathbf{n}_k \mathbf{n}_{k-1}^T\} \mathcal{E}\{\mathbf{n}_{k-1} \mathbf{n}_{k-1}^T\}^{-1} \tag{6.20}$$

is required for each individual frame k.

In a framework where the likelihood of the noise can be evaluated and a number of samples can be drawn, e.g. in the application of PFs, it becomes possible to estimate the two matrices $\mathbf{n}_k \mathbf{n}_{k-1}^T$ and $\mathbf{n}_{k-1} \mathbf{n}_{k-1}^T$ on the current $\mathbf{n}_k^{(s)}$ and previous $\mathbf{n}_{k-1}^{(s)}$ noise estimates for all samples $s = 1, 2, \ldots,$S [37]. To ensure that the prediction estimates which lead to a good noise estimate are emphasized and those predictions which lead to a poor estimate are suppressed, we have to weight the contribution of each noise estimate to the matrices due to their likelihood $p(\mathbf{y}_k|\mathbf{n}_k^{(s)})$ as described in Section 6.3. Thus, the matrices can be evaluated for each frame k by using

$$\mathcal{E}\{\mathbf{n}_k \mathbf{n}_{k-1}^T\} \;=\; \frac{1}{S} \sum_{s=1}^{S} w_k^{(s)} \mathbf{n}_k^{(s)} \; \mathbf{n}_{k-1}^{(s)}{}^T$$

and

$$\mathcal{E}\{\mathbf{n}_{k-1}\mathbf{n}_{k-1}^T\} \;=\; \frac{1}{S}\sum_{s=1}^{S} w^{(s)}\mathbf{n}_{k-1}^{(s)}\;\mathbf{n}_{k-1}^{(s)\;T}$$

to solve for (6.20). The weight of the different samples can be determined for example by

- the likelihood of the current observation

$$w_k^{(s)} = p(\mathbf{y}_k|\mathbf{n}_k^{(s)})$$

- or the likelihood of the previous and current observation

$$w_k^{(s)} = p(\mathbf{y}_{k-1}|\mathbf{n}_{k-1}^{(s)})p(\mathbf{y}_k|\mathbf{n}_k^{(s)})$$

  or

$$w_k^{(s)} = \sqrt{p(\mathbf{y}_{k-1}|\mathbf{n}_{k-1}^{(s)})p(\mathbf{y}_k|\mathbf{n}_k^{(s)})}.$$

A smoothing over previous frames might help to improve the reliability of the estimate. With the introduction of the forgetting factor ξ we can write the smoothed matrix $\overline{\mathbf{A}}_k$ with

$$\overline{\mathcal{E}}[\mathbf{n}_k\mathbf{n}_{k-1}^T] \;=\; \xi\mathcal{E}\{\mathbf{n}_k\mathbf{n}_{k-1}^T\} + (1-\xi)\overline{\mathcal{E}}[\mathbf{n}_{k-1}\mathbf{n}_{k-2}^T]$$

and

$$\overline{\mathcal{E}}[\mathbf{n}_{k-1}\mathbf{n}_{k-1}^T] \;=\; \xi\mathcal{E}\{\mathbf{n}_{k-1}\mathbf{n}_{k-1}^T\} + (1-\xi)\overline{\mathcal{E}}[\mathbf{n}_{k-2}\mathbf{n}_{k-2}^T].$$

The *sample variance* can now be calculated according to the normalized weight $w_k^{(s)}$, the likelihood of the investigated particle m divided by the summation over all likelihoods, as

$$\mathbf{\Sigma}_{\Delta\mathbf{n}} = \sum_{s=1}^{S} w_k^{(s)}(\mathbf{n}_k^{(s)} - \mathbf{A}_{k-1}\mathbf{n}_{k-1}^{(s)})(\mathbf{n}_k^{(s)} - \mathbf{A}_{k-1}\mathbf{n}_{k-1}^{(s)})^T \tag{6.21}$$

or with $\overline{\mathbf{A}}_k$ respectively.

The noise can now be predicted by

$$\mathbf{n}_k = \mathbf{A}_{k-1}\mathbf{n}_{k-1} + \boldsymbol{\epsilon}_k.$$

6.5 Noise Compensation

Given the noisy observation $\mathbf{y}$ and knowledge of the noise $\mathbf{n}$ requires a mapping function $\mathbf{x}_k = v(\mathbf{y}_k, \mathbf{n}_k)$ to derive enhanced or cleaned speech features $\mathbf{x}$. To solve for the underlying non-linear relation $\mathbf{y} \approx \ln(\mathbf{1} + e^{\mathbf{n}-\mathbf{x}})$—which directly follows from (6.14)—it has been suggested by Moreno *et al.* [148] to use an approximation by a truncated vector Taylor series as will be discussed in Section 6.5.1. A vector Taylor series is however not required in the Monte Carlo framework as the empirical density is modeled by individual samples in which the non-linear relationship can be applied directly as proposed in Section 6.5.3.

6.5.1 The Vector Taylor Series Approach

To solve for above mentioned non-linear relations Moreno *et al.* [148] proposed to use a VTS expansion around the oth Gaussian's mean μ_o. The number of a specific Gaussian in the Gaussian mixture $p_x(\mathbf{x}_k)$ with O Gaussians can be introduced as a hidden variable o, since $p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k)$ can be represented as the marginal density

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) = \sum_{o=1}^{O} p(\mathbf{x}_k, o|\mathbf{y}_{1:k}, \mathbf{n}_k).$$

With the equality

$$p(\mathbf{x}_k, o|\mathbf{y}_{1:k}, \mathbf{n}_k) = p(o|\mathbf{y}_{1:k}, \mathbf{n}_k) p(\mathbf{x}_k|o, \mathbf{y}_{1:k}, \mathbf{n}_k),$$

and pulling the sum over o out of the integral we yield the *vector Taylor series approach* (VTSA)

$$v^{(\mathrm{VTSA})}(\mathbf{y}_{1:k}, \mathbf{n}_k) = \sum_{o=1}^{O} p(o|\mathbf{y}_{1:k}, \mathbf{n}_k) \int \mathbf{x}_k p(\mathbf{x}_k|o, \mathbf{y}_{1:k}, \mathbf{n}_k) d\mathbf{x}_k.$$

6.5.2 The Gaussian Mixture Approach

In case of Monte Carlo sampling the noise variance is implicitly contained in the noise samples of the weighted empirical density. Thus a solution, which can be derived directly by considering the shift imposed by a single noise sample $\mathbf{n}_k$ to the oth Gaussian in the logarithmic spectral domain

$$\boldsymbol{\mu}'_o = \boldsymbol{\mu}_o + \underbrace{\ln(1 + e^{\mathbf{n}_k - \boldsymbol{\mu}_o})}_{\Delta_{\boldsymbol{\mu}_o, \mathbf{n}_k}}, \tag{6.22}$$

can be found without the need for a Taylor series expansion [179].

Instead of shifting the mean, we can conversely shift the corrupted spectrum $\mathbf{y}_k$ in the opposite direction to obtain the clean speech spectrum

$$\mathbf{x}_k = \mathbf{y}_k - \Delta_{\boldsymbol{\mu}_o, \mathbf{n}_k}. \tag{6.23}$$

This deterministic relationship yields

$$p(\mathbf{x}_k | o, \mathbf{y}_{1:k}, \mathbf{n}_k) = \delta_{\mathbf{y}_k - \Delta_{\boldsymbol{\mu}_k, \mathbf{n}_k}}(\mathbf{x}_k)$$

and hence we yield the *Gaussian mixture approach* (GMA)

$$\begin{aligned} v^{(\mathrm{GMA})}(\mathbf{y}_{1:k}, \mathbf{n}_k) &= \sum_{o=1}^{O} p(o|\mathbf{y}_{1:k}, \mathbf{n}_k) \int x \delta_{\mathbf{y}_k - \Delta_{\boldsymbol{\mu}_o, \mathbf{n}_k}}(\mathbf{x}_k) d\mathbf{x}_k \\ &= \sum_{o=1}^{O} p(o|\mathbf{y}_{1:k}, \mathbf{n}_k) \left(\mathbf{y}_k - \Delta_{\boldsymbol{\mu}_o, \mathbf{n}_k}\right) \\ &= \mathbf{y}_k - \sum_{o=1}^{O} p(o|\mathbf{y}_{1:k}, \mathbf{n}_k) \Delta_{\boldsymbol{\mu}_o, \mathbf{n}_k}. \end{aligned} \tag{6.24}$$

6.5.3 The Statistical Inference Approach

Noting that we have replaced the empirical density by Monte Carlo sampling let us directly use the non-linear relationship between $\mathbf{x}$, $\mathbf{n}$ and $\mathbf{y}$ [5].

Thus the marginal density $p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k)$ becomes deterministic, since $\mathbf{x}_k$ can be calculated from $\mathbf{y}_k$ and $\mathbf{n}_k$ as

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) = \delta_{\mathbf{y}_k + \ln(\mathbf{1} - e^{\mathbf{n}_k - \mathbf{y}_k})}(\mathbf{x}_k).$$

Substitution of $p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k)$ in $\int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{n}_k) d\mathbf{x}_k$ yields the *statistical inference approach* (SIA)

$$\begin{aligned} v^{(\mathrm{SIA})}(\mathbf{y}_{1:k}, \mathbf{n}_k) &= \int \mathbf{x}_k \delta_{\mathbf{y}_k + \ln(\mathbf{1} - e^{\mathbf{n}_k - \mathbf{y}_k})}(\mathbf{x}_k) d\mathbf{x}_k \\ &= \mathbf{y}_k + \ln(\mathbf{1} - e^{\mathbf{n}_k - \mathbf{y}_k}). \end{aligned} \tag{6.25}$$

6.6 Conclusion

The main focus of this chapter has been on the application of the particle filter in order to remove non-stationary additive distortions in the dimension reduced logarithmic frequency domain. The chapter has started by reviewing probability model based enhancement techniques and by introducing Bayesian filters for noise compensation. The latter sections have focused on the individual processing steps of the particle filter. Section 6.3 has discussed how to evaluate the noise estimate and suggested to replace the general speech model by phoneme dependent models. Section 6.4 copes with the prediction of noise. After reviewing the random walk and the predicted walk using a static autoregressive matrix it has been suggested to estimate the autoregressive matrix on a frame-by-frame basis within the particle filter framework. Finally, in Section 6.5 three different methods to remove the noise estimate from the noisy signal have been discussed.

Chapter 7

Compensation of Reverberation by Multi-Step Linear Prediction

In the previous section we have learned how to track and compensate for non-stationary additive distortions. In this section we review different ways to estimate and compensate for the second kind of distortions, namely *convolutive distortions*, which can be caused for example by reverberation. Probably the most prominent deconvolution algorithm with respect to *automatic speech recognition* (ASR) is *cepstral mean normalization* (CMN) [60]. Unfortunately it is only able to compensate for convolutive distortions which are no longer than the observation window. To satisfy the quasi-stationary assumption of speech signals this might not be longer than 32 ms in an ASR front-end. Petrick *et al.* [175], however, found that convolutions, caused by reflections which arrive between 100 ms after the direct path, and the time, where their sound level has decayed 40 dB below the level of the direct sound, have the strongest distortional effect on automatic classification. Those distortions appear significantly later than the time span covered by the observation window in CMN and thus can not be compensated by this or related techniques.

To estimate and compensate for those harmful late reflections several algorithms have been proposed. Probably one of the most promising family of methods as-

sumes that the reverberant power spectrum $\mathbf{r}_k$ is a scaled or weighted summation over previous frames

$$\mathbf{x}_k^{(\text{reverberant})} = \mathbf{x}_k + \mathbf{r}_k = \mathbf{x}_k + \sum_{m=1}^{M} \mathbf{s}_m \mathbf{x}_{k-m} \tag{7.1}$$

with frame index k and where the signal

$$\mathbf{x}_k = [|X(\Omega_1, k)|^2, |X(\Omega_2, k)|^2, \cdots, |X(\Omega_B, k)|^2]^T,$$

the reverberation

$$\mathbf{r}_k = [|R(\Omega_1, k)|^2, |R(\Omega_2, k)|^2, \cdots, |R(\Omega_B, k)|^2]^T$$

and the scale terms

$$\mathbf{s}_k = [S(\Omega_1, k), S(\Omega_2, k), \cdots, S(\Omega_B, k)]^T,$$

are dependent on the frequency Ω for bins $b = 1, 2, \cdots, B$. The scale terms can be determined for example by the Rayleight distribution [208] and adjusted by an estimate of the reverberation time or by more complex methods such as the one proposed by Sehr *et al.* [184].

In contrast to classical spectral subtraction methods [66] which estimate and subtract spectral energy caused by *additive distortions*, those methods estimate and subtract spectral energy caused by *reverberation*. The advantage of treating the reverberation also as additive in the power frequency domain is that those distortions, in this way, can be easily removed by simple subtraction without the need to estimate and invert the impulse response. In addition it has been shown by Lebart *et al.* [133] that those methods are not sensitive to fluctuations in the impulse response.

As an alternative to manipulating the input features, the acoustic models of the recognition system can be altered [119] by parallel model combination [102]. Even though significant performance improvements have been demonstrated [119], it stays unclear how to combine those methods with other model adaptation methods such as maximum likelihood linear regression. As we will demonstrate in the experimental section, feature enhancement techniques can be efficiently combined with model adaptation techniques to further reduce the recognition error. In addition feature enhancement techniques can readily be applied to other tasks such as speaker recognition.

Instead of estimating the reverberant power spectrum $\mathbf{r}_k$ by scaled versions of previous frames, as in (7.1), it has been proposed by Kinoshita *et al.* [129] to determine the reflection sequence in the time domain by *multi-step linear prediction* (MSLP) [104] and thereafter convert it into a reverberation estimate $\mathbf{r}_k$ by short-time spectral analysis. MSLP will be discussed in more detail in Section 7.3.

7.1 Knowing the Enemy: Harmful Effects of Reverberation

It is useful to gain an insight into the harmful effects of reverberation, in order to develop strategies for successfully combating them. Unfortunately, relatively little work has been published in this area focusing on automatic recognition. Pan and Waibel [172] have investigated the influence of room acoustics by comparing stereo data of close and distant recordings on the mel-scale logarithmic frequency domain derived from truncated mel-frequency cepstral coefficients. They observed that noise affects mainly the spectral valleys, while reverberation may also cause distortions at spectral peaks, i.e., at the fundamental frequency and its harmonics in voiced speech.

Although the definition varies from author to author, we will follow the probably most widely accepted definition [132] and consider *early reflections* to occur between the arrival of the direct signal and 50 ms thereafter. Similarly, we will take *late reflections* as any reflections or reverberation occurring after 50 ms. Tashev and Allred [195] found that reverberation between 50 ms and the time when the sound pressure has dropped 40 dB below its highest level, has the most damaging effect on the word accuracy of a far-field ASR system. Petrick *et al.* [175] have separately investigated early reflections, late reflections and reflections which are only present in low or high frequency regions in the context of ASR. They got slightly different results and concluded that late reflections, which appear between 100 ms after the direct path and the time where their sound level has decayed 40 dB below the level of the direct sound, have the most damaging effect on the classification accuracy. Furthermore, they found that reverberation in the frequencies between 250 Hz and 2.5 kHz leads to poor ASR accuracy while frequencies below 250 Hz and above 2.5 kHz do not have a significant impact on recognition accuracy.

Moreover, early reverberation in higher frequencies was found to improve automatic recognition performance. Similar results were found by Nishiura *et al.* [157], who reported that early reflections within approximately 12.5 ms of the direct signal actually improve recognition accuracy. This is significantly shorter than the 50 ms time frame wherein reflections were found to improve human recognition accuracy [132].

7.2 Problem in Speech Dereverberation

The transfer function between a speaker and a microphone in a room can be described by the *impulse response* $h[n]$, where n denotes the sample index, if we can assume that

- the environment is noise free and
- stationary.

The reverberant speech sequence $y[n]$ is then related to the clean input sequence $x[n]$ by a convolution with the impulse response $h[n]$, such that,

$$y[n] = \sum_{l=0}^{\infty} h[l]x[n-l]. \tag{7.2}$$

Deconvolution requires finding the inverse transfer function $h_{\text{inv}}[n]$ which would enable the estimation of the sequence $x[n-D]$, with some delay $D > 0$, given $y[n]$. Thus the ideal filter $h_{\text{inv}}[n]$ would have to satisfy

$$\sum_{l=0}^{\infty} h[l]_{\text{inv}} h[n-l] = \delta[n-D].$$

Perfect restoration of the sequence $x[n]$ would be possible in a noise free environment, if the impulse response would be known and if the transfer function would be minimum phase (in other words if the inverse of the impulse response would be stable and unique). In a realistic environment, however, we observe additive distortions, the system transfer function might be non-minimum phase and the impulse response is neither known nor can be well estimated—which is elementary as the success of filtering by the inverse of the transfer function is very sensitive to the correct estimate of $\mathbf{h}_{\text{inv}}$.

Apart from the noise the estimate of $\mathbf{h}$ is further complicated by the fact that the speech signal is not an i.i.d. sequence, as it has inherent features such as periodicity and as it follows a particular formant structure. This structure is due to the glottal, the vocal tract and the lip radiation filters, compare to Figure 3.2, which can be summarized to the speech production filter $\mathbf{h}_{\text{speech}}$.

Therefore, the observed reverberant speech signal

$$y[n] = \sum_{m=0}^{M} \underbrace{\sum_{l=0}^{L} h_{\text{room}}[m-l] h_{\text{speech}}[l]}_{h[m]} u[n-m] \tag{7.3}$$

is the convolution of the room impulse response $\mathbf{h}_{\text{room}}$ and the speech production filter $\mathbf{h}_{\text{speech}}$ with the excitation signal $\mathbf{u}$. Thus an inverse filter $\mathbf{h}_{\text{inv}}$, that converts a convolved sequence into a sequence where each component is independent, would not only filter out the impulse response of the channel $\mathbf{h}_{\text{room}}$, but also the impulse response of the speech production filter $\mathbf{h}_{\text{speech}}$ which is required for classification.

It follows that the success of combating reverberation depends mainly on the available knowledge sources and assumptions made. To develop strategies for dereverberation one should be aware of those parts in the reverberant speech signal which are particular harmful for the accuracy in ASR as already discussed in Section 7.1. In the literature most algorithms compensate for reverberation which starts around 50 ms; e.g. in [133]. In our own experiments we found that dereverberation algorithms, which start with the estimate of the reverberant energy 60 ms after the direct signal, provide the best recognition performance. For values between 50 up to 70 ms only a slight difference in the enhanced signal and thus recognition accuracy have been observed while values outside this range have led to significant lower recognition accuracies. The end time of the reverberation estimate should be sufficiently long to contain enough reverberation energy. This parameter, however, has only a limited effect on recognition accuracy and thus is not critical.

Separating the room impulse response into early and late reflections and assuming that the impulse response of the speech production filter is sufficiently short in comparison to the start time of the harmful intermediate to late reflection $M_{\text{early}} + 1$ let us express (7.3) as

$$y[n] \approx \sum_{m=0}^{M_{\text{early}}} \sum_{l=0}^{L} h_{\text{early}}[m-l] h_{\text{speech}}[l] u[n-m] + \sum_{m=M_{\text{early}}+1}^{\infty} h_{\text{late}}[m] u[n-m]. \tag{7.4}$$

7.3 Multi-Step Linear Prediction Estimation of Late Reflections

With the assumptions made in (7.4) it becomes possible to estimate and remove only those reflections which are not convolved with the transfer function of the speech production filter by estimating the correlation between the current observation $y[n]$ and the sequence $y[n+M_{\text{early}}+1], y[n+M_{\text{early}}+2], \ldots, y[n+M]$.

In order to estimate the correlation it has been proposed to use MSLP. In contrast to the well known *linear prediction* (LP), MSLP aims to predict a signal after a given delay D, the so called *step-size*. With the prediction error $e[n]$ we can formulate MSLP as

$$y[n] = \sum_{m=1}^{M} c_m y[n-m-D] + e[n]$$

where $c_1, \cdots, c_m$ represent the LP coefficients, $y[n]$ the observed signal and M the model order. The LP coefficients can be calculated by minimizing the mean square error of the error term $e[n]$. In matrix notation the solution for the MSLP coefficients $\mathbf{c} = [c_1, c_2, \cdots, c_M]^T$ is given by

$$\mathbf{c} = \left(E\left\{\mathbf{y}[n-D]\mathbf{y}[n-D]^T\right\}\right)^{-1} E\left\{\mathbf{y}[n-D]\mathbf{y}[n]^T\right\}$$

which can be efficiently solved using the Levinson-Durbin recursion.

An estimate of the reflection sequence $r[n]$ can be obtained by filtering the observed sequence $y[n]$ with the MSLP filter

$$r[n] = \sum_{m=1}^{M} c_m y[n-m-M_{\text{early}}+1] \tag{7.5}$$

where the delay D has been set to $M_{\text{early}} + 1$.

In order to remove the energy presented in the reflection sequence $r[n]$ it has been proposed by Lebart *et al.* [133] to use spectral subtraction [66]. Kinoshita *et al.* [129] have adopted this approach by converting the reflection sequence $r[n]$ into short-time power spectra $\mathbf{r}_{0:K}$. Note that in contrast to the removal of additive distortions, the late reverberation energy estimate is significantly changing for each frame k.

As the late reflection sequence $\mathbf{r}_{0:K}$ might still contain some correlation due to the speech production filter, it has been suggested to use pre-whitening prior to the estimation of the MSLP coefficients [128]. In our experiments, however,

we have not observed consistent gains and thus the pre-whitening filter has not been used for the experiments reported in this publication.

7.4 Conclusion

This chapter has started by investigating the harmful effects of reverberation on word accuracy. It has been found that reverberation caused by reflections which arrive between 50 ms after the direct path, and the time, where their sound level has decayed 40 dB below the level of the direct sound, have the strongest distortional effect on automatic classification. In Section 7.2 problems which might appear in finding the inverse of the room impulse response have been discussed. We have found that a couple of requirements can not fully be satisfied, including that the transfer function must be non-minimum phase, that the impulse response is neither known nor can be well estimated, that the signal is noise free and that the transfer function is stationary within a particular observation. Apart from the already mentioned problems the estimate of the inverse filter is further complicated by the fact that the speech signal is not an i.i.d. sequence which is due to the glottal, the vocal tract and the lip radiation filters, and that those properties are important for classification. To overcome some of the problems mentioned it has been suggested by Kinoshita *et al.*—instead of estimating the inverse impulse response—to estimate the energy caused by reverberation using multi-step linear prediction which has been reviewed in Section 7.3.

Chapter 8

Joint Compensation of Additive and Convolutive Distortions

While a lot of today's research in speech feature enhancement for *automatic speech recognition* (ASR) focus on compensating either stationary additive distortions such as background noise or convolutive distortions such as reverberation with a stationary room impulse response, most of the observed distortions are in reality *non*-stationary, additive *and* convolutive. These distortions, however, can neither be represented well under the stationary assumptions in the feature space by methods such as spectral subtraction [66] or constrained maximum likelihood linear regression nor in the model space by adaptation techniques such as maximum likelihood linear regression [137]. They are in fact one of the most significant problems in hands-free ASR.

To cope well with the non-stationary behavior of distortions, we have previously discussed approaches which track non-stationary noise on speech features in the logarithmic spectral domain. Although those algorithms work well with non-stationary noise, they are not able to remedy reverberation. In Section 7 an algorithm has been reviewed which is able to treat reverberation in the spectral domain. This algorithm, however, is not able to remove additive distortions.

To compensate for non-stationary noise as well as harmful reflections it is possible to simply concatenate the different processing steps. The full potential of speech feature enhancement, however, can only be reached by *jointly* estimating both kinds of distortions as they do in fact interact. For that we propose a generalized particle filter framework which is capable to jointly track additive noise and reverberation on a frame-by-frame basis. To integrate the reverberation estimate we have to extend the dimension of the particle filter. The first dimensions, equal to the number of spectral bins, represent estimates of additive noise while the additional dimensions model the scale of the reverberation estimate.

8.1 Tracking the Individual Distortion Types

To simplify modeling and tracking of the individual distortion types we aim to decompose the observed signal $\mathbf{y}$ into three parts

- the energy of the clean signal $\mathbf{x}$,
- the energy caused by additive noise $\mathbf{a}$ and
- the energy caused by reverberation $\mathbf{r}$.

Due to the large amount of free parameters which are required to estimate reverberation, we do not aim on tracking the room impulse response or late reverberation energy directly. Instead we aim on tracking the difference to the late reverberation energy estimate provided by an *auxiliary model*. The difference can be expressed in a scale term $\mathbf{s}$ of low dimensionality while the parameters of the auxiliary model are estimated over a larger time span such as an utterance.

Tracking of the additive noise $\mathbf{a}_k$ and the scale term $\mathbf{s}_k$, instead of the clean signal $\mathbf{x}_k$, becomes possible by the introduction of the two hidden variables $\mathbf{a}_k$ and $\mathbf{s}_k$ respectively

$$p(\mathbf{x}_k|\mathbf{y}_{1:k}) = \int\int p(\mathbf{x}_k, \mathbf{a}_k, \mathbf{s}_k|\mathbf{y}_{1:k}) d\mathbf{a}_k d\mathbf{s}_k.$$

With the relation

$$p(\mathbf{x}_k, \mathbf{a}_k, \mathbf{s}_k|\mathbf{y}_{1:k}) = p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{a}_k, \mathbf{s}_k) p(\mathbf{a}_k, \mathbf{s}_k|\mathbf{y}_{1:k})$$

and change in integration order we obtain

$$\mathrm{E}\{\mathbf{x}_k|\mathbf{y}_{1:k}\} = \int\int v(\mathbf{y}_{1:k}, \mathbf{a}_k, \mathbf{s}_k) p(\mathbf{a}_k, \mathbf{s}_k|\mathbf{y}_{1:k}) d\mathbf{a}_k d\mathbf{s}_k \tag{8.1}$$

where the function

$$v(\mathbf{y}_{1:k}, \mathbf{a}_k, \mathbf{s}_k) = \int \mathbf{x}_k p(\mathbf{x}_k|\mathbf{y}_{1:k}, \mathbf{a}_k, \mathbf{s}_k) d\mathbf{x}_k$$

maps the noisy observation sequence $\mathbf{y}_{1:k}$ given the distortion estimates $\mathbf{a}_k$ and $\mathbf{s}_k$ to the clean speech estimate $\mathbf{x}_k$. Note that due to the non-linear relationship between $\mathbf{a}$, $\mathbf{s}$, $\mathbf{y}$ and $\mathbf{x}$ in the chosen working domain, namely on the logarithmic mel-power coefficients, $v(\mathbf{y}_{1:k}, \mathbf{a}_k, \mathbf{s}_k)$ is also non-linear. Two solutions for $v(\mathbf{y}_{1:k}, \mathbf{a}_k, \mathbf{s}_k)$ have already been reviewed in Section 6.5 and can readily be applied using the mapping function $\mathbf{n}_k = u(\mathbf{a}_k, \mathbf{s}_k)$ which maps the additive distortion and scale term with the underlying late reverberation estimate $\mathbf{r}_k$ to the noise energy $\mathbf{n}_k$.

Folding the two vectors $\mathbf{a}$ and $\mathbf{s}$ into one "super" vector

$$\mathbf{d} = \begin{bmatrix} \mathbf{a} \\ \mathbf{s} \end{bmatrix}$$

let us express the filtering density $p(\mathbf{a}_k, \mathbf{s}_k|\mathbf{y}_{1:k})$ in (8.1) as $p(\mathbf{d}_k|\mathbf{y}_{1:k})$ and process as outlined in Section 6.2.1 and Section 6.2.2.

8.2 Modeling the Evolution of the Additive Noise and the Scale Term

To solve for (6.9) requires the prediction of the current distortion estimates $\mathbf{a}_k$ and $\mathbf{s}_k$ given the previous estimates $\mathbf{a}_{k-1}$ and $\mathbf{s}_{k-1}$ by the distortion transition probability

$$p(\mathbf{d}_k|\mathbf{d}_{k-1}) = \begin{bmatrix} p(\mathbf{a}_k|\mathbf{a}_{k-1}) \\ \cdots\cdots\cdots \\ p(\mathbf{s}_k|\mathbf{s}_{k-1}) \end{bmatrix}.$$

Various, previously proposed methods to model the transition probability $p(\mathbf{d}_k|\mathbf{d}_{k-1})$ have been reviewed in Section 6.4. However, only the random walk or the dynamic autoregressive model can be applied here as the noise term can not be independently learned prior to the application of the *particle filter* (PF). The estimate of the autoregressive matrix can be represented as a single matrix, however, we yielded better results by considering the additive noise and the scale terms as independent components

$$\mathbf{P}_k = \begin{bmatrix} \mathbf{A}_k & \vdots & \mathbf{0} \\ \cdots\cdots & \cdots & \cdots\cdots \\ \mathbf{0} & \vdots & \mathbf{S}_k \end{bmatrix} \tag{8.2}$$

where the additive distortion matrix $\mathbf{A}_k$ is recalculated for each frame k by the dynamic autoregressive process and the scale terms $\mathbf{s}_k$ are modeled by a random walk $\mathbf{S} = \mathbf{S}_k = \text{diag}(1)$.

8.3 Scaling the Reverberation Estimates

In order to compensate for estimation errors in the estimated reflection energy $\mathbf{r}_k$ which might be due to

- approximation of the reflection energy,
- additive noise in the estimate, as well as
- stationary assumption of the impulse response,

we introduce a scaling term

$$\mathbf{r}_k^{(\text{PF})} = \ln(\mathbf{s}_k)\mathbf{r}_k. \tag{8.3}$$

Note that the scaling term $\mathbf{s}_k$ is different to (7.1) as it changes for each frame k while the scale terms in (7.1) are usually constant over long observation windows such as an utterance. Thus $\mathbf{s}_k$ is able to adjust for changes in the room impulse response without updating the parameters of the reverberation models (7.1) or (7.5) which can only be estimated over a much longer time interval as they contain much more free variables; e.g. to cover a reverberation of 200 ms we need to estimate either 129 spectral bins multiplied by 20 frames = 2580 bins for model (7.1) or 16000 samples per second multiplied by 0.2 seconds = 3200 linear prediction coefficients for model (7.5).

The reverberation energy estimate in (8.3) can either

- be scaled by a single factor $s[1]$

$$r[b]_k^{(s)} = \ln\left(s[1]_k^{(s)}\right) r[b]_k \tag{8.4}$$

  for each frequency bin b, adding one dimension to the PF,

- be scaled by a single factor $s[1]$ and be tilted by $s[2]$ to scale lower and higher frequencies differently

$$r[b]_k^{(s)} = \ln\left(s[1]_k^{(s)} + s[2]_k^{(s)}(b - \bar{b})\right) r[b]_k \tag{8.5}$$

for each frequency bin b, where $\overline{b} = (B+1)/2$, adding two dimensions to the PF, or

- be scaled for each frequency bin individually $s[b]$

$$r[b]_k^{(s)} = \ln\left(s[b]_k^{(s)}\right) r[b]_k \tag{8.6}$$

for each frequency bin b, doubling the dimension of the PF.

As an individual scaling of each bin increases the search space significantly and thus the execution time, but could not outperform the alternative approaches with lower dimensionality, it will not be further investigated in this thesis, however it has been presented here for the sake of completeness.

8.4 Particle's Initialization

The first step in any PF framework is its initialization by drawing samples from the prior density. In the joint PF framework the particles have been drawn from the *prior distortion density*

$$p(\mathbf{p}_0) = \begin{bmatrix} p(\mathbf{a}_0) \\ \cdot\cdot\cdot\cdot\cdot \\ p(\mathbf{s}_0) \end{bmatrix} \tag{8.7}$$

which is a concatenation of the *prior additive distortion density* $p(\mathbf{a}_0)$ and the *prior scale density* $p(\mathbf{s}_0)$ of the estimated late reflection energies. Unfortunately, the prior additive distortion density $p(\mathbf{a}_0)$ can not be estimated directly as it is estimated between words or sentences and thus still contains significant energy due to reverberation. However, it can be decomposed into two densities which can be estimated:

- the *prior overall distortion density* $p(\mathbf{n}_0) = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ derived on silence regions of the input signal which contains additive *and* convolutive distortions and
- the *prior reverberation density* $p(\mathbf{r}_0) = \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ which is estimated over all frames derived on the late reflection energy sequence $\mathbf{r}_{0:K}$ estimated by *multi-step linear prediction* (MSLP) as described in Section 7.3.

With the prior overall distortion density and the prior reverberation density it is now possible to derive the prior additive distortion density as

$$p(\mathbf{a}_0) = \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_n)$$

by subtracting the mean value of the reverberation energy from the mean value of the noise energy

$$\boldsymbol{\mu}_a = \ln\left(e^{\boldsymbol{\mu}_n} - e^{\boldsymbol{\mu}_r}\right).$$

For simplicity the variance term $\boldsymbol{\Sigma}_a$ has been set to the variance term of the noise term $\boldsymbol{\Sigma}_n$ resulting in an overestimation of the variance which, however, is not critical here.

The prior scale density $p(\mathbf{s}_0)$ is given by a Gaussian $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ with $\mu_s = 1.0$ (for the actual scale terms) or $\mu_s = 0.0$ (for the actual tilt term) as we assume a correct estimate of the spectral energies which are due to reverberation. The variance term $\boldsymbol{\Sigma}_s$ is set to a small variable or can be learned from the data, however, in contrast to the mean values, the correct estimate of the variance is not critical.

8.5 Working Domain of Late Reverberation

We found that applying the dereverberation in the warped speech feature domain (129 dimensions), before feature reduction, is leading to slightly different estimates as a direct processing in the PF working domain. The improved accuracies of the speech recognition system, see Figure 8.1, can be explained by the higher accuracy of the subtraction due to the higher dimension of the working domain.

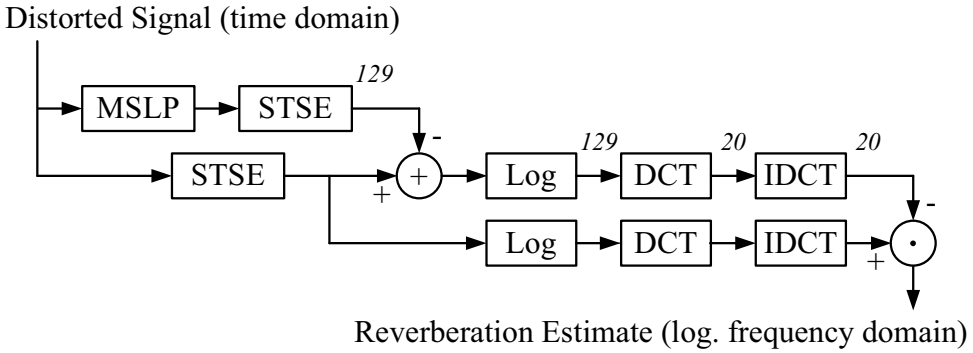


Figure 8.1: Flowchart of the reverberation estimate in the logarithmic frequency domain. STSE stands for short time spectral analysis, DCT and IDCT for discrete cosine transformation and its inverse respectively and MSLP for multi-step linear prediction.

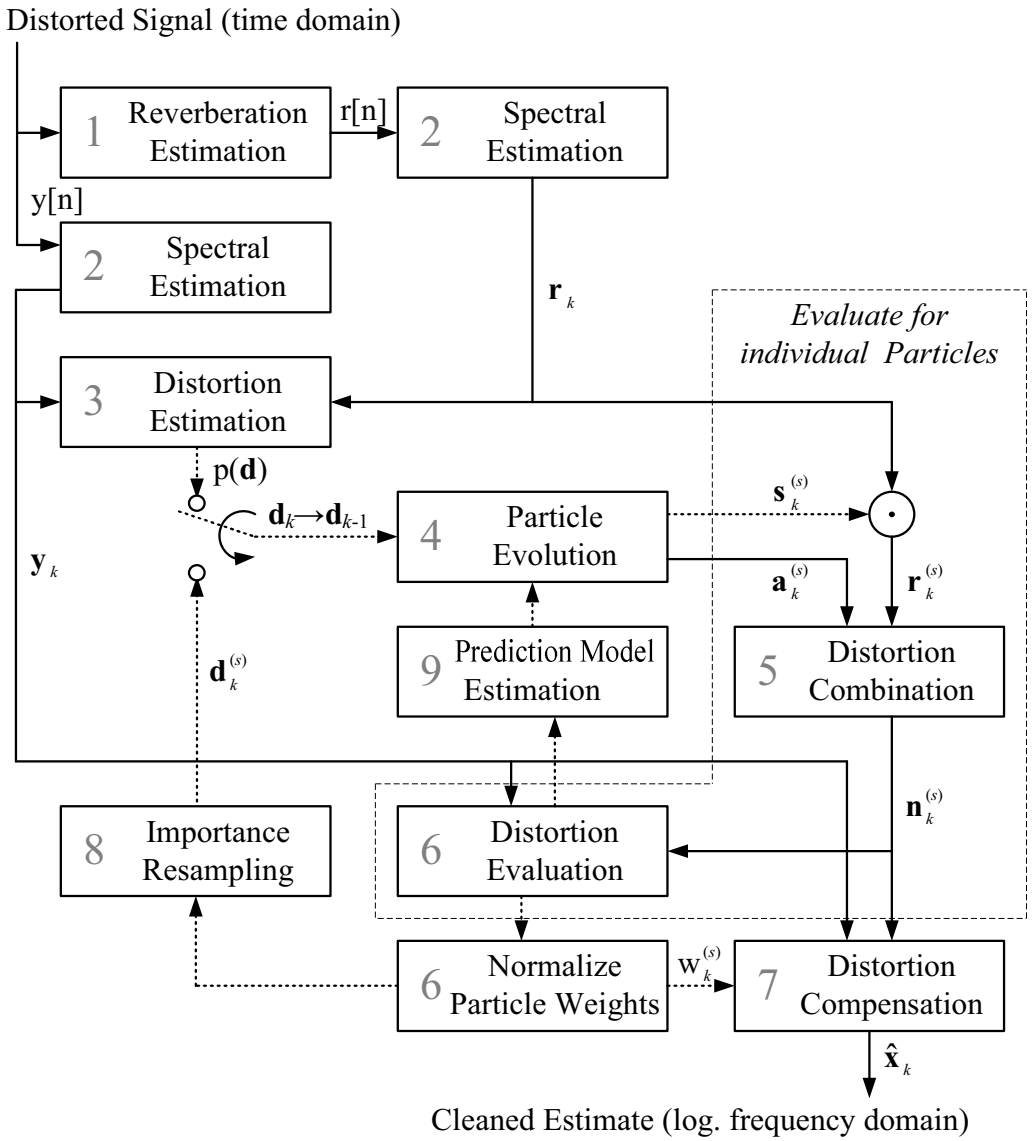


Figure 8.2: Flowchart of the joint particle filter approach for jointly estimating additive and convolutive distortions.

1. *Reverberation Estimation*
   The reverberation sequence is calculated by MSLP according to (7.5).
2. *Spectra Estimation*
   The reverberant and distorted short time power spectra are estimated for all frames.
3. *Distortion Estimation & Particle's Initialization*
   The prior additive distortion density $p(\mathbf{a}_0)$ and prior scale density $p(\mathbf{s}_0)$ are set as described in Section 8.4.
   Samples $\mathbf{d}_0^{(s)},\ s = 0, ..., S-1$, are drawn from the prior distortion density $p(\mathbf{d}_0)$ as defined in (8.7).
4. *Particle Evolution*
   All particles $\mathbf{d}_k^{(s)},\ s = 0, ..., S-1$, are propagated by the particle transition probability $p(\mathbf{d}_k|\mathbf{d}_{0:k-1})$ as defined in (8.2).
5. *Distortion Combination*
   The expected distortion $\mathbf{n} = u(\mathbf{a}, \mathbf{s})$ is calculated as

$$n[b]_k^{(s)} = \ln\left(e^{a[b]_k^{(s)}} + e^{r[b]_k^{(s)}}\right)\ \forall\ b\ \in\ B$$

   where $a[b]_k^{(s)}$ represents additive distortions and $r[b]_k^{(s)}$ represents the scaled spectral distortion due to reverberation as determined by either (8.4) or (8.5).
6. *Distortion Evaluation & Particle Weight Normalization*
   The distortion samples $\mathbf{n}$ are evaluated according to (6.15) and (6.16).
7. *Distortion Compensation*
   The clean feature is estimated according to either (6.24) or (6.25).
8. *Importance Resampling*
   Possibly the normalized weights are used to resample [181] among the noise particles $\mathbf{d}_k^{(s)},\ s = 1, ..., S$ to prevent the degeneracy problem.
9. *Prediction Model Estimation*
   The prediction matrix $\mathbf{A}_k$ in the dynamic transition probability model has to be updated according to (6.20).

Steps 4 until 9 are repeated with $k \mapsto (k+1)$ until either all frames are processed or the track is lost and has to be reinitialized with step 3.

Algorithm 8.1: Outline of the particle filter for speech feature enhancement to jointly estimate additive and convolutive distortions.

8.6 Overview of the Joint PF Approach

Figure 8.2 summarizes the joint particle filter estimation framework with its different components. In the image solid arrows represent the flow of the signal. Dotted arrows represent the flow of particle information such as the particle weight and the particle values which represent estimates for additive distortions for each frequency bin and a scaling factor for the convolutive distortion. The individual steps are described in Alg. 8.1.

8.7 Conclusion

In this chapter we have developed a framework which is able to jointly estimate and compensate for non-stationary additive distortions as well as convolutive distortions caused by late reverberation. This became possible by extending the dimensionality of the particle filter. The additional dimensions has been used to scale the reverberation estimate which has been provided by an auxiliary model as described in Section 7.3.

Chapter 9

Acoustic Channel Selection

Acoustic channel selection is important for *automatic speech recognition* (ASR) if acoustic channels with different acoustic qualities are available. This might happen, for example, in a lecture scenario where a close talk or lapel microphone is available for the main speaker and a room or hand held microphone is used to record questions or comments from the audience. In those cases, array processing techniques such as blind source separation or beamforming might not provide an enhanced signal as compared to the best single channel. Therefore techniques are required to select the channel which leads to the most accurate classification automatically.

Besides a direct application of acoustic channel selection the reference channel might also be determined in microphone array processing. Anguera *et al.* found that the success of microphone array processing is dependent on the quality of the reference channel [59].

In this chapter we review the classical *signal to noise ratio* (SNR) as an objective function for acoustic channel selection. The second class of objective functions for acoustic channel selection relies on knowledge which is provided by the decoder. Last but not least we propose the usage of class separability to improve multi-source far distance speech-to-text transcriptions. Class separability measures have the advantage, compared to other methods such as the SNR, that they are able to evaluate the channel quality on the actual features of

the recognition system and that they do not require time consuming decoding as decoder based methods. In addition using the class separability allows to evaluate channel quality even if silence regions are not present.

9.1 Review of Channel Selection Methods

Even though finding the channel which leads to the highest accuracy is an important and challenging task, it has not been a research topic which has drawn much attention over the last couple of years. To address this challenge, in the context of ASR, mainly two objective functions have been used which either rely on SNR or decoder information. We briefly review those methods:

- *Signal to noise ratio* is possibly the most widely used and is indeed a good indication for signal quality and proven to be useful in a broad variety of applications including channel selection for ASR [45]. It is handy and fast, but the quality of the result is strongly dependent on the estimate of speech and silence regions and in addition this measure is not considering any knowledge of the recognition system.

- *Decoder based* methods such as

  - *Maximum likelihood* that chooses the channel with the highest likelihood [185] or

  - *Difference in feature compensation* that compares the ASR hypothesis of uncompensated and compensated feature vectors for each channel and chooses the one with the smallest difference [164].

  The advantages of decoder based methods are the close coupling between the channel selection criteria and the recognition system, leading to more reliable estimations. The disadvantages are that for each individual channel, to not suffer from mismatch between the different channels, at least one recognition run is required — leading to a drastic increase in computation time. Obuchi [164] showed significant improvements, however, this method has two mayor drawbacks which makes it useless in real applications: first, it relies on word hypothesis, so for short utterances of no more than a couple of words, the quality of different channels would look identical and second, two recognition runs are required.

9.2 Class Separability in Channel Selection

In this section we introduce class separability as an objective function. It can be applied on different features and therefore allows to consider all possible information available in the recognition front-end. Furthermore, the classes required can be derived either as a stand alone or decoder based approach.

9.2.1 Scatter Matrices and Class Separability Measures

Class separability is a classical concept in pattern recognition, usually expressed using a scatter matrix. We can define

- the *within-class scatter matrix*

$$\mathbf{S}_w = \sum_{c=1}^{C} \left[\sum_{n}^{N_c} (\mathbf{x}_{cn} - \boldsymbol{\mu}_c)(\mathbf{x}_{cn} - \boldsymbol{\mu}_c)^T \right], \tag{9.1}$$

- the *between-class scatter matrix*

$$\mathbf{S}_b = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \tag{9.2}$$

- and the *total scatter matrix*

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b = \sum_{c=1}^{C} \left[\sum_{n}^{N_c} (\mathbf{x}_{cn} - \boldsymbol{\mu})(\mathbf{x}_{cn} - \boldsymbol{\mu})^T \right] \tag{9.3}$$

where N_c denotes the number of samples in class c, $\boldsymbol{\mu}_c$ is the mean vector for the cth class, and $\boldsymbol{\mu}$ is the global mean vector over all classes C.

Given the class scatter matrices, several separability measures are conceivable, probably the most widely used is

$$d = \text{trace}\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right) \tag{9.4}$$

which we have also used in our evaluations.

To not rely on the singularity of $\mathbf{S}_w$ it is also possible to calculate the separability measure as

$$d = \text{trace}\left(\mathbf{S}_b\right) / \text{trace}\left(\mathbf{S}_w\right), \tag{9.5}$$

another posibility is

$$d = \det\left(\mathbf{S}_w^{-1}\mathbf{S}_b\right).$$

The channel which maximizes the class separability

$$\widehat{ch} = \underset{\mathbf{ch}}{\operatorname{argmax}}\, d(\mathbf{ch})$$

is chosen to be used for classification.

9.2.2 Class Units to Calculate Class Separability

In the case of class separability and consequently in linear discriminant analysis it seems to be no consensus which class units should be used, e.g. phone, sub-phone, allophone or prototype level classes [110]. However, in large continuous speech recognition systems, where a lot of training data is available, it seems common to use sub-phone units.

In our opinion the ideal class unit might depend on the amount of available data to reliably estimate the scatter matrices. Due to very short utterances in our test set, some only one or two words long, we have limited our investigations to phone units (decoder based) and data driven units up to 32 classes (stand alone). To find the classes in the stand alone approach we have first separated between speech and silence frames by a simple voice activity detection. The speech frames have been further separated by classes derived by split and merge training (each Gaussian representing one class), either on the fly (on the utterance under investigation) or on the training data. A good classification is dependent on the separability between different phoneme classes only. Because the silence class is commonly not leading to confusion with a phoneme class, we have also considered cases where the silence class has been neglected in the calculation of the scatter matrices.

9.2.3 Feature Space

To determine reliable class separability measures one should aim to integrate as much knowledge about the human auditory system and to be as close as possible to the features as observed by the acoustic model of the ASR system. Therefore, we have used the 42 dimensional subspace

$$d(\mathrm{ch}) = \operatorname{trace}\left\{\left(\mathbf{W}^T\mathbf{S}_w^{\mathrm{ch}}\mathbf{W}\right)^{-1}\left(\mathbf{W}^T\mathbf{S}_b^{\mathrm{ch}}\mathbf{W}\right)\right\}$$

identical to the features as observed by the acoustic model of the ASR system. Here ch represents the investigated channel and $\mathbf{W}$ represents either the linear discriminant analysis matrix or the optimal feature space matrix. The trace is defined as the sum of the first n eigenvalues λ_i of a matrix (an n-dimensional subspace) and hence the sum of the variances in the principal directions.

9.3 Conclusion

Channel selection becomes important in those cases where blind source separation as well as beamforming techniques are not leading to improvement over the best single channel. Therefore, this chapter has reviewed various channel selection methods and suggested a new acoustic channel selection method based on class separability.

Chapter 10

Evaluation of the Proposed Methods

All theoretical developments are deemed to fail if they can not be applied on recordings captured with *real* speakers in *real* acoustic environments. In this chapter we, therefore, present numerous experiments on actual recordings—not artificially distorted—demonstrating the effectiveness of the proposed techniques. As such, the results of the experiments reported here are different from those reporting results on data that was originally captured with high signal quality, and thereafter artificially distorted by adding recorded noise, by convolving with a measured room impulse response or both. In our experience, the results obtained on such "artificial" data might fail to carry over because in real recordings the impulse response between the speaker's mouth and a microphone changes constantly.

In this thesis two objective functions are of particular interest (you optimize what you measure):

- *word error rate* to evaluate the proposed methods within the speech recognition system and
- *class separability* which is a good measure to compare different feature extraction techniques.

Besides the two frequently used objective functions we investigate the proposed methods also on other aspects to get a deeper understanding of the individual approaches and improvements.

10.1 The Janus Recognition Toolkit

Besides *Matlab* [199] for elementary experiments with fast turn around time and handy plotting functions, *Janus* has been the tool of first choice and exclusively used for the experiments reported in this thesis. Strictly speaking Janus is not only a speech recognition but also a speech-to-speech translation system. The Janus speech recognition engine has been around for more than 15 years and underwent a lot of transformation and extension in the past. By now there is probably not a single component of the original system that has not been rewritten several times or completely removed. After a complete reimplementation in 1995 the Janus recognition engine was renamed to *Janus Recognition Toolkit* (JRTk) to avoid further confusion. In 2001 the decoder has been changed to a single pass strategy [189]. Since 2002 the author has contributed to JRTk especially by working on the front-end, filter techniques and acoustic score functions. The author, however, wants to point out that JRTk is a joint effort of many dedicated researchers, developed jointly by the interACT located at the Universität Karlsruhe (TH) in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA. Without their contribution to this recognition system most parts of the experiments could not have been conducted.

10.1.1 Speech Recognition Setup

Most of the speech recognition experiments described in the following sections use a similar architecture. If not stated otherwise the experiments have been performed with a system architecture as described in the following sections.

10.1.1.1 Acoustic Front-End

The acoustic front-end provides features every 10 ms (first and second pass) or 8 ms (third pass) obtained by the specified spectral estimation technique. Vocal tract length normalization is applied on the estimated spectra either in the linear or warped domain respectively. In case of Fourier transformation a mel filterbank with 30 bands is used. In case of warped *minimum variance*

distortionless response (MVDR) or warped-twice MVDR either a linear filterbank with 30 bands (for model orders above 40) or no filterbank is used (for model orders below 40). In case of perceptual linear prediction the spectra are estimated according to [117]. The spectral estimation follows a discrete cosine transformation which is then truncated to either 13 or 20 cepstral coefficients. To compensate for the channel the cepstral sequence is mean and variance normalized. To include longer context 7 adjacent left and right frames are taken. To reduce the 195 or 300 dimensions to 42 dimensions and to maximize class separability, linear discriminant analysis is used. To further improve the representation of the acoustic information for speech recognition, a global semi-tied covariance transformation matrix [103] is estimated and multiplied with the dimension reduced features to obtain the final acoustic feature of order 42.

10.1.1.2 Phoneme and Filler Set

The phoneme and filler set is identical to our RT-06S [6] and RT-07S [53] evaluation systems. The phoneme set is an adapted version of the phoneme set used by the *Carnegie Mellon University* (CMU) dictionary which consists of 45 phonemes and allophones. Pronunciations of unknown words were generated automatically by Festival [64].

In comparison to previous systems, for example the RT-04S evaluation system [144], the used phoneme and filler set was augmented by additional noise models for laughter and other human noises to the existing breath and general noise models, and by a split of the filler model into a monosyllabic and a disyllabic filler model.

10.1.1.3 Acoustic Models

Acoustic model training was performed with fixed state alignments and fixed *vocal tract length normalization* (VTLN) factors. The acoustic models are represented by left-right *hidden Markov models* (HMM)s with three HMM states per phoneme without state skipping. Context information is introduced by using different sets of weights to differentiate between sub-phones that share a codebook. To derive initial codebooks, represented by up to 64 Gaussians with diagonal covariances each, we have used split and merge training (exact number of codebooks are indicated for each experiment but ranges between 2000 and 6000). The codebooks have been refined by two iterations of Viterbi training followed by four iterations of feature space speaker adaptive training [101]. The final codebooks were adjusted to compensate for recognition errors by two

iterations of maximum mutual information estimation training [177].

If not stated otherwise we used the following training material as summarized in Table 10.1. All the acoustic data is in 16 kHz, 16 bit quality and recorded with head-mounted microphones. Additional far-field data is available for ICSI, NIST and CHIL. Due to channel mismatch between ICSI and NIST data to the lecture meeting data we used only the far-field data provided by CHIL for supervised adaptation of the close-talking acoustic models to derive distant speech acoustic models.

Site	Type	Hours
CMU	meeting	11
ICSI	meeting	72
NIST	meeting	13
TED	lecture	13
CHIL	lecture	10
RT06	lecture & meeting	6

Table 10.1: Acoustic model training material.

10.1.1.4 Language Models

To train appropriate 3- or 4-gram language models we have combined corpora by linear interpolation as summarized in Table 10.2. The two inhouse web data collections were generated on queries on the most frequent n-grams in CHIL transcriptions and most frequent n-grams in the meeting transcripts, where irrelevant documents were skipped based on the perplexity on an in-domain *language model* (LM). For collecting the data we used the scripts provided by the University of Washington [95].

Site	Comments
TED transcripts	31 lectures from Eurospeech 1993
CHIL transcripts	subset
RT dev. & eval. data	subset
broadcast news	
proceedings	e.g. ICSLP, Eurospeech, ICASSP, ACL and ASRU
web data collection	University of Washington related to meetings
web data collection	inhouse related to lectures

Table 10.2: Language model training material.

The different LMs were build using the SRILM-toolkit [192]. For discounting we applied the Chen and Goodman's modified Kneser-Ney approach [79], and interpolation of discounted n-gram probability estimates with lower-order estimates was used. Pruning was performed after combining the different LM-components while the threshold was set with respect to a reasonable size of the LM.

The perplexity for the different test sets is around 125 for English and around 200 for German. The out of vocabulary rate of the different test sets where in all cases below 1.5%.

10.1.1.5 Adaptation

The speech recognition experiments conducted in the following sections used a multi pass strategy to allow for unsupervised adaptation. The adaptation parameters were estimated on the first best hypothesis of the prior pass.

The processing steps for the two-pass decoding strategy can be summarized as follows:

1. **Pass 1**
   Decode with the unadapted acoustic model.
2. **Adaptation**
   Estimate the *vocal tract length normalization* (VTLN) parameter, *constrained maximum likelihood linear regression* (CMLLR) parameters and *maximum likelihood linear regression* (MLLR) parameters for each speaker.
3. **Pass 2**
   Decode with the adapted acoustic model.

10.2 Objective Functions

To measure the quality of transcriptions, an objective measure is necessary. Since the early 1980s *word error rate* (WER) stabilized to be the measure of choice to compare between different *automatic speech recognition* (ASR) systems or to report improvements on the same system.

Even though the WER is widely accepted and used, a broad variety of additional *objective functions* or *cost functions* is necessary:

- *in the system itself*
  e.g. the Mahalanobis distance to evaluate the acoustic score
- *for system development*
  e.g. the perplexity for fast turn around times in language modeling
- *evaluation where the WER is not accessible*
  e.g. the maximum likelihood to adapt the acoustic models
- *outside the ASR system*
  e.g. signal to noise ratio or reverberation time to measure the signal quality.

Therefore, objective functions are required which have a high correlation to the WER. To optimize the performance of the overall system, one seeks to minimize or maximize the objective function instead of the WER. In the following section we introduce objective functions which are used throughout the experimental section.

10.2.1 Word Error Rate

There are typically three types of errors in text transcriptions, namely

- *insertions*: an extra word is added to the recognized word sequence,
- *substitutions*: a correct word in the word sequence is replaced by an incorrect word, and
- *deletions*: a correct word in the word sequence is omitted.

To determine the minimum error rate you have to align the string to score with the reference word string, which is known as *maximum substring matching*. After the alignment the WER can easily calculated by

$$\text{WER} = \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Number of Words in the Reference}}.$$

To evaluate our ASR system we have used the case-less sensitive lexical form known as the standard normalized orthographic representation. Furthermore non-lexical tokens such as breath or noise are not evaluated in scoring.

10.2.1.1 Differences between close- and distant-talking microphone transcriptions

Burger [71] has investigated the differences between close- and distant-talking microphone transcriptions. She found that the transcribers, to generate distant-talking microphone transcriptions from close-talking microphone transcriptions, had to remove an average of 4% of complete talk spurts, 2% of word tokens, 15% of word fragments and 12% of laughter annotations. The far-field transcriptions show an average of 60% more labels for non-identifiable utterances and 19% more word tokens tagged as hard to identify. The difference on breath has not been investigated.

10.2.2 Perplexity and Out of Vocabulary Rate

Perplexity is a measure in information theory and defined as

$$P(x) = 2^{H(x)} = 2^{-\sum_{k=1}^{n} p(k) log_2 p(k)}$$

where $H(x)$ is the entropy of the probability distribution of x, $p(k)$ is the probability of the k-th event in the distribution, and n is the number of possible events in the distribution.

In *natural language processing*, perplexity is a usual way of evaluating language models. The lower the perplexity a language model has, the easier it is to predict the next word given the previous n words and the language model. Domain-specific texts usually have lower perplexity (= less variation) than general language.

The *out-of-vocabulary rate* (OOV) defines the number of words which are not present in the dictionary. Usually a OOV word causes more than one error due to the relation of the language model.

10.2.3 Class Separability

We would like to derive acoustic feature vectors so that all vectors belonging to the same class (e.g. phoneme) are close together in feature space and well separated from the feature vectors of other classes (phonemes). Those properties can be expressed in the scatter matrices where a small within-class scatter and a large between-class scatter stand for large class separability. Therefore, an

approximate measure of class separability can be expressed by [109]

$$D_d = \text{trace}_d \left\{ \tilde{\mathbf{S}}_w^{-1} \cdot \tilde{\mathbf{S}}_b \right\} = \text{trace}_d \left\{ \left(\mathbf{W}^T \mathbf{S}_w \mathbf{W} \right)^{-1} \cdot \left(\mathbf{W}^T \mathbf{S}_b \mathbf{W} \right) \right\} \tag{10.1}$$

where $\mathbf{S}_w$ and $\mathbf{S}_b$ are already defined in Section 9.2.1. The trace_d is defined as the sum or the first d eigenvalues λ_i of $\tilde{\mathbf{S}}_w^{-1} \cdot \tilde{\mathbf{S}}_b$ (an d-dimensional subspace) and hence the sum of the variances in the principal directions, here $\mathbf{W}$ defines the linear discriminant matrix. It can also be interpreted as the radius of the scattering volume.

10.2.4 Signal to Noise Ratio

The *signal to noise ratio* (SNR) is a measure to compare between the level of a desired signal (such as speech) to the level of noise. Because many signals have a very wide dynamic range, SNRs are usually expressed in terms of the logarithmic decibel scale

$$\text{SNR} = 10 \log_{10} \frac{P_{\text{signal}}}{P_{\text{noise}}}$$

where P is an average power measured over the system bandwidth.

10.3 Feature Extraction and Adaptation

In order to evaluate the performance of the proposed warping (Section 3.4.6), scaling (Section 3.4.7), and model order adaptation (Section 5) of the MVDR envelope we conducted recognition experiments on the TED corpus. In contrast to the described approach for acoustic modeling (as the experiments have been conducted before the new setup) we have trained the acoustic models on the *Broadcast News* corpus [138], containing 104 hours of speech, and adapted on 31 speakers, 8 hours of speech, from the TED corpus by *maximum likelihood linear regression* (MLLR). The test set contained the final 8 speakers of the TED corpus.

The static cepstral coefficients were obtained by a spectral representation through a discrete cosine transform from either

- warped-LP(13), warped-MVDR(60) or warped and scaled-MVDR(60) envelope with fixed, the *model order* (MO) is given in brackets, or variable

MO followed by a filterbank consisting of 30 filters adapted to compensate for the differences between the bilinear transform and the mel-frequency, or

- the Fourier power spectrum, LP(20) or MVDR(80) envelope followed by a mel-filerbank.

The parameters of the MOs used in these experiments were tuned on a small development set without adaptation. The MOs of the different approaches can be explained by the characteristic of the different envelopes in combination with the filterbank following.

Front-End	WER %	RER %
Fourier	38.4	–
LP	39.7	-3.4
Perceptual LP	38.9	-1.3
Warped LP	38.7	-0.8
MVDR	38.6	-0.5
Warped MVDR	38.1	0.8
Warped & Scaled MVDR		
fixed	37.7	1.8
test	37.5	2.3
adaptation & test	37.0	3.6
training & adaptation & test	37.1	3.4

Table 10.3: The *word error rate* (WER) together with their *relative error reduction* (RER), in comparison to the Fourier power spectrum (the classical *mel-frequency cepstral coefficient* (MFCC) front-end), is given for different spectral representations.

The results of our speech recognition experiments, reported in absolute WER and relative WER reduction in Table 10.3, confirm our theoretical conclusions. We see a clear improvement going from LP to warped LP, whose performance is comparable to perceptual-LP and to MVDR. Going from MVDR to warped MVDR yields another significant improvement. The proposed MO adaptation leads to a further improvement resulting in a relative WER improvement of 3.6% in comparison to the widely used mel-frequency cepstral coefficients.

10.4 Signal Sensitive Feature Resolution

In order to evaluate the performance of the proposed warped-twice MVDR spectral estimation, Section 4.1, in combination with the steering factor, Section 4.2, we ran experiments on close-talking development and evaluation data of the Rich Transcription 2005 Spring Meeting Recognition Evaluation [158] lecture meeting task.

10.4.1 Class Separability

Comparing the class separability of different spectral estimation methods in Table 10.4 for *close-talking microphone* (CTM) and Table 10.5 for distant microphone recordings we first note that a higher number of cepstral coefficients always results in a higher class separability. Comparing the class separability, for 20 cepstral coefficients, on different front-ends we observe that class separability increases from *perceptual linear prediction* (PLP), *warped-twice linear prediction* (W2LP), *warped MVDR* (WMVDR), Fourier power spectrum to *warped-twice MVDR* (W2MVDR). The class separability is significant lower for PLP and significant higher for warped-twice MVDR, while warped-twice LP, warped MVDR and Fourier power spectrum have nearly the same value.

Test Set			Train	Develop	Eval
Front-End	Order	Cepstra	Class Separability		
Fourier	–	13	11.007	16.470	16.088
Fourier	–	20	11.620	17.929	16.299
PLP	13	13	10.699	17.110	15.152
PLP	20	20	11.029	18.059	16.068
WMVDR	60	13	10.768	16.813	16.261
WMVDR	60	20	11.337	18.022	16.614
WMVDR	30	13	10.900	17.675	16.702
WMVDR	30	20	11.386	18.630	17.318
W2LP	20	13	10.772	17.038	16.254
W2LP	20	20	11.333	17.864	16.436
W2MVDR	60	13	10.893	17.673	16.456
W2MVDR	60	20	11.473	18.510	16.818

Table 10.4: Class separability for different front-end types and settings on close-talking microphone recordings (note that in the WMVDR front-end with model order 30 applies no smoothing and dimension reduction by a filterbank).

Test Set			Develop	Eval
Front-End	Order	Cepstra	Class Separability	
Fourier	–	13	14.786	13.470
Fourier	–	20	15.806	13.944
PLP	13	13	15.121	12.917
PLP	20	20	15.399	12.975
WMVDR	60	13	13.836	13.885
WMVDR	60	20	14.487	14.161
WMVDR	30	20	15.111	14.155
W2LP	20	13	14.524	13.393
W2LP	20	20	15.119	13.803
W2MVDR	60	13	14.895	13.901
W2MVDR	60	20	15.380	14.116

Table 10.5: Class separability for different front-end types and settings on distant microphone recordings.

On CTM recordings in Table 10.4, we observe that warped-twice MVDR provides features with the highest separability on the development as well as the evaluation set. Averaging development and evaluation set the warped MVDR(30) is followed by warped-twice MVDR(60), warped MVDR(60), warped-twice LP(20), Fourier power spectrum and PLP. On distant microphone recordings, where the distance between speakers and microphones varies between approximately one and three meters, the Fourier power spectrum has the highest class separability on the development set. On the evaluation set, warped-twice MVDR performs equally well as warped MVDR, see Table 10.5. Averaging development and evaluation set on the distant data, the Fourier power spectrum provides the highest class separability followed by warped-twice MVDR(60), warped MVDR(30), warped-twice LP(20), warped MVDR(60) and PLP.

10.4.2 Word Error Rate

The error rates of speech recognition experiments for different spectral estimation techniques and passes are presented in Table 10.6 for close-talking and Table 10.7 for distant microphone recordings.

Comparing the averaged WERs over close and distant talking of different spectral estimation methods we observe that a higher number of cepstral coefficients does not always result in a lower WER. Power spectra, warped and warped-twice MVDR envelopes tend to better performance with 20 cepstral coefficients while

Test Set			Develop			Eval		
Pass			1	2	3	1	2	3
Front-End	Order	Cepstra	Word Error Rate %					
Fourier	–	13	36.1	30.3	28.0	35.3	29.7	27.7
Fourier	–	20	36.0	29.7	27.7	37.2	31.3	28.4
PLP	13	13	34.7	29.3	27.2	34.2	29.6	27.1
PLP	20	20	34.7	29.5	27.7	34.9	30.3	27.9
WMVDR	60	13	35.0	30.0	28.2	35.5	29.9	27.6
WMVDR	60	20	34.5	29.1	27.3	35.3	29.6	27.3
WMVDR	30	13	34.6	29.8	27.8	34.7	29.6	27.2
WMVDR	30	20	33.9	29.1	27.4	34.9	29.2	26.9
W2LP	20	13	35.3	30.5	28.5	36.2	29.8	27.1
W2LP	20	20	34.4	29.5	27.4	37.1	29.4	26.8
W2MVDR	60	13	34.5	29.5	27.5	34.1	29.2	27.0
W2MVDR	60	20	34.1	28.8	26.8	35.4	29.0	26.3

Table 10.6: Word error rates for different front-end types and settings on close-talking microphone recordings (note that in the WMVDR front-end with model order 30 applies no smoothing and dimension reduction by a filterbank).

PLP performs better with 13 cepstral coefficients. The following discussion always refers to the lower WER. In average warped-twice MVDR provides the lowest WER followed by warped-twice LP and warped MVDR which perform equally well. PLP has a lower WER on the first and second pass which equals on the third compared to the Fourier power spectrum. PLP provides the lowest feature resolution which seems to be an advantage on the first pass, however, after model adaptation the lower feature resolution seems to be a disadvantage.

Investigating the WER on CTM recordings only, Tabel 10.6, we observe that the warped-twice MVDR(60) front-end provides the best performance, followed by PLP, warped MVDR(30) and warped-twice LP(20) which are equally off (statistically). Warped MVDR(60) ranks before the Fourier power spectrum which finishes last.

On distant microphone recordings, Table 10.7, the warped-twice MVDR(60) front-end shows robust performance and has, in average, the lowest WER. On the development set, however, the Fourier power spectrum and warped MVDR(30) have the lowest WER. In average the warped-twice MVDR(60) is followed by warped MVDR(60), than warped-twice LP(20), thereafter the Fourier power spectrum due to a weak performance on the evaluation set and PLP on the last place.

Test Set			Develop			Eval		
Pass			1	2	3	1	2	3
Front-End	Order	Cepstra	Word Error Rate %					
Fourier	–	13	61.9	52.0	51.1	60.8	54.2	51.1
Fourier	–	20	59.8	50.4	48.9	61.0	55.0	51.7
PLP	13	13	60.7	51.8	50.5	59.9	53.4	51.8
PLP	20	20	59.8	52.1	50.2	59.6	54.4	52.7
WMVDR	60	13	62.9	53.7	52.0	60.7	52.8	50.7
WMVDR	60	20	60.9	51.2	49.7	59.6	51.7	49.5
WMVDR	30	20	59.0	50.5	48.9	59.3	52.1	49.9
W2LP	20	13	62.8	53.8	52.1	61.1	54.5	50.9
W2LP	20	20	58.9	50.8	49.3	59.9	53.0	50.2
W2MVDR	60	13	63.1	53.6	51.6	60.7	52.7	49.3
W2MVDR	60	20	60.3	51.1	49.8	59.9	50.4	47.9

Table 10.7: Word error rates for different front-end types and settings on distant microphone recordings.

The reduced improvements of the warped-twice MVDR in comparison to the warped MVDR on distant microphone recordings can be explained by the fact that, in comparison to CTM recordings, the range of the values φ_i over all i is reduced. Therefore, the effect of spectral resolution steering is attenuated and consequently warped-twice MVDR envelopes behave more similarly to warped MVDR envelopes.

Comparing the W2MVDR front-end with model order 60 and filterbanks to its warped MVDR counterpart, we observe a constant gain of at least 0.5% in word accuracy. If we wish to neclect the filterbank we have to compensate for its smoothing behaviour by reducing the model order to 30 and—for best performance—we have to increase the number of cepstral coefficients to 20. This leads to an improvement of at least 0.6% in word accuracy compared to the warped MVDR with filterbanks.

10.4.3 Phoneme Confusability

We investigate the confusability between phonemes by calculating the minimum distances, on the final features, between different phoneme pairs. In order to account for the range of variability of the sample points in both phoneme classes Ω_p and Ω_q, expressed by the covariance matrices $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$, we extend the well

known Mahalanobis distance by a second covariance matrix

$$D_{p,q} = \sqrt{(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T \left(\boldsymbol{\Sigma}_p + \boldsymbol{\Sigma}_q\right)^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)}.$$

Here $\boldsymbol{\mu}_p$ denotes the sample mean of phoneme class Ω_p and $\boldsymbol{\mu}_q$ denotes the sample mean of phoneme class Ω_q respectively.

As the comparison of the confusion matrix itself would be impractical we limit our investigations on the comparison of the *distance* between the *nearest* phoneme to a given *phoneme* for different spectral estimation techniques as plotted in Tabel 10.8. Note that the PLP front-end is excluded from this analysis as it, due to a different scale, can not be directly compared. By comparing the nearest phoneme pairs over different phonemes and spectral estimation methods we observe that different spectral representations result in slightly different phoneme pairs. In addition we observe that, in average, phonemes with a small value of φ are easier confused (smaller distance) with other phonemes than phonemes with a high φ value. This can be explained by the energy of the different phoneme classes where the phoneme classes belonging to small φ values contain less energy and are thus stronger distorted by background noise.

Comparing the power spectrum with the warped MVDR envelope we observe that the power spectrum tends to provide lower confusability for lower φ values and higher confusability for higher φ values. The warped-twice LP and warped-twice MVDR envelopes have a similar distance structure over φ, with in average larger distances for the warped-twice MVDR envelopes. While the warped-twice MVDR envelopes, compared to the warped MVDR envelope, provide a lower confusability for small values of φ, the confusability is higher for larger values of φ. While the warped MVDR envelope is not capable to provide a lower confusability over the whole range of φ in comparison to the power spectrum, the warped-twice MVDR envelope provides, in average, a lower confusability over the whole range of φ in comparison to the power spectrum.

phoneme	S	SH	CH	Z	JH	ZH	F	TH	T	K	HH	D	DH	⋯
φ	0.51	0.55	0.60	0.62	0.73	0.78	0.80	0.81	0.85	0.89	0.93	0.93	0.93	⋯
spectrum	power spectrum													
nearest	Z	CH	JH	S	CH	JH	T	T	TH	P	T	T	D	⋯
distance	2.41	1.56	0.81	2.27	1.36	1.55	2.36	2.04	1.75	2.33	1.99	1.95	2.45	⋯
spectrum	warped MVDR													
nearest	Z	CH	JH	S	CH	JH	T	T	TH	P	T	T	D	⋯
distance	2.32	1.56	0.86	2.21	1.65	1.49	2.26	2.03	1.74	2.36	2.15	2.05	2.56	⋯
spectrum	warped-twice LP													
nearest	Z	CH	JH	S	CH	JH	K	T	TH	P	T	T	D	⋯
distance	2.46	1.58	0.87	2.26	1.78	1.5	2.38	2.09	1.72	2.37	2.04	1.93	2.47	⋯
spectrum	warped-twice MVDR													
nearest	Z	CH	JH	S	CH	JH	T	T	TH	P	T	T	D	⋯
distance	2.43	1.6	0.85	2.24	1.75	1.58	2.35	2.08	1.74	2.35	2.11	1.97	2.49	⋯

phoneme	⋯	B	AXR	UH	OW	OY	W	UW	XL	NG	N	XN	M	XM
φ	⋯	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
spectrum	power spectrum													
nearest	⋯	G	R	UW	XL	OW	B	UH	L	N	M	N	N	L
distance	⋯	2.64	3.22	3.28	3.19	3.55	3.04	2.97	2.94	3.32	2.83	3.59	3.04	4.88
spectrum	warped MVDR													
nearest	⋯	G	R	UW	XL	AY	B	UH	L	N	M	N	N	XL
distance	⋯	2.83	3.45	3.44	3.49	3.8	3.29	3.19	3.18	3.52	3.01	3.65	3.3	5.07
spectrum	warped-twice LP													
nearest	⋯	G	R	UW	XL	OW	B	UH	L	N	M	N	N	XL
distance	⋯	2.67	3.14	3.19	3.22	3.47	3.06	2.93	2.96	3.36	2.77	3.57	3.03	4.97
spectrum	warped-twice MVDR													
nearest	⋯	G	R	UW	XL	OW	B	UH	L	N	M	N	N	XL
distance	⋯	2.72	3.28	3.21	3.26	3.59	3.1	2.99	3	3.36	2.83	3.56	3.12	5.02

Table 10.8: Nearest phoneme distance for different phonemes (ordered by φ) and spectral estimation methods.

10.5 Non-Stationary Noise Compensation

In order to evaluate the performance of the baseline PF with different spectral representations and feedback as proposed in Section 6.3.2 under realistic conditions we have chosen approximately 45 minutes of lecture speech which has been used in the Rich Transcription 2005 Spring Meeting Recognition Evaluation [158]. To perform experiments with different SNRs we have artificially added dynamic noise with a broad variety of sounds coming from a truck, slamming rubbish containers, distant voices, and shouts [198].

SNR			∞ dB		10 dB		5 dB		0 dB	
Pass			1	2	1	2	1	2	1	2
Front-End	Model	PF Adp.	Word Error Rate %							
Fourier	none	-	31.7	25.5	42.6	30.3	48.7	34.2	62.7	44.7
warped MVDR	none	-	31.0	25.4	39.4	29.2	48.1	33.8	60.2	42.4
Fourier	GSM	-	-	-	41.0	29.6	46.2	33.7	60.1	43.8
warped MVDR	GSM	-	-	-	38.5	28.4	45.6	33.5	57.0	42.1
warped MVDR	PSM	ref.	-	-	36.9	28.3	41.9	30.0	51.0	36.7
warped MVDR	PSM	hypo.	-	-	-	28.5	-	34.7	-	43.0
warped MVDR	MM	ref.	-	-	37.1	28.5	43.7	32.2	53.9	39.5
warped MVDR	MM	hypo.	-	-	-	28.4	-	31.8	-	40.3

Table 10.9: Word error rates for different front-ends, different or no *particle filter* (PF) and *signal to noise ratios* (SNR)s. The PF can either use the *general speech model* (GSM), the *phoneme-specific speech model* (PSM) or the *mixture model* (MM). PF adaptation of the PSM is either based on the hypothesis (unadapted recognition output) or the reference. The adapted speech recognizer pass has always been adapted with the output of the corresponding unadapted recognition pass.

Table 10.9 shows WERs for unadapted and adapted passes. Vocal tract length normalization, in contrast to the results reported otherwhere in this thesis, has not been used in these experiments (due to the implementation of the PF at the time of the experiments). The following discussion concentrates, if not stated otherwise, on the more relevant adapted results only.

For clean features the two different front-ends perform equally well. For decreasing SNRs the MVDR based features clearly outperform the Fourier based ones. The "traditional" PF shows good improvements for the unadapted recognition pass which reduces to marginal improvements on the adapted recognition pass. At 0 db, the MVDR based features can even improve accuracy over Fourier based features in combination with a "traditional" PF. The combination of

MVDR based features and the "traditional" PF can further improve the good result.

As most of the gain seen on the unadapted pass levels off on the adapted pass, we conclude that the adaptation of the speech recognition system compensates for most of the noise cleaned by the "traditional" PF. The good result, a gain in accuracy of more than 5% relative, of the proposed phone-specific PF on the reference and the proposed mixture on the hypotheses states, that the phone-specific PF is able to compensate for noises which can not be compensated by the adaptation of the speech recognition system. Note that the phoneme-specific PF failes in the case where no mixture model has been used on the hypotheses of the ASR engine. This demonstrates the problem of "model tying".

We have observed that approximately 3 to 5 percent of the frames get lost, because all particles got a likelihood value of zero. The number of dropouts seems to increase for a decrease in SNR. The dropout rate of the Fourier transformation was 10 percent higher than the one based on the warped MVDR.

10.6 Joint Compensation of Non-Stationary Additive Distortions and Reverberation

In order to evaluate the performance of the proposed joint PF algorithm, Section 8, under realistic conditions we have recorded and transcribed 35 minutes of lecture speech (continuous, freely spoken) by an English speaker with different microphone types and speaker to microphone distances (similar to NIST's RT-06s development and evaluation data [93], however, including a lapel microphone). To demonstrate and confirm our results we have made an additional recording, yet in German, containing 45 minutes of lecture speech. The German set, however, does not contain as much dynamic background noise as the steerable cameras in the CHIL room have been turned off.

10.6.1 Data and Algorithm Analysis

In this section we analyze the speech signals recorded with different speaker to microphone distances on the English test set. We start our analysis by estimating the signal-to-additive-distortion (labeled with Additive), signal-to-reverberation (labeled with Reverberation) and signal-to-distortion (labeled with Together) ratio calculated within the joint estimation framework. Comparing the different estimates in Table 10.10 to the signal-to-noise (labeled with

SNR) estimate based on voice activity detection we immediately observe that the distortion estimates are significantly higher within the joint estimation framework which becomes more expressed for higher SNR values.

On the CTM recordings the energy estimates of additive distortions and the energy estimates of late reverberation are nearly alike. The distortion energy estimates of the lapel microphone are higher for late reverberations than for additive distortions. This is also true for the table-top microphone, however, the difference between additive distortions and late reverberation energies is much smaller. On the wall mounted microphone the energy estimates of additive distortions and late reverberation become again nearly similar. In addition we observe that the energy estimates of late reverberation only slightly increase between the lapel, table-top and the wall mounted microphone.

Microphone	CTM	Lapel	Table-Top	Wall
Distance	1 cm	20 cm	1.5–2 m	3–4 m
Estimate	Average Energy vs Cleaned Estimate dB			
SNR	24	23	17	10
Additive	15.1	13.7	12.0	11.3
Reverberation	15.5	11.6	11.5	11.1
Together	12.3	9.5	8.7	8.2

Table 10.10: Average energy of non-stationary additive distortions and late reverberation vs cleaned speech estimate.

Figure 10.1 presents the average energy over all energy bands of the observed speech signal, the non-stationary additive distortion estimate, the late reverberation estimate and the cleaned estimate. Comparing the energies of the additive distortion and late reverberation estimated over different frames we can clearly observe the time dependent characteristics of the distortions. Furthermore, we note that the reverberation estimate has a significantly higher fluctuation than the additive distortion estimate, except for the CTM.

Microphone	CTM	Lapel	Table-Top	Wall
Distance	1 cm	20 cm	1.5–2 m	3–4 m
Pair	Normalized Correlation			
distorted signal - add. estimate	0.213	0.103	0.321	0.321
distorted signal - rev. estimate	0.511	0.586	0.536	0.569
add. estimate - rev. estimate	0.261	0.115	0.150	0.171

Table 10.11: Normalized correlation, averaged over all frequency bands in the logarithmic frequency domain, between the distorted signal and the two estimated distortions.

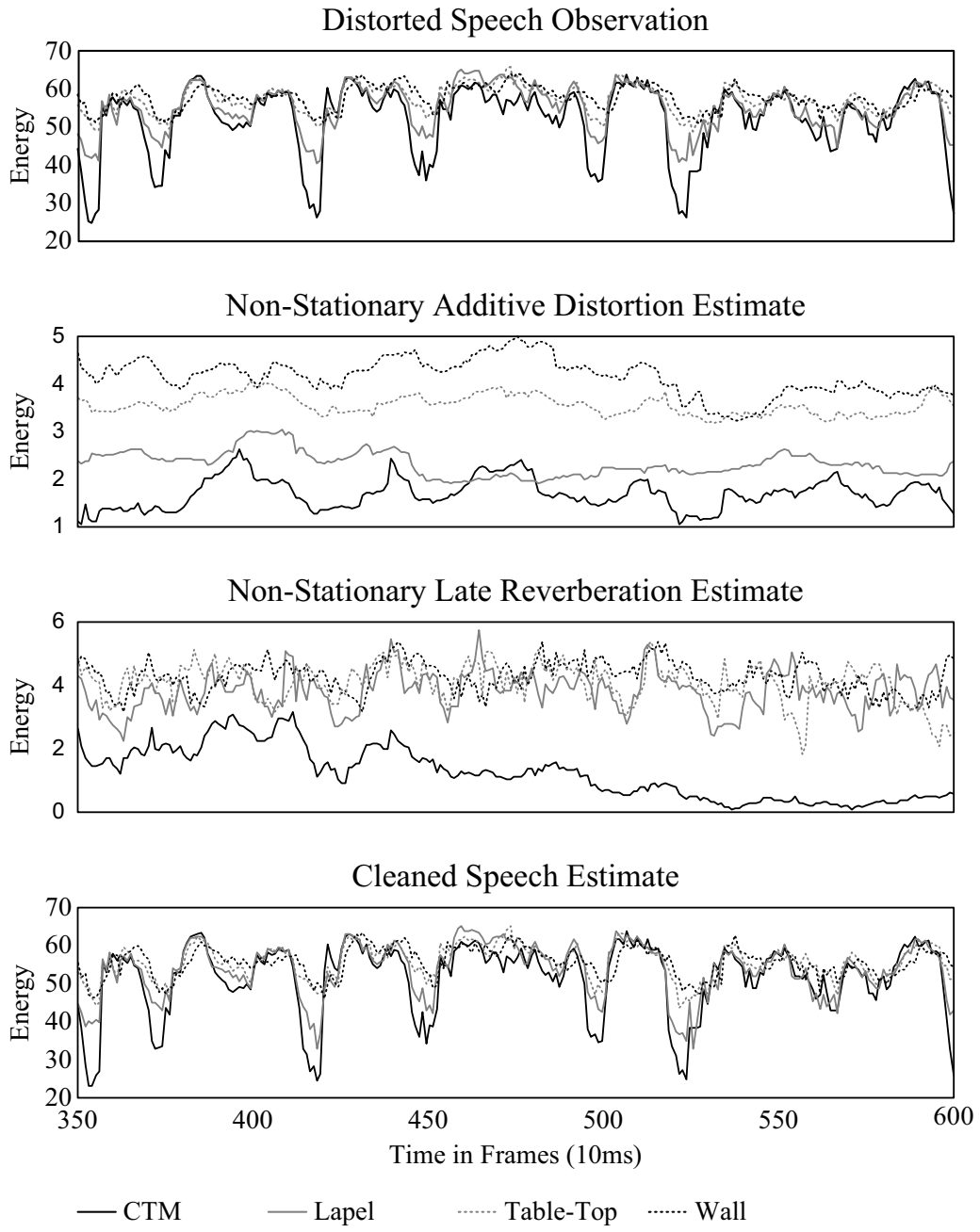


Figure 10.1: Average energies over all frequency bands vs. time of the distorted speech frames, the additive distortion estimate frames, late reverberation estimate frames and cleaned speech estimate frames.

Tabel 10.11 compares the correlation between the distorted signal and the two estimated distortions averaged over all frequency bands in the logarithmic frequency domain. The correlation between the distorted signal and the additive distortion estimate is between 10 and 32 percent for the different channels. Those values are surprisingly high as background noise is assumed to be uncorrelated to the clean speech signal. The correlation (delay adjusted) between the distorted signal and the late reverberation is always above 50. This is not surprising as the late reverberation is assumed to be a delayed version of the signal itself. The correlation (delay adjusted) between the additive distortion estimate and the late reverberation estimate increases by taking the microphone away from the speaker, except for the CTM, and is in average below the correlation between the distorted signal and the additive distortion.

10.6.2 Experiments English Set

Since it has been shown that the *mel-frequency cepstral coefficients* (MFCC)s are outperformed by warped MVDR cepstral coefficients, in distorted conditions in combination with and without speech feature enhancement, see Section 10.5, we decided to use the warped MVDR front-end exclusively for the following speech enhancement experiments.

The *Gaussian mixture model* (GMM) representing clean speech in the *model-combination-based acoustic mapping* (MAM) [203] as well as in the PF has been trained with 64 Gaussians on the same acoustic training material as used to train the acoustic model of the ASR system. For the second pass experiments the GMM has been trained on features which have been normalized by VTLN. The GMM of noise to initialize the PF has been trained for each individual utterance on silence regions found by voice activity detection. To train the static *autoregressive* (AR) matrix we have collected approximately 150 seconds of noise only pieces on the test set, found by voice activity detection, prior to the actual system evaluation. To let the PF settle it has been given two extra seconds in front of each utterance.

The optimal step-size D, in *multi-step linear prediction* (MSLP), has been set to 60 ms for all utterances. This value has been determined on additional acoustic material and has not appeared to be strongly dependent on the acoustic conditions such as different channel types or distance between the microphone and the speaker's mouth. This is in contrast to the 1000 LP coefficients $\mathbf{c}$ which are strongly utterance and channel dependent and thus have been estimated individually for each utterance and channel.

Microphone		CTM		Lapel		Table-Top		Wall	
Distance		5 cm		20 cm		1.5–2 m		3–4 m	
SNR		24 dB		23 dB		17 dB		10 dB	
Pass		1	2	1	2	1	2	1	2
Prediction	Compensation	Word Error Rate %							
no	no	11.2	9.1	10.9	9.2	18.6	14.0	45.4	28.6
no	MAM[1]	10.9	10.4	11.2	10.2	18.8	14.3	43.5	29.0
no	MAM[2]	10.9	10.2	10.9	10.1	18.5	14.1	45.1	27.9
random walk	GMA	11.8	9.4	12.1	9.2	20.9	15.4	49.2	31.6
random walk	SIA	11.6	9.4	12.0	9.2	20.1	15.0	48.6	29.6
static AR[3]	GMA[3]	10.9	9.2	11.3	9.6	18.6	13.7	44.2	26.9
static AR	SIA	10.8	8.9	11.2	9.4	18.5	13.2	42.5	25.3
dynamic AR	GMA	10.8	9.0	11.0	9.2	17.3	13.1	43.5	25.3
dynamic AR	SIA	10.6	9.0	10.7	9.0	17.8	13.2	42.8	25.4

Table 10.12: Word error rates for no compensation, static compensation (lines 2 and 3) and different particle filter enhancement techniques (lines 4 to 9) for different speaker to microphone distances.

[1] *original approach of model-combination-based acoustic mapping.*
[2] *model-combination-based acoustic mapping using a vector Taylor series approximation as suggested in Section 6.1.*
[3] *baseline particle filter as proposed by Raj et al. [179].*

We start our speech recognition experiments by comparing the WERs for no compensation, static compensation by MAM and different PF variants as previously described. Comparing static compensation, line two in Table 10.12, with dynamic compensation techniques, lines three, five and seven, it becomes obvious that the capacity of tracking the noise is able to improve over static compensation. It is clear from Table 10.12 that the random work (6.17) is significant worse than the two investigated predictive walk methods on all microphone conditions. The dynamic AR model (6.18) yields small improvements over the static AR model (6.19) which are more pronounced if the additive distortion is compensated by the GMA (6.24). The *statistical interference approach* (SIA) (6.25) provides higher reduction in error rates than the *Gaussian mixture approach* (GMA) for random and predicted walks with a static AR matrix, while in the case of a dynamic AR matrix the performance of the two methods is nearly alike. In average SIA outperforms GMA and the dynamic AR model outperforms the static AR model. Thus, for the experiments presented in Table 10.13, the PF uses a dynamic AR matrix to predict the noise and SIA to compensate for the noise exclusively. Note that the static AR model can not be directly applied in our further investigations as the additive distortion term

can not be determined a-priori in the joint framework.

Microphone			CTM		Lapel		Table-Top		Wall	
Distance			5 cm		20 cm		1.5–2 m		3–4 m	
SNR			24 dB		23 dB		17 dB		10 dB	
Pass			1	2	1	2	1	2	1	2
Front-End	Compensation Additive	Rev.	Word Error Rate %							
Fourier	no	no	11.3	9.5	12.3	10.3	18.0	14.2	45.9	30.0
Warped MVDR	no	no	11.2	9.1	10.9	9.2	18.6	14.0	45.4	28.6
Warped MVDR	yes[1,2]	no	10.6	9.0	10.7	9.0	17.8	13.2	42.8	25.4
Warped MVDR	no	yes[3]	14.4	9.5	15.1	9.6	17.7	13.4	39.2	23.9
Warped MVDR	yes[2]	yes[3]	12.1	9.3	13.4	9.5	17.7	13.3	38.3	23.3
Warped MVDR	joint 1		11.7	9.3	11.8	9.3	17.4	12.8	37.9	22.7
Warped MVDR	joint 2		11.5	8.6	11.9	9.0	16.9	12.6	38.4	22.2

Table 10.13: Word error rates without compensation and with different compensation approaches for different speaker to microphone distances of an English speaker.

[1] *identical to line 9 in Table 10.12*
[2] *additive distortion compensation by particle filter*
[3] *convolutive distortion compensation by multi-step linear prediction*

Comparing the first two lines in Table 10.13 confirms that the warped MVDR front-end outperforms the Fourier front-end. Thus the warped MVDR front-end is exclusively used in our further experiments. Comparing the second with the third line we observe that the compensation of non-stationary additive distortions by the PF is able to improve the recognition performance in all investigated cases. This comes as a little surprise, as it was not expected that the nearly clean CTM and lapel microphone can profit from enhancement techniques. Compensating for the reverberation using MSLP, the fourth line, we see a different picture: On the close-talk and lapel microphones, where no reverberation is expected, the word accuracy collapses if the acoustic models are not adapted. If unsupervised adapted this collapse is partly compensated. On the table-top microphone the reductions in WER are comparable to those of the PF. On the wall mounted microphone, where more reverberation is expected, MSLP is able to significantly outperform the PF approach. Both approaches, the PF as well as MSLP, are able to compensate for distortions which can not be treated exclusively by MLLR and constrained MLLR. This is apparent by comparing the second pass results. Applying both approaches, MSLP followed by PF, can either keep or further lower the error in the cases where the speech signal is significantly distorted. On the close-talk and lapel microphones the PF

can compensate for some distortions introduced by MSLP. The last two lines in Table 10.13 present results for the proposed joint approach. While *joint 1* shares a single scaling term as determined by (8.4), *joint 2* in addition has a tilt term as determined by (8.5). It is clear upon comparing the two variants that the introduction of tilt improves the recognition accuracy over the first variant. This approach leads to the best accuracy (equal on the lapel microphone) on all channels after unsupervised model adaptation. Note that this is in contrast to a variety of feature enhancement techniques which improve the accuracy on distorted signals, however are deemed to reduce the accuracy on distortion free signals; e.g. the MSLP approach. Thus the proposed joint approach can be applied without constrains to all microphone conditions.

10.6.2.1 Timing Studies

In this section we compare the runtime performance of two systems, namely the baseline system with the warped MVDR front-end without feature enhancement and a system using feature enhancement based on the proposed approach (joint 2). The runtime factors have been measured on an Intel Xeon processor with 3.2 GHz and 3.5 GByte of RAM running on SUSE Linux 10.3.

Table 10.14 presents the runtime factors on different microphones which belong to different speaker to microphone distances respectively. We can observe, by comparing the first with the second row for the first pass (p1) or adaptation (adp), that the whole feature enhancement process is running in roughly real time with similar execution times for all channels. Due to more difficult environment the decoding in the ASR system takes longer for channels which are further away from the speaker's mouth. While the feature enhancement step nearly doubles the computation time on the CTM, on the wall microphone the execution time is increased by less than 30%. If speed is the mayor requirement, one could remove the second run of the enhancement step with only a small loss in word accuracy on the second pass (p2).

Microphone	CTM			Lapel			Table-Top			Wall		
Distance	5 cm			20 cm			1.5–2 m			3–4 m		
Step	p1	adp	p2	p1	adp	p2	p1	adp	p2	p1	adp	p2
Compens.	Real Time Factor											
no	1.18	0.31	0.94	1.48	0.30	0.89	2.70	0.29	1.34	4.71	0.32	2.31
joint 2	2.34	1.28	0.91	2.63	1.14	1.03	3.77	1.14	1.51	5.85	1.28	2.23

Table 10.14: Timing experiments of the different steps in the speech recognition system for different speaker to microphone distances.

10.6.3 Experiments German Set

The German acoustic model has been trained the same way as the English acoustic model. Due to the lack of proper acoustic material we have trained a speaker dependent system using 10 hours of acoustic training material.

Similar to the English system we have collected data in the Internet for language modeling. This data has been merged with corpora taken from inhouse lecture transcriptions, news, talks, presentation slides and concept papers. In contrast to the English system only a limited number of technical texts is available which makes it more difficult to find adequate data. In addition the German language has a large amount of compound words and inflections. For a better vocabulary coverage and a more robust estimation of the statistical language model we used compound splitting based on a big German vocabulary containing all possible inflection forms. The variables in the enhancement algorithms have not been altered and thus are identical to the description in Section 10.6.2.

Comparing the first two lines in Table 10.15 once more confirms that the warped MVDR front-end is leading to better results than the Fourier front-end (the classical MFCC features), in particular for severe acoustic environments. Comparing the second with the third line we observe that the PF is not always able to improve the recognition performance as observed on the English data. As already mentioned before, the German recording is different as less additive non-stationary distortions are present. As a consequence the improvements by the PF vanish in the second pass.

Compensating for the reverberation using MSLP, the fourth line, we observe that the close-talk and lapel microphones are not as much degraded in performance as the English system. This can, once more, be explained by the fact that less additive distortions are present in the German recordings and thus the estimate of the MSLP is more accurate. While the improvements by dereverberation vanish in the second pass on the Lapel and table-top microphones it is still present on the wall mounted microphone. Applying both approaches, MSLP followed by PF, can not lower the error rate as MSLP. This can be explained by the poor performance of the PF. The last two lines in Table 10.15 present results for the proposed joint approach. Those approaches are able to further lower the WER. Those reductions show also on the second pass for the Lapel, table-top and wall mounted microphones however vanish on the CTM.

To conclude, we can confirm the trend already seen on the English test set. The gains observed, however, are smaller as the introduced distortions are not as severe as in English. This is also reflected in the WER. Comparing the different channels for the English and the German test set we observe that the

English system has a lower WER on CTM, while it has a higher WER on the wall mounted microphone. Comparing the presented results with a cheating experiment where the text transcripts have been added to the language model we have observed that a better language model is leading to higher reductions in WER.

Microphone			CTM		Lapel		Table-Top		Wall	
Distance			5 cm		20 cm		1.5–2 m		3–4 m	
SNR			26 dB		23 dB		17 dB		12 dB	
Pass			1	2	1	2	1	2	1	2
Front-End	Compensation Additive	Rev.	Word Error Rate %							
Fourier	no	no	13.5	13.6	14.0	14.0	27.0	21.1	40.8	26.9
Warped MVDR	no	no	14.0	13.8	13.6	13.6	27.2	20.3	39.9	25.1
Warped MVDR	yes[1]	no	13.4	13.4	13.4	13.7	26.7	20.6	38.3	26.3
Warped MVDR	no	yes[2]	14.0	13.7	14.2	13.6	27.0	20.8	37.6	24.8
Warped MVDR	yes[1]	yes[2]	13.9	13.9	14.2	13.9	28.1	21.5	37.3	25.9
Warped MVDR	joint 1		13.6	13.5	13.9	13.6	24.6	20.1	35.3	25.1
Warped MVDR	joint 2		13.4	13.6	13.2	13.1	24.6	19.7	35.1	24.6

Table 10.15: Word error rates without compensation and with different compensation approaches for different speaker to microphone distances of a German speaker using a weak language model.

[1] *additive distortion compensation by particle filter*
[2] *convolutive distortion compensation by multi-step linear prediction*

10.7 Acoustic Channel Selection

In order to evaluate the performance of the proposed channel selection method, see Section 9, we have first investigated the collected English and German lecture data. Due to the small channel quality difference between two channels and large channel quality difference to the other channels those experiments have not been able to exhibit significant differences in WER in comparison to the selection of the channel by SNR. We have, therefore, turned our attention to a more difficult problem where a couple of microphones with nearly similar signal quality has been available.

10.7.1 Lecture Data on Different Channel Types

Figure 10.2 plots the number of utterances selected for each channel and selection method. As we can observe the SNR as well as the class separability measure is choosing most of the time either the CTM or the lapel microphone channel. Only very seldom the table-top or wall mounted microphone is selected. Investigating the cases where either the table-top or wall mounted microphone is selected reveals that in most of the cases an empty utterance has been analyzed and thus an influence on the recognition performance is not expected. This is confirmed by baseline experiments. A decoding over the automatically selected channel is leading to the identical WER for class separability as well as SNR of 11.2%, which is also identical to the CTM channel. This is not surprising as the used microphone is nearly exclusively selected from either the CTM or the lapel microphone which have a similar WER. The experiments performed on German bear no surprise, the CTM reference channel has a WER of 13.3% (on the weak LM) while the automatically selected channel provids a WER of 13.4% for class separability as well as SNR.

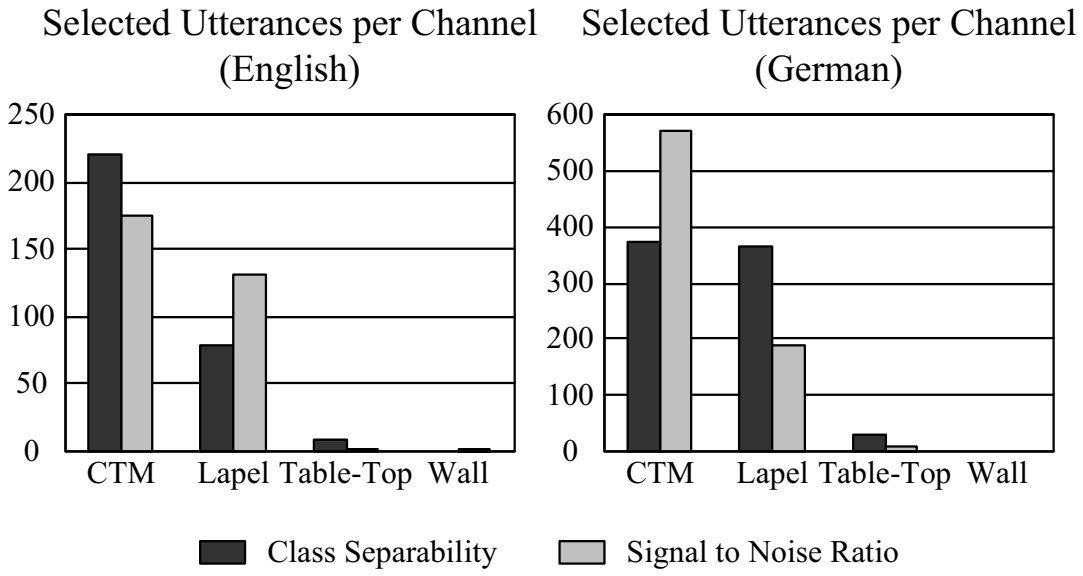


Figure 10.2: Number of utterances for each channel selected by either class separability or signal to noise ratio for English and German lectures.

10.7.2 Speech Recognition Experiments on NIST's RT-07 Lecture Meeting Data

NIST's RT-07 lecture meeting data [160] contain multiple distant microphone recordings and therefore enable the realistic evaluation of multi-source far-distant speech recognition technologies. The far-field data captured by table-top microphones are exacerbated, in comparison to CTM recordings, by the much poorer acoustic signal quality caused by reverberation, background noise and overlapping speech.

The severe acoustic condition reveals differences in the way the class separability is used. On preliminary experiments four observations are made:

- Direct comparisons between (9.4) and (9.5) have showed a small difference in accuracy, where (9.4) has always been ahead. Therefore, our further investigations are limited to (9.4).
- Classes which have been determined on the investigated utterance (on the fly) have always led to slightly higher recognition errors as compared to classes which have been predetermined on the training data (identical to the acoustic training data for the acoustic models of the ASR system). In addition, on the fly classes take longer to process. Therefore, our further investigations are limited to predetermined classes.
- The knowledge of the vocal tract length, determined by the ASR system, can also be considered [109] in the calculation of the scatter matrices and is leading to slightly different scores, which, in some cases, might lead to the selection of a different channel. However, we found that it has a minor effect on the classification result and therefore is not treated in the experiments separately—on first pass experiments no information about the vocal tract length is available, on second and third pass experiments the vocal tract length has always been considered.
- Experiments with different number of classes in the scatter matrix have led to slightly different accuracy. On our data set we found that eight classes are leading to the best classification results.

Comparing the WERs provided by different channel selection techniques in Table 10.16 we observe that any of the investigated *class separability measures* (CSM)s are superior to SNR. Note that for the decoder based approach the phone classes are derived by a forced alignment on hypothesis of a previous pass and thus a first pass experiment can not be performed. Comparing the third pass, we observe an absolute difference of 4.4% which is a relative improvement

Channel Selection	Word Error Rate %		
Pass	1	2	3
Signal to Noise Ratio	73.0	62.3	59.5
Class Separability - stand alone[1,3]	68.6	59.1	56.7
Class Separability - stand alone[2,3]	68.1	58.4	55.9
Class Separability - stand alone[2,4]	67.4	57.8	55.1
Class Separability - decoder based[1,3]	–	58.5	57.1

Table 10.16: Influence of different channel selection techniques, signal to noise and a variety of class separability, on the word error rates.

[1] *class selection on combined channel*
[2] *class selection on individual channels*
[3] *classes on all frames*
[4] *classes only on speech frames*

of 7.4%. Taken the CTM performance as a lower bound, 31.3%, we gain back 15.6% of the accuracy lost by using multi-channel distant microphones by replacing SNR channel selection with the proposed CSM channel selection. Note that even though CSM based methods take a little bit longer to compute as SNR based methods, the reported improvements are established with an overall *decrease* in computation time, as decodings (which eat up most of the computation) run faster on channels with a better quality.

Comparing stand alone and decoder based CSM approaches we observe that the decoder based approach is not improving the stand alone approach. This might be a bit surprising, possible reasons could be the high number of 46 classes as determined by the number of phonemes and that the decoding has only be performed on one channel, resulting in a mismatch if evaluated on other channels. For an improved performance one could run decodings for each channel, as recommended in decoder based methods, and/or cluster the phonemes to reduce the number of classes.

Comparing between the different stand alone CSM approaches we can conclude that each channel should be treated separately and that the performance has improved by ignoring the silence class.

A direct comparison between delay-and-sum channel combination and the proposed channel selection technique on the final pass of the RT07 evaluation system [44] with two front-ends (the described and a warped-twice MVDR front-end [35]) shows a relative improvement of 3.6%, from 52.4% to 50.5% WER.

Chapter 11

Conclusion and Outlook

This sections reviews the mayor contributions of this thesis and highlights developments by other researchers who followed ideas or algorithms presented in previous publications by the present author or in this publication.

The proposed warped *minimum variance distortionless response* (MVDR) front-end has been proved successful in a broad number of applications and many evaluations, e.g. [159]. On clean data it provides at least the same accuracy as the *mel-frequency cepstral coefficient* (MFCC) or perceptual linear prediction front-end and is always able to improve the accuracy in more challenging environments. Thus, the warped MVDR front-end has replaced the MFCC front-end—which has nearly been used exclusively in the Janus Recognition Toolkit—to extract speech cues in various languages such as English, German and Spanish. It needs to be determined if the warped MVDR front-end is also superior for tonal languages such as Mandarin. Additional gains in accuracy are possible by combining the warped MVDR front-end with different front-ends by confusion network combination; e.g., [24, 53].

Recently published work by Dharanipragada *et al.* [91] using perceptual motivated MVDR methods on the matched Aurora-2 task, the Wall Street Journal task and the Switchboard task support some of our findings. Following our approach Chen *et al.* [80] have combined the proposed warped MVDR feature extraction with feature normalization techniques; namely *progressive his-*

togram normalization. By comparison to MFCC features they concluded that "The results indicated that both the MVDR-based features and the normalization processes are very helpful." Muralishankar and O'Shaughnessy [149] have compared the phoneme accuracy for warped MVDR cepstrum, *warped discrete Fourier transform cepstrum* (WDFTC) [150] and MFCC on six different noise conditions at various SNRs and concluded that "In general, it can be easily concluded on the basis of all the results presented above that the WDFTC and PMVDR outperform MFCC in different noise types and SNRs. [...] It may be useful to note that the better performance of the PMVDR and WDFTC are attributed to their noise robustness and lower feature variance."

The proposed warped MVDR features have been compared to other features for capturing timbral information of music signals in connection with genre classification applications and concluded that "MFCCs based on fixed order, signal independent linear prediction and MVDR spectral estimators did not exhibit any statistically significant improvement over MFCCs based on the simpler Fourier transform." [121]. The equal performance of the different features in their experiments is probably due to a very simple acoustic model. It would be interesting to repeat the experiments with more advanced acoustic models.

As the information needed to discriminate between different phonemes is provided in different frequency regions we have followed Nakatoh *et al.* [154] who have used two bilinear transformations in combination with linear prediction to steer spectral resolution to lower or higher frequencies. We have, however, applied the two bilinear transformations on MVDR spectral envelopes. On noisy data we have observed a better class separability and a decreased word error.

We have demonstrated that the optimal choice of the model order depends on the fundamental-frequency of the speaker and that the correct choice of the model order per speaker can help to lower the word error rate. This has probably driven Hegde *et al.* [116] to develop a robust approach for modeling voiced speech by combining a family of MVDR estimates (MVDR estimates of different orders).

To compensate for non-stationary additive distortions we have adopted a particle filter approach proposed by Sing and Raj [187]. We have improved their original approach by replacing the vector Taylor series by the statistical interference approach which led to significant performance improvements. The introduction of the dynamic autoregressive process eliminates the requirement to learn the prediction matrix prior to the application of the particle filter.

While feature enhancement techniques have been developed to compensate for either additive or convolutive distortions we have developed a particle filter framework which is capable to track and to remove non-stationary additive dis-

tortions and late reverberation with a non-stationary room impulse response. In a series of experiments with different speaker to microphone distances we have demonstrated that compensating for additive as well as for convolutive distortions helps to improve the accuracy of an automatic speech recognition system. Furthermore we have been able to gain in accuracy by jointly estimating additive and convolutive distortions over their individual estimation. In addition we have argued that the compensation of non-stationary distortions in the feature space is able to compensate for distortions which can not be treated well with those techniques assuming stationary distortions. This argument is confirmed by demonstrating additional improvement in word accuracy by combining the proposed joint particle filter with feature and model adaptation techniques.

Last but not least we have proposed to use class separability as a measure for channel quality. While no significant performance difference could have been observed in contrast to signal-to-noise ratio on channels with different characteristics such as close talk, table-top and wall mounted microphones, significant improvements could be demonstrated in those cases where the channel quality was very similar; e.g. between different table-top microphones.

A combination of the proposed techniques can lead to—on realistic recordings in noisy and reverberant environments—relative reductions in WER by up to 26.0% compared to the mel-frequency cepstral coefficient front-end without feature enhancement after unsupervised acoustic model adaptation.

The development of the algorithms, except for the warped MVDR front-end; e.g. on Switchboard [50], has exclusively focused on the automatic transcription of lectures in clean and demanding acoustic environments. Most of the developed techniques improve the robustness of the features against acoustic distortions and signal variations. Thus, the proposed algorithms might be useful in other scenarios. For example the particle filter, as developed in this thesis, can be readily applied to remove the motor noises of a humanoid robot, see [10].

It would be also interesting to investigate if the proposed feature extraction and enhancement techniques are useful in related fields such as acoustic speaker verification or classification.

Glossary

Notational Convention

$\mathbf{a}, \mathbf{b}, \mathbf{c}, \cdots$	all vectors are column vectors and written in boldface
a_i	i^{th} element of $\mathbf{a}$
$a^{(i)}$	vector/prediction of order i
$\mathbf{a}^T$	transposition operator
$\mathbf{a}^H$	conjugate transposition operator
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \cdots$	all matrices are capitalized and written in boldface
$A_{i,j}$	$[i, j]^{\text{th}}$ element of $\mathbf{A}$
$\mathbf{A}^T$	transpose of matrix
a^*	complex conjugate of a
$\hat{a}$	estimate of a
$\tilde{a}$	warped value of a
$(\cdot)$	continuous
$[\cdot]$	discrete
$/\cdot/$	denotes a phoneme
%	modulo

Principal Symbols

α	warp parameter or forgetting factor
$\boldsymbol{\mu}$	mean vector

ω	angular frequency, $\omega = 2\pi f$
φ	normalized first autocorrelation coefficient
γ	mixture weight
ξ	smoothing parameter
$\boldsymbol{\phi}$	correlation matrix
$\boldsymbol{\Sigma}$	covariance matrix
a	additive distortion
$\mathbf{a}_k$	additive distortion vector
$\mathbf{A}$	linear prediction matrix, transformation matrix
b	spectral bin
c	linear prediction coefficient, class
d	distance or distortion
$\mathbf{d}_k$	distortion "super" vector
D	delay
e	error term
$\mathcal{E}\{\cdot\}$	estimation
f	frequency
$f(\cdot)$	transfer function
$f_{\text{mode}}(x, y, z)$	room modes
$g(\cdot)$	observation function
h	impulse response
H	transfer function
k	frame index
l_x, l_y, l_z	dimensions
L	sound pressure level or number of harmonics
n	noise signal or discrete time index
$\mathcal{N}(x; \mu, \Sigma)$	Gaussian distribution
$p(\mathbf{x})$	prior distribution of speech
$p(\mathbf{x}_{k+1}\|\mathbf{x}_k)$	(state) transition probability, evolution
$p(\mathbf{x}_k\|\mathbf{y}_{1:k})$	filtering density
$p(\mathbf{y})$	prior distribution of speech
$p(\mathbf{y}_k\|\mathbf{x}_k)$	output probability, likelihood function
$p(x, y, z)$	sound pressure
r	reflection sequence
$\mathbf{r}_k$	reverberation vector
R	correlation coefficients
s	scale term
$\mathbf{s}_k$	reverberation scale vector
S	power spectrum
$\mathbf{S}$	scatter matrix
t	continuous time
u	excitation signal
$\mathbf{u}_k$	process noise
w	weight

$\mathbf{w}_k$	measurement noise
W	weighting filter, word string
x	clean signal, input signal
$\mathbf{x}_k$	state vector
y	distorted signal (noise and reverberation), output signal
$\mathbf{y}_k$	observation vector

Abbreviations

AR	Autoregressive
ASR	Automatic Speech Recognition
CDCN	Codeword-Dependent Cepstral Normalization
CHIL	Computers in the Human Interaction Loop
CMLLR	Constrained Maximum Likelihood Linear Regression
CMN	Cepstral Mean Normalization
CMU	Carnegie Mellon University
CNC	Confusion Network Combination
CSM	Class Separability Measures
CTM	Close-Talking Microphone
dB	deciBel
DCT	Discrete Cosine Transformation
EPPS	European Parliament Plenary Session
FCDCN	Fixed Codeword-Dependent Cepstral Normalization
GMA	Gaussian Mixture Approach
GMM	Gaussian Mixture Model
GSM	General Speech Model
HMM	Hidden Markov Model
IDCT	Inverted Discrete Cosine Transformation
LM	Language Model
LP	Linear Prediction
LPC	Linear Prediction Coefficient
MA	Microphone Array
MAM	Model-combination-based Acoustic Mapping
MDE	Minimum Discriminant Estimation
MFCC	Mel-Frequency Cepstral Coefficients
MLLR	Maximum Likelihood Linear Regression
MM	Mixture Model
MMSE	Minimum Mean Squared Error
MO	Model Order
MSLP	Multi-Step Linear Prediction

MVDR	Minimum Variance Distortionless Response
NIST	National Institute of Standards and Technology
OCW	Open Course Ware
OOV	Out Of Vocabulary
OYC	Open Yale Courses
pdf	probability density function
PF	Particle Filter
PLP	Perceptual Linear Predictive
PLSA	Probabilistic Latent Semantic Analysis
PMVDR	Perceptual Minimum Variance Distortionless Response
PSM	Phoneme-Specific speech Model
RER	Relative Error Reduction
SIA	Statistical Inference Approach
SNR	Signal-to-Noise Ratio
SPLICE	Stereo-Based Piecewise Linear Compensation for Environments
TC-STAR	Technology and Corpora for Speech to Speech Translation
TED	Translanguage English Database
UKA	Universität Karlsruhe
VTLN	Vocal Tract Length Normalization
VTS	Vector Taylor Series
W2MVDR	Warped-Twice Linear Prediction
W2MVDR	Warped-Twice Minimum Variance Distortionless Response
WDFTC	Warped Discrete Fourier Transform Cepstrum
WER	Word Error Rate
WMVDR	Warped Minimum Variance Distortionless Response

Bibliography

Conferences

ASRU	Automatic Speech Recognition and Understanding
ESSV	Elektronische Sprachsignalverarbeitung
Eurospeech	European Conference on Speech-Communication and Technology
EUSIPCO	European Signal Processing Conference
HSCMA	Hands-free Speech Communication and Microphone Arrays
ICASSP	IEEE International Conference on Acoustic, Speech, and Signal Processing
ICME	IEEE International Conference on Multimedia & Expo
ICMI	International Conference on Multimodal Interfaces
ICSLP	International Conference on Spoken Language Processing
Interspeech	Fusion of the two biyearly conferences Eurospeech and ICSLP
IROS	EEE/RSJ International Conference on Intelligent RObots and Systems
IWSLT	International Workshop on Spoken Language Translation
LREC	International Conference on Language Resources and Evaluation
MLMI	Machine Learning for Multimodal Interaction
SLT	IEEE Workshop on Spoken Language Technology

Journals

ASA	Journal of the Acoustic Society of America
ASLP	IEEE Transactions on Audio, Speech and Language Processing
ASSP	IEEE Transactions on Acoustics, Speech and Signal Processing
SAP	IEEE Transactions on Speech and Audio Processing

My Publications

[1] S. Burger, K. Laskowski, and M. Wölfel. A comparative cross-domain study of the occurrence of laughter in meeting and seminar corpora. *6th International Conference on Language Resources and Evaluation (LREC2008)*, 2008.

[2] M. Dambier, M. Wölfel, and C. Fügen. Robuste Spracherkennung im Cockpit von Luftfahrzeugen. *Proc. of ESSV*, 2004.

[3] M. Dambier, M. Wölfel, and J. Hinkelbein. Spracherkennung im Cockpit. *Fliergermagazin 3/2005*, 2005.

[4] F. Faubel and M. Wölfel. Coupling particle filters with automatic speech recognition for speech feature enhancement. *Proc. of Interspeech*, 2006.

[5] F. Faubel and M. Wölfel. Overcoming the vector tailor series approximation in speech feature enhancement – a particle filter approach. *Proc. of ICASSP*, 2007.

[6] C. Fügen, S. Ikbal, F. Kraft, K. Kumatani, K. Laskowski, J. McDonough, M. Ostendorf, S. Stüker, and M. Wölfel. The ISL RT-06S speech-to-text system. *In Proc. of the Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop (RT-06S), Bethesda, MD, USA*, 2006.

[7] C. Fügen, M. Wölfel, S. Ikbal, F. Kraft, K. Laskowski, M. Ostendorf, S. Stüker, and K. Kumatani. Advances in lecture recognition: The ISL RT-06S evaluation system. *Proc. of Interspeech*, 2006.

[8] T. Gehrig, U. Klee, J. McDonough, S. Ikbal, M. Wölfel, and C. Fügen. Tracking and beamforming for multiple simultaneous speakers with probabilistic data association filters. *Proc. of Interspeech*, 2006.

[9] M. Kolss, M. Wölfel, M. Kraft, J. Niehues, M. Paulik, and A. Waibel. Simultaneous german-english lecture translation. *Proc. of IWSLT*, 2008.

[10] F. Kraft and M. Wölfel. Humanoid robot noise suppression by particle filters for improved automatic speech recognition accuracy. *Proc. of IROS*, 2007.

[11] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel. Adaptive beamforming with a minimum mutual information criterion. *Trans. on ASLP*, 15:2527–2541, 2007.

[12] K. Kumatani, J. McDonough, U. Mayer, T. Gehrig, E. Stoimenov, and M. Wölfel. Minimum mutual information beamforming for simultaneous active speakers. *Proc. of ASRU*, 2007.

[13] K. Laskowski, M. Wölfel, M. Heldner, and J. Edlund. Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems. *155th Meeting of the Acoustical Society of America, 5th EAA Forum Acusticum, and 9th SFA Congrés Français d'Acoustique (Acoustics2008)*, 2008.

[14] D. Macho, J. Padrell, A. Abad, C. Nadeu, J. Hernando, J. McDonough, M. Wölfel, U. Klee, M. Omologo, A. Brutti, P. Svaizer, G. Potamianos, and S.M. Chu. Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus. *Proc. of ICME*, 2005.

[15] J. McDonough, K. Kumatani, T. Gehrig, E. Stoimenov, U. Mayer, S. Schacht, M. Wölfel, and D. Klakow. To separate speech! A system for recognizing simultaneous speech. *Proc. of MLMI*, 2007.

[16] J. McDonough and M. Wölfel. Distant speech recognition: Bridging the gaps. *Proc. of HSCMA*, 2008.

[17] J. McDonough, M. Wölfel, K. Kenichi, R. Barbara, F. Faubel, and D. Klakow. Distant speech recognition: No black boxes allowed. *Proc. of Sprachkommunikation 2008, 8. ITG-Fachtagung*, 2008.

[18] J. McDonough, M. Wölfel, and E. Stoimenov. On maximum mutual information speaker-adapted training. *Comput. Speech Lang.*, 22(2):130–147, 2008.

[19] F. Metze, P. Gieselmann, H. Holzapfel, T. Kluge, I. Rogina, A. Waibel, M. Wölfel, J. Crowley, P. Reignier, D. Vaufreydaz, F. Berard, B. Cohen, J. Coutaz, S. Rouillard, V. Arranz, M. Bertran, and H. Rodriguez. The "fame" interactive space. *Proc. of MLMI*, 2005.

[20] S. Ochs, M. Wölfel, and S. Stüker. Verbesserung der automatischen Transkription von englischen Wörtern in deutschen Vorlesungen. *Proc. of ESSV*, 2008.

[21] G. Potamianos, L. Lamel, M. Wölfel, J. Huang, E. Marcheret, C. Barras, X. Zhu, J. McDonough, J. Hernando, D. Macho, and C. Nadeu. Automatic speech recognition. *In Computers in the Human Interaction Loop by A. Waibel and R. Stiefelhagen (Editors), Springer*, pages 271–287, 2009.

[22] D. Raub, J. McDonough, and M. Wölfel. A cepstral domain maximum likelihood beamformer for speech recognition. *Proc. of ICSLP*, 2004.

[23] R. Stiefelhagen, K. Bernardin, H.K. Ekenel, J. McDonough, K. Nickel, M. Voit, and M. Woelfel. Audio-visual perception of a lecturer in a smart seminar room. *Signal Processing - Special Issue on Multimodal Interfaces, Elsevier*, 86(12), Dec. 2006.

[24] S. Stüker, C. Fügen, S. Burger, and M. Wölfel. Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end. *Proc. of Interspeech*, 2006.

[25] S. Stüker, C. Fügen, K. Florian, and M. Wölfel. The ISL 2007 english speech transcription system for european parliament speeches. *Proc. of Interspeech*, 2007.

[26] S. Stüker, C. Fügen, R. Hsiao, S. Ikbal, Q. Jin, F. Kraft, M. Paulik, M. Raab, Y.C. Tam, and M. Wölfel. The ISL TC-STAR spring 2006 ASR evaluation systems. *In Proc. of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain*, 2006.

[27] A. Waibel, K. Bernardin, and M. Wölfel. Computer-supported human-human multilingual communication. *Proc. of Interspeech*, 2007.

[28] A. Waibel, K. Bernardin, and M. Wölfel. Computer-supported human-human multilingual communication. *in 50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence by M. Lungarella, F. Iida, J. Bongard and R. Pfeifer (Editors), Springer*, pages 271–287, 2007.

[29] M. Wölfel. Mel-Frequenzanpassung der Minimum Varianz Distortionless Response Einhüllenden. *Proc. of ESSV*, 2003.

[30] M. Wölfel. Minimum variance distortionless response spectral estimation and subtraction for robust speech recognition. *Diploma-Thesis, Universität Karlsruhe (TH), Karlsruhe, Germany*, Jan. 2003.

[31] M. Wölfel. Multiquellentraining: Chancen für kleine Trainingsmengen in der automatischen Spracherkennung. *Proc. of ESSV*, 2004.

[32] M. Wölfel. Speaker dependent model order selection of spectral envelopes. *Proc. of ICSLP*, 2004.

[33] M. Wölfel. Frame based model order selection of spectral envelopes. *Proc. of Interspeech*, 2005.

[34] M. Wölfel. Warped & warped-twice MVDR spectral estimation with and without filterbanks. *Proc. of MLMI*, 2006.

[35] M. Wölfel. Warped-twice minimum variance distortionless response spectral estimation. *Proc. of EUSIPCO*, 2006.

[36] M. Wölfel. Channel selection by class separability measures for automatic transcriptions on distant microphones. *Proc. of Interspeech*, 2007.

[37] M. Wölfel. Integration of the predicted walk model estimate into the particle filter framework. *Proc. of ICASSP*, 2008.

[38] M. Wölfel. A joint particle filter and multi-step linear prediction framework to provide enhanced speech features prior to automatic recognition. *Proc. of HSCMA*, 2008.

[39] M. Wölfel. Predicted walk with correlation in particle filter speech feature enhancement for robust automatic speech recognition. *Proc. of ICASSP*, 2008.

[40] M. Wölfel. Enhanced speech features by single channel joint compensation of noise and reverberation. *Trans. on ASLP*, 17(2):312–323, 2009.

[41] M. Wölfel. Signal adaptive spectral envelope estimation for robust speech recognition. *accepted for publication in Speech Communication*, 2009.

[42] M. Wölfel and S. Burger. The ISL baseline lecture transcription system for the TED corpus. *Technical Report TR0001, http://isl.ira.uka.de/~wolfel*, 2005.

[43] M. Wölfel and H.K. Ekenel. Feature weighted Mahalanobis distance: Improved robustness for Gaussian classifiers. *Proc. of EUSIPCO*, 2005.

[44] M. Wölfel and F. Faubel. Considering uncertainty by particle filter enhanced speech features in large vocabulary continuous speech recognition. *Proc. of ICASSP*, 2007.

[45] M. Wölfel, C. Fügen, S. Ikbal, and J. McDonough. Multi-source far-distance microphone selection and combination for automatic transcription of lectures. *Proc. of Interspeech*, 2006.

[46] M. Wölfel, M. Kolss, M. Kraft, J. Niehues, M. Paulik, and A. Waibel. Simultaneous machine translation of German lectures into English: Investigating research challenges for the future. *Proc. of SLT*, 2008.

[47] M. Wölfel and J. McDonough. Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination. *Proc. of Interspeech*, 2005.

[48] M. Wölfel and J. McDonough. Minimum variance distortionless response spectral estimation, review and refinements. *IEEE Signal Processing Magazine: Special Issue on Speech Technology and Systems in Human-Machine Communication*, 22(5):117–126, Sept. 2005.

[49] M. Wölfel and J. McDonough. *Distant Speech Recognition*. John Wiley & Sons, 2009.

[50] M. Wölfel, J. McDonough, and A. Waibel. Minimum variance distortionless response on a warped frequency scale. *Proc. of Eurospeech*, 2003.

[51] M. Wölfel, J. McDonough, and A. Waibel. Warping and scaling of the minimum variance distortionless response. *Proc. of ASRU*, 2003.

[52] M. Wölfel, K. Nickel, and J. McDonough. Microphone array driven speech recognition: Influence of localization on the word error rate. *Proc. of MLMI*, 2005.

[53] M. Wölfel, S. Stüker, and F. Kraft. The ISL RT-07 speech-to-text system. *In Proc. of the Rich Transcription 2007 Meeting Recognition Evaluation Workshop (RT-07), Baltimore, USA*, 2007.

Other Publications

[54] A. Acero. *Acoustical and environmental robustness in automatic speech recognition*. PhD thesis, Carnegie Mellon University, September 1990.

[55] A. Acero and R.M. Stern. Environmental robustness in automatic speech recognition. *Proc. of ICASSP*, pages 849—552, 1990.

[56] A Acero and R.M. Stern. Robust speech recognition by normalization of the acoustic space. *Proc. of ICASSP*, pages 893—896, 1991.

[57] AMI – Augmented Multi-party Interaction. http://www.amiproject.org.

[58] Andreas Andreou, Theresa Kamm, and Jordan Cohen. Experiments in vocal tract normalization. *Proc. of CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[59] X. Anguera, C. Wooters, and J. Hernando. Speaker diarization for multi-party meetings using acoustic fusion. *Proc. of ASRU*, 2005.

[60] B.S. Atal. Effictiveness of linear prediction characteristics on the speech wave for automatic speaker identificationand verification. *Jour. of ASA*, 55:1304–1312, 1974.

[61] J. Barker and M.P. Cooke. Modelling the recognition of spectrally reduced speech. *Proc. of Eurospeech*, pages 2127–2130, 1997.

[62] H.E. Bass, H.-J. Bauer, and L.B. Evans. Atmospheric absorption of sound: analytical expression. *Jour. of ASA*, pages 821–825, 1972.

[63] R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961.

[64] A.W. Black and P.A. Taylor. The festival speech synthesis system: System documentation. *Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK*, 1997.

[65] K. Boakye and A. Stolcke. Improved speech activity detection using cross-channel features for recognition of multiparty meetings. *Proc. of Interspeech*, 2006.

[66] S.F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Trans. on ASSP*, 27:113–120, Apr. 1979.

[67] C. Braccini and A. V. Oppenheim. Unequal bandwidth spectral analysis using digital frequency warping. *Trans. of ASSP*, 22:236–244, Aug. 1974.

[68] Lecture Browser. http://web.sls.csail.mit.edu/lectures.

[69] A. Brutti, M. Omologo, and P. Svaizer. Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays. *Proc. of Interspeech*, 2005.

[70] J.P. Burg. The relationship between maximum entropy and maximum likelihood spectra. *Geophysics*, 37:375–376, Apr. 1972.

[71] S. Burger. The CHIL RT07 Evaluation Data. R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans, Joint Proceedings of the Second International Evaluation workshop on Classification of Events, Activities and Relationships, CLEAR 2007 and the Spring 2007 Rich Transcription Meeting Evaluation*, Lecture Notes in Computer Science, No. 4625, Baltimore, USA, 2007. Springer.

[72] S. Burger and Z. Sloan. The isl meeting corpus: Categorical features of communicative group interactions. *Proc. of ICASSP: Meeting Recognition Workshop*, 2004.

[73] J.P. Campbell Jr. and D.A. Reynolds. Corpora for the evaluation of speaker recognition systems. *Proc. of ICASSP*, 1999.

[74] J. Capon. High-resolution frequency-wavenumber spectrum analysis. *Proc. of the IEEE*, 57:1408–1418, August 1969.

[75] O. Cappé. Elimination of the musical noise phonemenon with the ephraim and malah noise suppressor. *Trans. on SAP*, 2(2), April 1994.

[76] Press Release Carnegie Mellon University. Carnegie Mellon and University of Karlsruhe To Demonstrate Breakthroughs In Cross Lingual Communication and Speech-to-Speech Translation. *http* : $//www.cmu.edu/PR/releases05/051013_alex.html$, 2005.

[77] M. Cettolo, F. Brugnara, and M. Federico. Advances in the automatic transcription of lectures. *Proc. of ICASSP*, 2004.

[78] Research Channel. http://www.researchchannel.com.

[79] S.F. Chen and J. Goodman. An empirical study of smoothing. *Techniques for Language Modeling," TR-10-98, Computer Science Group, Harvard University*, 1998.

[80] Y. Chen and L.S. Lee. Robust features for speech recognition using MVDR spectrum estimation and feature normalization techniques. *Proc. of International Symposium on Chinese Spoken Language Processing*, 2004.

[81] CHIL – Computers In the Human Interaction Loop. http://chil.server.de.

[82] W.T. Chu and A.C.C. Warnock. Detailed directivity of sound fields around human talkers. *Technical Report, National Research Council Canada, IRC-RR-104*, 2002.

[83] W.T. Chu and A.C.C. Warnock. Voice and background noise levels measured in open offices. *Internal Report, National Research Council Canada, IR-837.*, 2002.

[84] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Trans. on ASSP*, 28:357–366, Aug. 1980.

[85] J.R. Deller, J.H.L. Hansen Jr., and J.G. Proakis. *Discrete-time processing of speech signals.* IEEE Press, 2000.

[86] J.R. Deller Jr., J.G. Proakis, and J.H.L. Hansen. *Discrete-time processing of speech signal.* Macmillan, 1993.

[87] L. Deng, A. Acero, M. Plumpe, and X. D. Huang. Large vocabulary speech recognition under adverse acoustic environments. *Proc. of ICSLP*, 2000.

[88] L. Deng, J. Droppo, and A. Acero. A bayesian approach to speech feature enhancement using the dynamic cepstral prior. *Proc. of ICASSP*, 2002.

[89] L. Deng, J. Wu, J. Droppo, and A. Acero. Analysis and comparison of two speech feature extraction/compensation algorithms. *IEEE Signal Proceeding Letters*, 12(6):477–480, 2005.

[90] S. Dharanipragada and B.D. Rao. MVDR based feature extraction for robust speech recognition. *Proc. of ICASSP*, 1:309–312, 2001.

[91] S. Dharanipragada, U.H. Yapanel, and B. Rao. Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method. *Trans. on SAP*, 15(1):224–234, January 2007.

[92] B. Edler and G. Schuller. Audio coding using a psychoacoustic pre- and postfilter. *Proc. of ICASSP*, 2:881–884, 2000.

[93] J.G. Fiscus, J. Ajot, M. Michel, and J.S. Garofolo. The rich transcription 2006 spring meeting recognition evaluation. *Proc. of Machine Learning*

for Multimodal Interaction, S. Renals, S. Bengio, and J.G. Fiscus (Eds.), LNCS vol. 4299, Springer, pages 309–322, 2006.

[94] D. Focken and R. Stiefelhagen. Towards vision-based 3-d people tracking in a smart room. *IEEE Int. Conf. Multimodal Interfaces*, 2002.

[95] Scripts for web data collection provided by University of Washington. http://ssli.ee.washington.edu/projects/ears/WebData/web_data_collection.html.

[96] C. Fügen. *A System for Simultaneous Translation of Lectures and Speeches*. PhD thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, 2008.

[97] M. Fujimoto and S. Nakamura. Particle filter based non-stationary noise tracking for robust speech feature enhancement. *Proc. of ICASSP*, 2005.

[98] S. Furui. Speaker–independent isolated word recognition using dynamic features of speech spectrum. *Trans. on ASSP*, 34:52–59, 1986.

[99] S. Furui, K. Maekawa, and H. Isahara. Toward the realization of spontaneous speech recognition – introduction of a japanese priority program and preliminary results –. *Proc. of ICSLP*, 2000.

[100] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998.

[101] M. J. F. Gales. Adaptive training schemes for robust asr. *Proc. of ASRU*, 2001.

[102] M.J.F. Gales. *M.J.F. Gales (1996). Model-Based Techniques for Noise Robust Speech Recognition.* PhD thesis, University of Cambridge, 1996.

[103] M.J.F. Gales. Semi-tied covariance matrices for hidden Markov models. *Trans. on SAP*, 7:272–281, 1999.

[104] D. Gespert and P. Duhamel. Robust blind identification and equalization based on multi-step predictors. *Proc. of ICASSP*, 26(5):3621–3624, 1997.

[105] D. Gildea and T. Hofmann. Topic-based language models using em. *Proc. of Eurospeech*, pages 2167–2170, 1999.

[106] J. Glass, L. Hazen, T.J. Hetherington, and C. Wang. Analysis and processing of lecture audio data: Preliminary investigations. *Proc. of HLT-NAACL, Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, 2004.

[107] J.R. Glass, T.J. Hazen, D.S. Cyphers, K. Schutte, and A. Park. The MIT spoken lecture processing project. *Proc. of HLT*, 2005.

[108] D. Graff. An overview of broadcast news corpora. *Speech Communication*, 37:15–26, 2002.

[109] R. Haeb-Umbach. Investigations on inter-speaker variability in the feature space. *Proc. of ICASSP*, 1:397 – 400, 1999.

[110] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. *Proc. of ICASSP*, 1992.

[111] R. Haeb-Umbach and J Schmalenstroeer. A comparison of particle filtering variants for speech feature enhancement. *Proc. of Interspeech*, 2005.

[112] T. Hain, J. Dines, G. Garau, M. Karafiat, M. Moore, V. Wan, R. Ordelman, and S. Renals. Transcription of conference room meetings: an investigation. *Proc. of Interspeech*, 2005.

[113] A. Härmä, M. Karjalainen, L. Savioja, U. K. Välimäki, V. Laine, and J. Huopaniemi. Frequency-warped signal processing for audio applications. *J. Audio Eng. Soc.*, 48(11):1011–1031, Nov. 2000.

[114] A. Härmä and U.K. Laine. A comparison of warped and conventional linear predictive coding. *Trans. on SAP*, 9(5):579–588, 2001.

[115] S. Haykin. *Adaptive filter theory—3th ed.* Prentice Hall, 1991.

[116] R. M. Hegde, Y. Jin, and B. D.; Rao. Spectral estimation of voiced speech using a family of MVDR estimates. *Proc. of ICASSP*, 2007.

[117] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Jour. of ASA*, 87(4):1738–1752, Apr. 1990.

[118] R. Hincks. *Computer Support for Learneres of Spoken English.* Doctor thesis, KTH School of Computer Science and Communication, Stockholm, Sweden, 2005.

[119] H.G. Hirsch and H. Finster. A new approach for the adaptation of hmms to reverberation and background noise. *Speech Communication*, 50:244–263, 2008.

[120] X. Huang, A. Acero, and H.W. Hon. *Spoken language processing.* Prentice Hall, 2001.

[121] J.H. Jensen, M.G. Christensen, M.N. Murthi, and S.H. Jensen. Evaluation of MFCC estimation techniques for music similarity. *Proc. of EUSIPCO*, 2006.

[122] C. Johns-Lewis. *Prosodic differentiation of discourse models.* Croom Helm, 1986.

[123] M. Karjalainen. Auditory interpretation and application of warped linear prediction. *Proc. of Consistent & Reliable Acoustic Cues for Sound Analysis*, Sep. 2001.

[124] K. Kato, H. Nanjo, and T. Kawahara. Automatic transcription of lecture speech using topic-independent language modeling. *Proc. of ICSLP*, 2000.

[125] T. Kawahara, Y. Nemoto, and Y. Akita. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. *Proc. of ICASSP*, 2008.

[126] R. Khun and R. De Mori. A cache-based natural language model for speech recognition. *Trans. on Pattern Analysis and Machine Intelligenc*, 12(6):570–583, 1990.

[127] N. S. Kim. Imm-based estimation for slowly evolving environments. *IEEE Signal Processing Letters*, 5(6):146–149, Jun. 1998.

[128] K. Kinoshita, T. Nakatani, and Miyoshi M. Efficient dereverberation framework for automatic speech recognition. *Proc. of Interspeech*, pages 3145–3148, 2005.

[129] K. Kinoshita, T. Nakatani, and Miyoshi M. Spectral subtraction steered by multi-step forward linear prediction for single channel speech dereverberation. *Proc. of ICASSP*, pages 817–820, 2006.

[130] W. Koenig, H.K. Dunn, and L.Y. Lacy. The sound spectrograph. *Jour. of ASA*, 18:19–49, 1946.

[131] H. Kuttruff. Sound in enclosures. *Encyclopedia of Acoustics*, 1997.

[132] H. Kuttruff. *Room Acoustics*. Elsevier Applied Science, 2000.

[133] K. Lebart, J.M. Boucher, and P.N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, 87(3):359–366, May/June 2001.

[134] Lecturefinder. http://lecturefinder.com.

[135] Li Lee and Richard C. Rose. Speaker normalization using efficient frequency warping procedures. *Proc. of ICASSP*, volume I, pages 353–356, 1996.

[136] E. Leeuwis, M. Federico, and M. Cettolo. Language modeling and transcription of the TED corpus lectures. *ICASSP*, 2003.

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

[138] Linguistic Data Consortium (LDC). English broadcast news speech (Hub-4). www.ldc.upenn.edu/ Catalog/LDC97S44.html.

[139] K. Linhard and H. Klemm. Noise reduction with spectral subtraction and median filtering for suppression of musical tones in robust speech recognition using unknown communication channels. *ESCA-NATO Tutorial and Research Workshop*, pages 159–162, April 1997.

[140] J. Makhoul. Linear prediction: A tutorial review. *Proc. of the IEEE, Vol. 63, No. 4*, April 1975.

[141] H. Matsumoto and M. Moroto. Evaluation of mel-LPC cepstrum in a large vocabulary continuous speech recognition. *Proc. of ICASSP*, 1:117–120, 2001.

[142] M. Matsumoto, Y. Nakatoh, and Y. Furuhata. An efficient mel-LPC analysis method for speech recognition. *Proc. of ICSLP*, pages 1051–1054, 1998.

[143] J. McDonough, W. Byrne, and X. Luo. Speaker normalization with all–pass transforms. *Proc. of ICSLP*, 1998.

[144] F. Metze, C. Fügen, Y. Pan, T Schultz, and H. Yu. The ISL rt-04s meeting transcription system. *Proc. of ICASSP: ICASSP-2004 Meeting RecognitionWorkshop*, 2004.

[145] T. Misu, T. Kawahara. A bootstrapping approach for developing language model of new spoken dialogue systems by selecting web texts. *Proc. of Interspeech*, 2006.

[146] A. Moreno and J. Pfretzschner. Human head directivity in speech emission: A new approach. *Acoustics Letters*, Vol. 1:pp 78–84, 1978.

[147] P.J. Moreno, B. Raj, and R.M. Stern. Multivariate-gaussian-based cepstral normalization for robust speech recognition. *Proc. of ICASSP*, 1995.

[148] P.J. Moreno, B. Raj, and R.M. Stern. A vector taylor series approach for environment-independent speech recognition. *Proc. of ICASSP*, 1996.

[149] R. Muralishankar and D. O'Shaughnessy. A comparative analysis of noise robust speech features extracted from all-pass based warping with mfcc in a noisy phoneme recognition. *The Third International Conference on Digital Telecommunications*, 2008.

[150] R. Muralishankar, A. Sangwan, and D. O'Shaughnessy. Warped discrete cosine transform cepstrum: A new feature for speech processing. *Proc. of EUSIPCO*, 2005.

[151] M.N. Murthi and B.D. Rao. Minimum variance distortionless response (MVDR) modeling of voiced speech. *Proc. of ICASSP*, April 1997.

[152] M.N. Murthi and B.D. Rao. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *Trans. on SAP*, 8(3):221–239, May 2000.

[153] B.R. Musicus. Fast MLM power spectrum estimation from uniformly spaced correlations. *Trans. on ASSP*, 33:1333–1335, 1985.

[154] Y. Nakatoh, M. Nishizaki, S. Yoshizawa, and M. Yamada. An adaptive mel-LP analysis for speech recognition. *Proc. of ICSLP*, 2004.

[155] H. Nanjo and T. Kawahara. Unsupervised language model adaptation for lecture speech recognition. *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[156] H. Nanjo and T. Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *Trans. on SAP*, 12(4):391–400, 2004.

[157] T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama. Investigations into early and late reflections on distant–talking speech recognition toward suitable reverberation criteria. *Proc. of Interspeech*, 2007.

[158] NIST, Rich Transcription 2005 Spring Meeting Recognition Evaluation. http://www.nist.gov/speech/tests/rt/rt2005/spring.

[159] NIST, Rich Transcription 2006 Spring Meeting Recognition Evaluation. http://www.nist.gov/speech/tests/rt/rt2006/spring.

[160] NIST, Rich Transcription 2007 Meeting Recognition Evaluation. http://www.nist.gov/speech/tests/rt/rt2007.

[161] N. Nocerino, F.K. Soong, L.R. Rabiner, and D.H. Klatt. Comparative study of several distortion measures for speech recognition. *Proc. of ICASSP*, 1985.

[162] A.M. Noll. Short–time spectrum and 'cepstrum' technique for vocal–pitch detection. *Jour. of ASA*, 36:296–302, 1964.

[163] M. Novak and R. Mammone. Improvement of non-negative matrix factorization based language model using exponential models. *Proc. of ASRU*, 2001.

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

[165] OCW – Open Course Ware. http://ocw.mit.edu/index.html.

[166] J.P. Olive. *Acoustics of American English speech : A dynamic approach.* Springer, 1993.

[167] A.V. Oppenheim, D.H. Johnson, and K. Steiglitz. Computation of spectra with unequal resolution using the fast fourier transform. *IEEE Proc. Letters, Vol. 59, No. 2, pp. 229-301*, February 1971.

[168] A.V. Oppenheim and R.W. Schafer. *Discrete-time signal processing.* Prentice-Hall Inc., 1989.

[169] D. O'Shaughnessy. *Speech communications: Human and machine.* IEEE Press, 2000.

[170] T. Ottmann, S. Trahasch, and T. Lauer. Systems support for virtualizing traditional courses in science and engineering. *Proc. of IFIP Quality Education @ a Distance (QE@D)*, 2003.

[171] OYC – Open Yale Courses. http://oyc.yale.edu.

[172] Y. Pan and A. Waibel. The effects of room acoustics on MFCC speech parameter. *Proc. of ICSLP*, 2000.

[173] A. Papoulis and S.U. Pillar. *Probability, Random Variables, and Stochastic Processes.* 4nd ed., McGraw-Hill, 2002.

[174] A. Park, T.J. Hazen, and J.R. Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. *Proc. of ICASSP*, 2005.

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

[176] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig. fMPE: Discriminatively trained features for speech recognition. *Proc. of ICASSP*, Philadelphia, Pennsylvania, USA, 2005.

[177] D. Povey and P.C. Woodland. Improved discriminative training techniques for large vocabulary continuous speech recognition. *Proc. of ICASSP*, Salt Lake City, UT, USA, May 2001.

[178] L.R. Rabiner. On the use of autocorrelation analysis for pitch detection. *Trans. on ASSP*, 1977.

[179] B. Raj, R. Singh, and R. Stern. On tracking noise with linear dynamical system models. *Proc. of ICASSP*, 2004.

[180] B. Raj and R. M. Stern. Missing feature methods for robust automatic speech recognition. *IEEE Signal Processing Magazine: Special Issue on Speech Technology and Systems in Human-Machine Communication*, 2005.

[181] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, Boston, MA, 2004.

[182] J.M. Sachar and H.F. Silverman. A baseline algorithm for estimating talker orientation using acoustical data from a large-aperture microphone array. *Proc. of ICASSP*, 2004.

[183] S. Saito and K. Nakata. *Fundamentals of speech signal processing*. Academic Press, 1985.

[184] A. Sehr and W. Kellermann. Towards robust distant–talking automatic speech recognition in reverberant environments. E. Hänsler and G. Schmidt, editors, *Topics in Speech and Audio Processing in Adverse Environments*. Springer, 2008.

[185] Y. Shimizu, S. Kajita, K. Takeda, and F. Itakura. Speech recognition based on space diversity using distributed multi-microphones. *Proc. of ICASSP*, 2000.

[186] K. Shitaoka, H. Nanjo, and T. Kawahara. Automatic transformation of lecture transcription into document style using statistical framework. *Proc. of ICSLP*, 2004.

[187] R. Singh and B. Raj. Tracking noise via dynamical systems with a continuum of states. *Proc. of ICASSP*, 2003.

[188] J. O. Smith III and J. S. Abel. Bark and ERB bilinear transforms. *Trans. on SAP*, 7(6):697–708, 1999.

[189] H. Soltau, F. Metze, C. Fügen, and A. Waibel. A one pass-decoder based on polymorphic linguistic context assignment. *Proc. of ASRU*, 2001.

[190] V. Stanford and J. Garofolo. Beyond close-talk – issues in distant speech acquistion, conditioning classification, and recognition. *Proc. of ICASSP: Meeting Recognition Workshop*, 2004.

[191] S.S. Stevens, J. Volkman, and E.B. Newman. The mel scale equates the magnitude of perceived differences in pitch at different frequencies. *Jour. of ASA*, 8(3):185–190, 1937.

[192] A. Stolcke. SRILM - an extensible language modeling toolkit. *Proc. of ICSLP*, 2002.

[193] STR-DUST – Speech Translation for Domain Unlimited Spontaneous Communication Tasks. http://isl.ira.uka.de/index.php?id=17.

[194] H.W. Strube. Linear prediction on a warped frequency scale. *Jour. of ASA*, 68(8):1071–1076, 1980.

[195] I. Tashev and D. Allred. Revereberation reduction for improved speech recognition. *Proc. of HSCMA*, 2005.

[196] TC-STAR – Technology and Corpora for Speech to Speech Translation. http://www.tc-star.org.

[197] Presseinformationen Universität Karlsruhe (TH). InterACT zeigte erstmalig simultane Übersetzung eines freien Vortrags Kommunikation über Grenzen hinweg. *http://www.presse.uni-karlsruhe.de/4187.php*, 2005.

[198] The Freesound Project. garbage.coll.serv.ds70p.mp3. http://freesound.iua.upf.edu/samplesViewSingle.php?id=6986.

[199] The Mathworks, Matlab. http://www.mathworks.com.

[200] J. Tierney. A study of LPC analysis of speech in additive noise. *Trans. on ASSP*, 28(4), 1980.

[201] K. Tokuda, T. Kobayashi, and S. Imai. Adaptive cepstral analysis of speech. *Trans. on SAP*, 3(6):481–489, 1995.

[202] Videolectures. http://videolectures.net.

[203] M. Westphal and A. Waibel. Model-combination-based acoustic mapping. *Proc. of ICASSP*, 2001.

[204] D. Willed, T. Niesler, E. McDermott, Y Minami, and S. Katagiri. Pervasive unsupervised adaptation for lecture speech transcription. *Proc. of ICASSP*, 2003.

[205] World-Lecture-Project. http://www.world-lecture-project.org.

[206] J. Wu. *Disciminative Speaker Adaptation and Environment Robustness in Automatic Speech Recognition*. PhD thesis, Univ. of Hong Kong, 2004.

[207] J. Wu and Q. Huo. An environment compensation minimum classification error training approach and its evaluation on aurora2 database. *Proc. of ICSLP*, pages 453–456, 2002.

[208] M. Wu and D. Wang. A two-stage algorithm for one-microphone reverberant speech enhancement. *Trans. on ASLP*, 14(3):774–784, 2006.

[209] K. Yao and S. Nakamura. Sequential noise compensation by sequential monte carlo methods. *Adv. Neural Inform. Process. Syst.*, 14, Sep. 2002.

[210] B. Yegnanarayana and T. K. Raja. Performance of linear prediction analysis on speech with additive noise. *Proc. of ICASSP*, 1977.

[211] J. Yuan, M. Liberman, and C. Cieri. Towards an integrated understanding of speaking rate in conversation. *Proc. of ICSLP*, 2006.

[212] J.J. Zhang, H.Y. Chan, and P. Fung. Improving lecture speech summarization using rhetorical information. *Proc. of ASRU*, 2007.

Index

R

S

T

U

V

W

Automatic transcription of lectures and oral presentations is becoming an important task. Possible applications can be found in the fields of automatic translation, automatic summarization, information retrieval, digital libraries, education and communication research. Ideally those systems would operate on distant recordings, freeing the presenter from wearing body-mounted microphones. This task, however, is surpassingly difficult, given that the speech signal is severely degraded due to the larger distance between the mouth of the speaker and the microphone by both, background noise and reverberation. The main goal of this book is to investigate, invent and present methods to improve automatic transcription of lectures in the previous mentioned environments.

ISBN: 978-3-86644-394-5

www.uvka.de

ISBN 978-3-86644-394-5

9 783866 443945 >