Signals and Communication Technology

For further volumes:
http://www.springer.com/series/4748
Tobias Herbig, Franz Gerl, and Wolfgang Minker

Self-Learning Speaker Identification
A System for Enhanced Speech Recognition

Springer
Self-Learning Speaker Identification for Enhanced Speech Recognition

**Summary.** A self-learning speech controlled system has been developed for unsupervised speaker identification and speech recognition. The benefits of a speech controlled device which identifies its main users by their voice characteristics are obvious: The human-computer interface may be personalized. New ways for interacting with a speech controlled system may be developed to simplify the handling of the device. Furthermore, speech recognition accuracy may be significantly improved if knowledge about the current user can be employed. The speech modeling of a speech recognizer can be improved for particular speakers by adapting the statistical models on speaker specific data. The adaptation scheme presented captures speaker characteristics after a very few utterances and transits smoothly to a highly specific speaker modeling. Speech recognition accuracy may be continuously improved. Therefore it is quite natural to employ speaker adaptation for a fixed number of users. Optimal performance may be achieved when each user attends a supervised enrollment and identifies himself whenever the speaker changes or the system is reset. A more convenient human-computer communication may be achieved if users can be identified by their voices. Whenever a new speaker profile is initialized a fast yet robust information retrieval is required to identify the speaker on successive utterances. Such a scenario presents a unique challenge for speaker identification. It has to perform well on short utterances, e.g. commands, even in adverse environments. Since the speech recognizer has a very detailed knowledge about speech, it seems to be reasonable to employ its speech modeling to improve speaker identification. A unified approach has been developed for simultaneous speech recognition and speaker identification. This combination allows the system to keep long-term adaptation profiles in parallel for a limited group of users. New users shall not be forced to attend an inconvenient and time-consuming enrollment. Instead new users should be detected in an unsupervised manner while operating the device. New speaker profiles have to be initialized based on the first occurrences of a speaker. Experiments on the evolution of such a system were carried out on a subset of the SPEECON database. The results show that in the long run the system produces adaptation profiles which give continuous improvements in speech recognition and speaker identification rate. A variety of applications may benefit from a system that adapts individually to several users.
Contents

1 Introduction ............................................. 1
   1.1 Motivation ......................................... 2
   1.2 Overview ......................................... 4

2 Fundamentals ............................................ 5
   2.1 Speech Production ..................................... 5
   2.2 Front-End ............................................ 8
   2.3 Speaker Change Detection .............................. 16
      2.3.1 Motivation ..................................... 16
      2.3.2 Bayesian Information Criterion .................... 18
   2.4 Speaker Identification ................................. 21
      2.4.1 Motivation and Overview ......................... 22
      2.4.2 Gaussian Mixture Models ......................... 25
      2.4.3 Detection of Unknown Speakers ................... 29
   2.5 Speech Recognition .................................... 30
      2.5.1 Motivation ..................................... 30
      2.5.2 Hidden Markov Models ......................... 31
      2.5.3 Implementation of an Automated Speech
            Recognizer ..................................... 35
   2.6 Speaker Adaptation .................................... 37
      2.6.1 Motivation ..................................... 38
      2.6.2 Applications for Speaker Adaptation in Speaker
            Identification and Speech Recognition ............... 39
      2.6.3 Maximum A Posteriori ......................... 41
      2.6.4 Maximum Likelihood Linear Regression .......... 44
      2.6.5 Eigenvoices .................................... 45
      2.6.6 Extended Maximum A Posteriori ................. 51
   2.7 Feature Vector Normalization .......................... 53
      2.7.1 Motivation ..................................... 53
## Contents

2.7.2 Stereo-based Piecewise Linear Compensation for Environments .................................. 54  
2.7.3 Eigen-Environment .................................................. 56  
2.8 Summary ................................................................. 57

3 Combining Self-Learning Speaker Identification and Speech Recognition ........................................ 59  
3.1 Audio Signal Segmentation .................................................. 59  
3.2 Multi-Stage Speaker Identification and Speech Recognition .................................................. 63  
3.3 Phoneme Based Speaker Identification .................................................. 65  
3.4 First Conclusion ............................................................. 69

4 Combined Speaker Adaptation .................................................. 71  
4.1 Motivation ................................................................. 71  
4.2 Algorithm ................................................................. 72  
4.3 Evaluation ................................................................. 76  
4.3.1 Database ............................................................... 76  
4.3.2 Evaluation Setup ....................................................... 77  
4.3.3 Results ................................................................. 79  
4.4 Summary ................................................................. 82

5 Unsupervised Speech Controlled System with Long-Term Adaptation ........................................ 83  
5.1 Motivation ................................................................. 83  
5.2 Joint Speaker Identification and Speech Recognition .................................................. 85  
5.2.1 Speaker Specific Speech Recognition .................................................. 86  
5.2.2 Speaker Identification .................................................. 91  
5.2.3 System Architecture .................................................... 93  
5.3 Reference Implementation ................................................ 95  
5.3.1 Speaker Identification .................................................. 95  
5.3.2 System Architecture .................................................... 96  
5.4 Evaluation ................................................................. 97  
5.4.1 Evaluation of Joint Speaker Identification and Speech Recognition .................................................. 98  
5.4.2 Evaluation of the Reference Implementation .................................................. 105  
5.5 Summary and Discussion .................................................. 111

6 Evolution of an Adaptive Unsupervised Speech Controlled System ........................................ 115  
6.1 Motivation ................................................................. 115  
6.2 Posterior Probability Depending on the Training Level .................................................. 116  
6.2.1 Statistical Modeling of the Likelihood Evolution .................................................. 117  
6.2.2 Posterior Probability Computation at Run-Time .................................................. 122  
6.3 Closed-Set Speaker Tracking .................................................. 125  
6.4 Open-Set Speaker Tracking .................................................. 128
6.5 System Architecture ............................................... 132
6.6 Evaluation ............................................................ 134
    6.6.1 Closed-Set Speaker Identification ....................... 134
    6.6.2 Open-Set Speaker Identification ......................... 139
6.7 Summary ............................................................ 143

7 Summary and Conclusion ............................................. 145

8 Outlook ............................................................... 149

A Appendix ...................................................................... 153
    A.1 Expectation Maximization Algorithm ...................... 153
    A.2 Bayesian Adaptation ............................................. 155
    A.3 Evaluation Measures ............................................ 158

References ................................................................. 161

Index ............................................................................... 171
Nomenclature

ADC  Analog to Digital Converter
ASR  Automated Speech Recognition
BIC  Bayesian Information Criterion
CDHMM Continuous Density HMM
CMS  Cepstral Mean Subtraction
cos  cosine
DCT  Discrete Cosine Transform
DFT  Discrete Fourier Transform
EM   Expectation Maximization
EMAP Extended Maximum A Posterior
EV   Eigenvoice
FFT  Fast Fourier Transform
GMM  Gaussian Mixture Model
HMM  Hidden Markov Model
Hz   Hertz
ID   Identity
iid  independent and identically distributed
kHz  kilo Hertz
LDA  Linear Discriminant Analysis
LLR  Log Likelihood Ratio
log  logarithm
LR   Lip Radiation
MAP  Maximum A Posteriori
MFCC Mel Frequency Cepstral Coefficients
min  minute
ML   Maximum Likelihood
MLLR Maximum Likelihood Linear Regression
MLP  Multilayer Perceptron
MMSE Minimum Mean Squared Error
NN   Neural Networks
PCA  Principal Component Analysis
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBF</td>
<td>Radial Basis Functions</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operator Characteristics</td>
</tr>
<tr>
<td>SCHMM</td>
<td>Semi-Continuous HMM</td>
</tr>
<tr>
<td>sec</td>
<td>second</td>
</tr>
<tr>
<td>SPEECON</td>
<td>Speech-Driven Interfaces for Consumer Devices</td>
</tr>
<tr>
<td>SPLICE</td>
<td>Stereo-based Piecewise Linear Compensation for Environments</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform</td>
</tr>
<tr>
<td>UBM</td>
<td>Universal Background Model</td>
</tr>
<tr>
<td>VAD</td>
<td>Voice Activity Detection</td>
</tr>
<tr>
<td>VQ</td>
<td>Vector Quantization</td>
</tr>
<tr>
<td>VT</td>
<td>Vocal Tract</td>
</tr>
<tr>
<td>VTLN</td>
<td>Vocal Tract Length Normalization</td>
</tr>
<tr>
<td>WA</td>
<td>Word Accuracy</td>
</tr>
<tr>
<td>WER</td>
<td>Word Error Rate</td>
</tr>
</tbody>
</table>
Automatic speech recognition has attracted various research activities since the 1950s. To achieve a high degree of user-friendliness, a natural and easy to use human-computer interface is targeted for technical applications. Since speech is the most important means of interpersonal communication, machines and computers can be operated more conveniently with the help of automated speech recognition and understanding [Furui, 2009].

The history of speech recognition and speaker identification is characterized by steady progress. Whereas in the 1950s the first realizations were mainly built on heuristic approaches, more sophisticated statistical techniques have been established and continuously developed since the 1980s [Furui, 2009]. Today a high level of recognition accuracy has been achieved which allows applications of increasing complexity to be controlled by speech.

Speech recognition has attracted attention for a variety of applications such as office systems, manufacturing, telecommunication, medical reports or infotainment systems in automobiles [Rabiner and Juang, 1993]. Speech recognition can increase labor efficiency in call-centers or for dictation tasks of special occupational groups with extensive documentation duty. For in-car applications both the usability and security can be increased for a wide variety of users. The driver can be supported to safely participate in road traffic and to operate technical devices, e.g. navigation systems or hands-free sets.

Despite significant advances during the last few decades, there still exist some deficiencies that limit the wide-spread application of speech recognition in technical applications [Furui, 2009]. For example, recognition accuracy can be negatively affected by changing environments, speaker variability and natural language input [Junqua, 2000].

The goal of this book is to contribute to a natural human-computer communication. An automatic personalization through a self-learning speech controlled system is targeted. An integrated implementation of speech recognition and speaker identification has been developed which adapts individually to several users. Significant progresses for speech recognition are exemplified for in-car applications.
1.1 Motivation

Infotainment systems with speech recognition, in general navigation, telephone or music control, typically are not personalized to a single user. The speech signal may be degraded by varying engine, wind and tire noises, or transient events such as passing cars or babble noise. For embedded systems computational efficiency and memory consumption are important design parameters. Nevertheless, a very large vocabulary, e.g. city or street names, needs to be accurately recognized.

Speaker independent speech recognizers are trained on a large set of speakers. Obviously, the trained speech pattern does not fit the voice of each speaker perfectly [Zavaliagkos et al., 1995]. To achieve high recognition rates in large vocabulary dictation systems, the statistical models can be perfectly trained to a particular user [Thelen, 1996]. However, the user is known in this case and the environment in general does not change.

In a system without speaker tracking all information acquired about a particular speaker is lost with each speaker turn. However, often one can expect that a device is only used by a small number of users, e.g. 5 recurring speakers. Therefore, it seems to be natural to employ speaker adaptation separately for several speakers and to integrate speaker identification. In addition to an enhanced speech recognition accuracy by incremental speaker adaptation, the usability can be improved by tracking personal preferences and habits.

A simple implementation would be to impose the user to identify himself whenever the system is initialized or the speaker changes. A more natural human-computer communication will be achieved by identifying the current user in an unsupervised way. No additional intervention of the user should be required.

Automatic personalization through a self-learning speech controlled system is targeted in this book. An integrated implementation comprising speech recognition, speaker identification, detection of new users and speaker adaptation has been developed. Speech recognition is enhanced by individually adapting a speech recognizer to 5-10 recurring users in an unsupervised manner. Several tasks have to be accomplished by such a system:

The various speaker adaptation schemes which have been developed, e.g. by Gauvain and Lee [1994]; Kuhn et al. [2000]; Stern and Lasry [1987], mainly differ in the number of parameters which have to be estimated. In general, a higher number of parameters allows more individual representation of the speaker’s characteristics leading to improved speech recognition results. However, such approaches are only successful when a sufficient amount of utterances proportional to the number of adaptation parameters can be reliably attributed to the correct speaker. Especially during the initialization of a new speaker profile, fast speaker adaptation that can converge within a few utterances is essential to provide robust statistical models for efficient speech recognition and reliable speaker identification. When prior knowledge about
1.1 Motivation

Speaker variability can be used only a few parameters are required to efficiently adapt the statistical models of a speech recognizer. However, the capability to track individual characteristics can be strongly limited. In the long run speaker characteristics have to be optimally captured. Thus, a balanced strategy of fast and individual adaptation has to be developed by combining the respective advantages of both approaches.

Speech recognition has to be extended so that several individually trained speaker profiles can be applied in parallel. Optimal recognition accuracy may be achieved for each user by an efficient profile selection during speech recognition. Recognition accuracy and computational complexity have to be addressed.

Speaker identification is essential to track recurring users across speaker turns to enable optimal long-term adaptation. Speakers have to be identified reliably despite limited adaptation data to guarantee long-term stability and to support the speech recognizer to build up well trained speaker profiles. Unknown speakers have to be detected quickly so that new profiles can be initialized in an unsupervised way without any intervention of the user. An enrollment to train new profiles should be avoided for the sake of a more natural operation of speech controlled devices.

Since applications for embedded systems are examined, the system shall be designed to efficiently retrieve and represent speech and speaker characteristics, especially for real-time computation. Multiple recognitions of speaker identity and spoken text have to be avoided. A unified statistical modeling of speech and speaker related characteristics will be presented as an extension of common speech recognizer technologies.

Beyond the scope of this book, the knowledge about the speaker is not only useful for speech recognition. The human-computer interface may be personalized in different ways:

Feedback of the speaker identity to the speech dialog engine allows special habits and knowledge of the user about the speech controlled device to be taken into consideration. For instance, two operator modes for beginners and more skilled users are possible. Beginners can be offered a comprehensive introduction or assistance concerning the handling of the device. Advanced users can be pointed to more advanced features.

Knowledge about the speaker identity enables the system to prefer or pre-select speaker specific parameter sets such as an address book for hands-free telephony or a list of frequently selected destinations for navigation. Confusions of the speech recognizer can be reduced and in doubt the user can be offered a list of reasonable alternatives leading to an increased usability.

Barge-in detection is another application where techniques for speaker identification and speech recognition can be of help as shown by Ittycheriah and Mammone [1999]; Ljolje and Goffin [2007]. Barge-in allows the user to interrupt speech prompts, e.g. originating from a Text-To-Speech (TTS) system, without pressing a push-to-talk button. A more natural and efficient control of automated human-computer interfaces may be achieved.
1.2 Overview

First, the fundamentals relevant for the scope of this book are discussed. Speech production is explained by a simple model. It provides a first overview of speech and speaker variability. The core components of a speech controlled system, namely signal processing and feature extraction, speaker change detection, speaker identification, speech recognition and speaker adaptation, are introduced and discussed in detail. Step by step more sophisticated strategies and statistical models are explained to handle speech and speaker characteristics.

Then a brief survey of more complex approaches and systems concerning the intersection of speaker change detection, speaker identification, speech recognition and speaker adaptation is given. Speech and speaker characteristics can be involved in several ways to identify the current user of a speech controlled system and to understand the content of the spoken phrase. The target system of this book is sketched and the main aspects are emphasized.

The first component of the target system is a speaker adaptation scheme which combines short-term adaptation as well as individual adjustments of the speaker profiles in the long run. This speaker adaptation method is able to capture speaker characteristics after a few utterances and transits smoothly to a highly specific speaker modeling. It is used subsequently to initialize and continuously adapt the speaker profiles of the target system and forms the identification basis of an integrated approach for speaker identification and speech recognition. Remarkable improvements for speech recognition accuracy may be achieved under favorable conditions.

Another important component of the target system is the unified realization of speaker identification and speech recognition which allows a compact statistical representation for both tasks. Speaker identification enables the system to continuously adapt the speaker profiles of its main users. Speech recognition accuracy is continuously improved. Experiments have been conducted to show the benefit for speech recognition and speaker identification.

The basic system can be extended by long-term speaker tracking. Speaker tracking across several utterances significantly increases speaker identification accuracy and ensures long-term stability of the system. The evolution of the adaptive system can be taken into consideration so that a strictly unsupervised system may be achieved. New users may be detected in an unsupervised way.

Finally, a summary and conclusion are given which repeat the main aspects and results. Several future applications and extensions are discussed in an outlook.
The fundamentals of speech production, automatic speech processing and the components of a complete system for self-learning speaker identification and speech recognition are introduced in this chapter.

In the first section discussion starts with speech production to gain insight into speech and the problem of speaker variability. A simple yet effective model is provided to simplify the understanding of the algorithms handling speaker variability and speech characteristics.

Then feature extraction is considered as the inverse process. A technique is introduced which extracts the relevant speech features from the recorded speech signal that can be used for further automated processing.

Based on these features the modules speaker change detection, speaker identification and speech recognition are described. They are essential for a speech controlled system operated in an unsupervised manner. Starting from a basic statistical model, each module introduces an additional extension for a better representation of speech signals motivated by speech production theory.

Finally, two competing approaches are presented to handle unseen situations such as speaker variability or acoustic environments. Speaker adaptation allows modifying the statistical models for speech and speaker characteristics whereas feature vector normalization or enhancement compensates mismatches on a feature level. A continuous improvement of the overall performance is targeted.

Based on the fundamentals, several realizations of complete systems known from literature are presented in the next chapter.

2.1 Speech Production

Fant’s model [Fant, 1960] defines speech production as a source-filter model\(^1\). The air stream originating from the lungs flows through the vocal cords and

\(^1\) The subsequent description of speech production follows O’Shaughnessy [2000] if not indicated otherwise.
generates the excitation [Campbell, 1997]. The opening of the glottis determines the type of excitation and whether voiced or unvoiced speech is produced. Unvoiced speech is caused by turbulences in the vocal tract whereas voiced speech is due to a quasi-periodic excitation [Schukat-Talamazzini, 1995]. The periodicity of the excitation is called fundamental frequency F0 or pitch. It depends on speaker and gender specific properties of the vocal cords, e.g. length, tension and mass [Campbell, 1997]. The fundamental frequencies of male speakers fall in the range between 80 Hz and 160 Hz and are on average about 132 Hz. Female speakers have an average fundamental frequency of 223 Hz. Children use even higher frequencies.

The Vocal Tract (VT) generally comprises all articulators above the vocal cords that are involved in the speech production process [Campbell, 1997]. It can be roughly approximated by a series of acoustic tubes producing characteristic resonances known as formant frequencies [Campbell, 1997; Rabiner and Juang, 1993]. As speakers differ anatomically in shape and length of their vocal tract, formant frequencies are speaker dependent [Campbell, 1997]. The length of the vocal tract is given by the distance between the glottis and the lips. Male speakers have an average length of 17 cm whereas the vocal tract of female speakers is 13 cm long on average.

The modeling of speech production is based on the three main components comprising excitation source, vocal tract and lip radiation [Schukat-Talamazzini, 1995]. The characteristics of glottis, vocal tract and lip radiation are modeled by filters with the impulse responses $h_G$, $h_{VT}$ and $h_{LR}$. In this context $u(\tau)$ denotes the excitation signal at time instant $\tau$. The speech signal $s(\tau)$ is given by the convolution

$$s(\tau) = u(\tau) * h_G(\tau) * h_{VT}(\tau) * h_{LR}(\tau) \quad (2.1)$$

$$u(\tau) * h(\tau) = \int_{\tilde{\tau} = -\infty}^{\tau} u(\tilde{\tau}) \cdot h(\tau - \tilde{\tau}) \, d\tilde{\tau}. \quad (2.2)$$

Fig. 2.1 depicts the associated block diagram of the source-filter model. A simplified model can be given by

$$s(\tau) = u(\tau) * h_{VT}(\tau) \quad (2.3)$$

if the glottis and the lip radiation are neglected [Wendemuth, 2004].

Additive noise and channel characteristics, e.g. the speech transmission from the speaker’s mouth to the microphone, can be integrated by $n(\tau)$ and a further room impulse response $h_{Ch}$. The speech signal $s_{Mic}$ recorded by the microphone is given by

$$s_{Mic}(\tau) = u(\tau) * h_{VT}(\tau) * h_{Ch}(\tau) + n(\tau). \quad (2.4)$$

Speaker variability can be viewed as a result of gender and speaker dependent excitation, anatomical differences in the vocal tract and acquired
speaking habits [Campbell, 1997]. Low-level acoustic information is related to the vocal apparatus whereas higher-level information is attributed to learned habits and style, e.g. prosody, word usage and conversational style [Reynolds et al., 2003].

Noisy environments can have an additional influence on the manner of articulation which is known as Lombard effect [Rabiner and Juang, 1993]. Since speakers try to improve communication intelligibility, significant changes in their voice patterns can occur degrading automated speech recognition and speaker identification [Goldenberg et al., 2006]. The effects are highly speaker-dependent and cause an increase in volume, fundamental frequency and vowel duration as well as a shift of the first two formants and energy distribution [Junqua, 1996].

Speaker change detection and speaker identification have to provide appropriate statistical models and techniques which optimally capture the individual speaker characteristics.

Fig. 2.1 Source-filter model for voiced and unvoiced speech production as found by Rabiner and Juang [1993]; Schukat-Talamazzini [1995]. The speech production comprises either a quasi-periodic or noise-like excitation and three filters which represent the characteristics of the glottis, vocal tract and the lip radiation.

The focus of speech recognition is to understand the content of the spoken phrase. Therefore further aspects of speech are important. Speech can be decomposed into phonemes\(^2\) which can be discriminated by their place of articulation as exemplified for vowels, fricatives, nasal consonants and plosives:

- Vowels are produced by a quasi-periodic excitation at the vocal cords and are voiced. They are characterized by line spectra located at multiples of the fundamental frequency and their intensity exceeds other phonemes.

\(^2\) Phonemes denote the smallest linguistic units of a language. Many languages can be described by 20 – 40 phonemes. The physical sound generated by the articulation of a phoneme is called phone [O’Shaughnessy, 2000].
Nasal consonants are characterized by a total constriction of the vocal tract and a glottal excitation [Rabiner and Juang, 1993]. The nasal cavity is excited and attenuates the sound signals so that nasals are less intensive than vowels.

Fricatives are produced by constrictions in the vocal tract that cause a turbulent noisy airflow. The constriction produces a loss of energy so that fricatives are less intensive. They are characterized by energy located at higher frequencies.

Plosives (stops) are produced by an explosive release of an occlusion of the vocal tract followed by a turbulent noisy air-stream.

The phoneme based speaker identification discussed in Sect. 3.3 exploits the knowledge about the speaker variability of phonemes to enhance the identification accuracy. For example, no information about the vocal tract is present if the place of articulation is given by the teeth or lips as in /f/ or /p/. In consequence, little speaker discrimination is expected.

In the next section an automated method for feature extraction is presented. The goal is to extract the relevant spectral properties of speech which can be used for speaker change detection, speaker identification and speech recognition.

2.2 Front-End

The front-end is the first step in the speech recognition or speaker identification processing chain. In this book it consists of a signal processing part, feature extraction and post-processing as displayed in Fig. 2.2.

The signal processing applies a sampling and preprocessing to the recorded microphone signals. The feature extraction transforms discrete time signals into a vector representation which is more feasible for pattern recognition algorithms. The following post-processing comprises feature vector normalization to compensate for channel characteristics. In the case of speech recognition a discriminative mapping is used in addition. The three parts of the front-end are described in more detail.

![Fig. 2.2](image-url) Block diagram of a front-end for speech recognition and speaker identification. The recorded microphone signal is preprocessed to reduce background noises. A vector representation is extracted from the speech signal. The feature vectors are normalized in the post processing to compensate for channel characteristics.
2.2 Front-End

Signal Processing

The signal processing receives an input signal from the microphone and delivers a digitized and enhanced speech signal to feature extraction. The principle block diagram is displayed in Fig. 2.3.

![Block diagram of the signal processing in the front-end.](image)

**Fig. 2.3** Block diagram of the signal processing in the front-end. The recorded speech signal is sampled at discrete time instances and noise reduction is performed. Speech pauses are excluded for subsequent speech processing.

The Analog to Digital Converter (ADC) samples the incoming time signals at discrete time instances

\[ s_{\text{Mic}}(l) = s_{\text{Mic}}(\frac{l}{f_s}), \quad l = 0, 1, 2, 3, \ldots \]

and performs a quantization. In this book a sampling rate of \( f_s = 11.025 \text{ kHz} \) and a 16 bit quantization are used. In this context \( l \) denotes the discrete time index.

Especially in automotive applications, speech signals are superimposed with background noises affecting the discrimination of different speakers or speech decoding. Noise reduction targets to minimize environmental influences which is essential for a reliable recognition accuracy.

Noise reduction can be achieved by a spectral decomposition and a spectral weighting as described by Vary et al. [1998]. The time signal is split into its spectral components by an analysis filter bank. A noise estimation algorithm calculates the noise spectrum as found by Cohen [2003]; Cohen and Berdugo [2002]. In combination with an estimate of the disturbed speech spectrum, the power of the undisturbed or clean speech signal is estimated by a spectral weighting. The Wiener filter, spectral subtraction and Ephraim-Malah are well-known weighting techniques which are described by Cappé [1994]; Ephraim and Malah [1984] in more detail. The phase of the speech signals remains unchanged since phase distortions seem to be less critical for human hearing [Vary et al., 1998]. After spectral weighting the temporal speech signals are synthesized by a synthesis filter bank. The enhanced speech signal is employed in the subsequent feature extraction. Fig. 2.4 displays the noise reduction setup as described above.

As noise reduction algorithms do not play an important role in this book, only the widely-used Wiener filter is applied. Interested readers are referred to Benesty et al. [2005]; Hängsler and Schmidt [2004, 2008] for further detailed information.
Fig. 2.4 Block diagram of the noise reduction. The signal is split into its spectral components by an analysis filter bank. Noise and speech power are estimated to calculate a spectral weighting. The power of the clean speech signal is estimated by a spectral weighting of the disturbed speech spectrum. The enhanced speech signal is synthesized by the synthesis filter bank.

Voice Activity Detection (VAD) or speech segmentation excludes speech pauses so that the subsequent processing is done on portions of the microphone signal that are assumed to comprise only speech utterances. The computational load and the probability of false classifications can be reduced for speaker identification and speech recognition. Conventional speech segmentation algorithms rely on speech properties such as the zero-crossing rate, rising or falling energy\(^3\) \(E\{(s_{i}^{\text{Mic}})^2\}\) [Kwon and Narayanan, 2005]. Furthermore, the harmonic structure of voiced speech segments may be used, especially for vowels. Adverse environments complicate the end-pointing because background noises mask regions of low speech activity. Finally, speech segmentation can shift the detected boundaries of the beginning and end of an utterance to guarantee that the entire phrase is enclosed for speech recognition. For further reading, the interested reader is referred to literature such as Espi et al. [2010]; Ramírez et al. [2007].

Feature Extraction

Feature extraction transforms the enhanced speech signals into a vector representation which reflects discriminatory properties of speech. The Mel Frequency Cepstral Coefficients (MFCC) are frequently used in speech recognition and speaker identification [O’Shaughnessy, 2000; Quatieri, 2002]. MFCCs are physiologically motivated by human hearing\(^4\).

The reader may wonder why the MFCC features are also used for speaker identification. Even though MFCCs are expected to cause a loss of speaker information, they seem to capture enough spectral information for speaker identification. The vocal tract structure may be considered as the

---

\(^3\) \(E\{\}\) denotes the expectation value.

\(^4\) For the comparison of human and machine in speaker identification and speech recognition it is referred to Hautamäki et al. [2010]; Liu et al. [1997]; O’Shaughnessy [2000] and Meyer et al. [2006].
dominant physiological feature to distinguish speakers. This feature is reflected by the speech spectrum [Reynolds and Rose, 1995]. Kinnunen [2003] provides a thorough investigation of several spectral features for speaker identification and concludes that the best candidates range at the same level. However, features directly computed by a filter bank such as the MFCCs are obviously better suited for noisy applications [Reynolds and Rose, 1995]. MFCCs enable a compact representation and are suitable for statistical modeling by Gaussian mixtures [Campbell, 1997] as explained later in Sect. 2.4. Reynolds [1995a,b] obtains excellent speaker identification rates even for large speaker populations. According to Quatieri [2002] the mel-cepstrum can be regarded as one of the most effective features for speech-related pattern recognition.

Hence, the description in this book is restricted to the extraction of the MFCC features. High-level features [Quatieri, 2002; Reynolds et al., 2003] are not considered since a command and control scenario is investigated. An exemplary MFCC block diagram is shown in Fig. 2.6.

First, the Short Time Fourier Transform (STFT) divides the sequence of speech samples in frames of predetermined length, applies a window function and splits each frame into its spectral components. Common window functions are the Hamming and Hann window [Vary et al., 1998]. In this book the window function $\tilde{h}$ is realized by the Hann window

$$\tilde{h}_l = \frac{1}{2} \left( 1 + \cos \left( \frac{2\pi l}{N_{Frame}} \right) \right), \quad l = -\frac{1}{2} N_{Frame}, \ldots, \frac{1}{2} N_{Frame}, \quad (2.6)$$

where the length of one frame is given by $N_{Frame}$. Subsequently, $t$ denotes the discrete frame index. The frame shift $N_{shift}$ describes the time which elapsed between two frames. The frame length is 20 ms and frame shift is given as half of the frame length. The windowing

$$\tilde{s}^w_{t,l} = s^\text{Mic}_t \cdot N_{shift} + l \cdot \tilde{h}_l, \quad t > 1, \quad (2.7)$$

combined with the Discrete Fourier Transform (DFT) or Fast Fourier Transform (FFT)$^5$ realizes a filter bank [Vary et al., 1998]. The filter characteristics are determined by the window’s transfer function shifted in the frequency domain [Hänsl and Schmidt, 2004; Vary et al., 1998]. The incoming microphone signals are split into $N_{FFT}$ narrow band signals $S(f_b, t)$ [Hänsl and Schmidt, 2004; Rabiner and Juang, 1993]:

$$S(f_b, t) = \mathcal{F}\mathcal{F}\mathcal{T} \{ \tilde{s}^w_{t,l} \} = \frac{1}{2} N_{Frame} \sum_{l=-\frac{1}{2} N_{Frame}}^{\frac{1}{2} N_{Frame}} \tilde{s}^w_{t,l} \cdot \exp(-i \frac{2\pi}{N_{FFT}} f_b l), \quad N_{FFT} \geq N_{Frame}. \quad (2.8)$$

$^5$ The FFT represents an efficient implementation of the DFT [Kammeyer and Kroschel, 1998]. Details on DFT and FFT are given by Kammeyer and Kroschel [1998]; Oppenheim and Schafer [1975], for example.