ROBUST SPEECH RECOGNITION

K.K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
(e-mail: K.Paliwal@me.gu.edu.au)

ABSTRACT

Performance of an automatic speech recognition system degrades drastically when there is a mismatch between training and testing conditions. The aim of robust speech recognition is to overcome the mismatch problem so as to result in a moderate and graceful degradation in recognition performance. In this paper, we provide a brief overview of an automatic speech recognition system, describe sources of speech variability that cause mismatch between training and testing, and discuss some of the current techniques to achieve robust speech recognition.

1. INTRODUCTION

Automatic speech recognition is a difficult problem and has a long history with initial papers appearing in 50's [1, 2]. However, thanks to significant progress made in recent years in this area [3, 4], the speech recognition technology which was confined earlier to the research laboratories is now being applied to the real-world applications and a number of commercial speech recognition products (from Dragon, IBM, Philips, etc.) have appeared in the market. The main factors that have made it possible are the advances made in digital signal processing (DSP) techniques and stochastic modeling algorithms. The signal processing techniques are important for extracting reliable acoustic features from the speech signal, and the stochastic modeling algorithms are useful for representing speech utterances in the form of efficient models, such as the hidden Markov models (HMMs), which simplify the task of speech recognition. The other factors responsible for the commercial success of the speech recognition technology include the availability of fast processors (in the form of DSP chips) and high density memories at relatively low cost.

With the current state-of-the-art in the speech recognition technology, it is relatively easy to accomplish fairly complex speech recognition tasks reasonably well in a controlled laboratory environments. For example, it is possible to achieve 1% error rate in a speaker-dependent isolated word recognition task with 20,000 word vocabulary [5] and less than 0.5% word error-rate in a speaker independent digit recognition task [6]. Even the continuous speech from any speaker and from a vocabulary of a few thousand words can be recognized with a word-error rate of between 5% and 10% [7]. This high level of speech recognition performance is achievable only when the training and the test data are matched. When there is a mismatch between the training and the test data, the speech recognition performance degrades drastically.

The mismatch between the training and test sets may occur due to changes in acoustic acoustic environments (background, channel mismatch, etc.), speakers, task domains, speaking styles, etc. [8]. Each of these sources of mismatch can cause severe distortion in recognition performance. For example, an isolated word recognizer that can recognize 10 English digits perfectly when spoken in laboratory environment, recognizes only 30% of the spoken digits when white noise is added to the signal with 10 dB signal-to-noise ratio (SNR) [9]. Similar degradations in recognition performance are observed due to channel mismatch. The recognition accuracy of the SPHINX speech recognition system on a speaker independent alphanumeric task dropped from 85% correct to 20% correct when the close-talking Sennheiser microphone used in training was replaced by the omnidirectional Crown desktop microphone [10]. Similarly, when a digit recognition system is trained tested for a particular speaker, its recognition accuracy can be easily 100%. But its recognition performance can go down to as low as 50% when it is tested on a new speaker.

In order to understand the effect of mismatch between training and test conditions, we show in Fig. 1 the performance of a speaker-dependent, isolated word recognition system on speech corrupted by additive white noise. The recognition system uses a a 9-word English e-set alphabet vocabulary where each word is represented by a single-mixture continuous Gaussian density HMM with five states. Figure 1 shows the recognition accuracy as a function of the signal-to-noise ratio (SNR) of the test speech under the following two types of conditions: 1) Mismatched conditions where the recognition system is trained on clean speech and tested on noisy speech, and 2) Matched conditions where the training and the test speech data have the same SNR. It can be seen from this figure that the additive noise causes a drastic degradation in recognition performance under the mismatched conditions; but with the matched conditions, the degradation is moderate and graceful. It may be noted here that if the SNR becomes too low (such as -10 dB), the system results in a very poor recognition performance even when it is operated under matched noise conditions. This is because the signal is completely swamped by noise and no useful information can be extracted from it neither during training nor in testing.

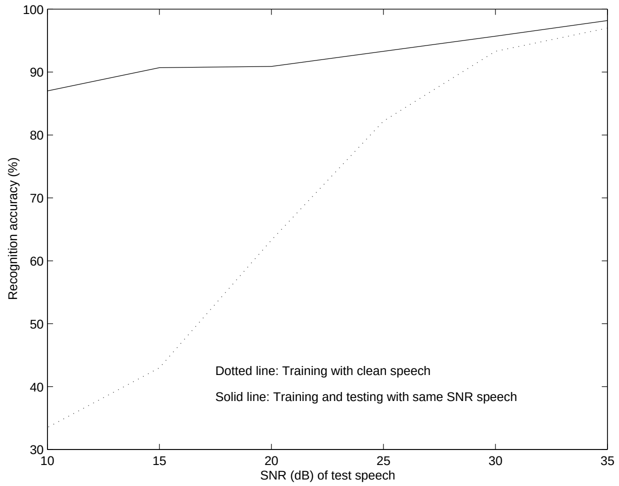

Figure 1: Effect of additive white noise on speech recognition performance under matched and mismatched conditions.

When a speech recognition system is deployed in practice in a real-life situation, there is bound to be a mismatch between training and testing. This mismatch causes severe deterioration in the recognition performance. The aim of a robust speech recognition system is to remove the effect of this mismatch and achieve a recognition performance that is as graceful as obtained under matched conditions.

In this paper, we provide only a glimpse of robust speech recognition and describe briefly some of the currently popular techniques used for this purpose. (For more details, see [11, 12, 13, 14, 15, 16, 17].) We will focus here mainly on techniques used to handle mismatches resulting from changes in acoustic environments (e.g., due to channel and noise distortions). Some of these techniques are equally applicable to mismatches resulting from speaker variability. This paper is organized as follows: Section 2 provides a brief overview of the automatic speech recognition process. Different sources of variability in the speech signal are discussed in Section 3. Robust speech recognition techniques are briefly described in Section 4. Section 5 summarizes the paper.

2. SPEECH RECOGNITION: AN OVERVIEW

Objective of an automatic speech recognition system is to take the speech waveform of an unknown (input) utterance, and classify it as one of a set of spoken words, phrases, or sentences. Typically, this is done in two steps (as shown in Fig. 2). In the first step, an acoustic front-end is used to perform feature analysis where the speech signal is analyzed at the rate of about 100 frames per second (i.e., every 10 ms) to extract a set of features. This produces a sequence of feature vectors that characterizes the speech utterance sequentially in time. Second step deals with pattern classification where the sequence of feature vectors is compared against the machine's knowledge of speech (in the form of acoustics, lexicon, syntax, semantics, etc.) to arrive at a transcription of the input utterance.

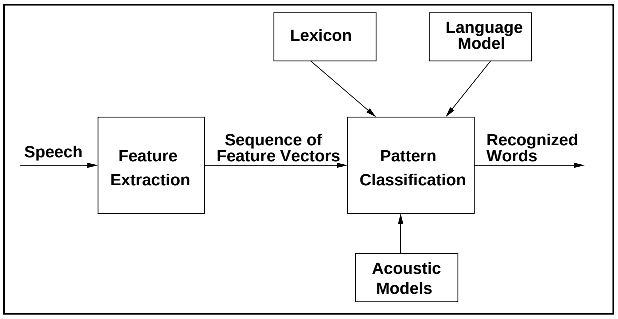

Figure 2: A block diagram of an automatic speech recognition system.

Currently, most of the speech recognition systems use a statistical framework to carry out the pattern classification task, and they generally recognize the input speech utterance as a sequence of words. Consider a sequence of feature vectors,

$$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\},$$

representing the T frames of the input speech utterance. The task of the speech recognition system is to find a word sequence,

$$W = \{w_1, w_2, \ldots, w_K\},$$

that maximizes the *a posteriori* probability of the observation sequence $\mathbf{Y}$; i.e., the recognized word sequence,

$$\hat{W} = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_{\hat{K}}\},$$

is given by the following equation:

$$\hat{W} = \operatorname*{argmax}_{W} \Pr(W|\mathbf{Y}). \tag{1}$$

In this equation, maximization of the a posteriori probability $\Pr(W|\mathbf{Y})$ is done over all possible word sequences $\{w_1, w_2, \ldots, w_K\}$ for all possible values of K. For a large vocabulary continuous speech recognition system, this is a computationally exorbitant task. Fast search algorithms are available in the literature to carry out this task [18, 19, 20].

Applying the Bayes rule and noting that $\Pr(\mathbf{Y})$ is independent of W, Eq. (1) can be written as

$$\hat{W} = \operatorname*{argmax}_{W} \Pr(\mathbf{Y}|W) \cdot \Pr(W). \tag{2}$$

This equation is known as the maximum a posteriori probability (MAP) decision rule in the statistical pattern recognition literature [21].

Equation (2) indicates that we need two probabilities $\Pr(\mathbf{Y}|W)$ and $\Pr(W)$ to carry out the recognition task. These are computed through the acoustic and the language models, respectively. These models are briefly described below.

- **Acoustic models**: The acoustic models are used to compute the probability $\Pr(\mathbf{Y}|W)$. In order to compute this probability, we need the probability of an observed sequence of feature vectors for each of the words in the vocabulary. This is done by representing each word by a hidden Markov model (HMM) [23] and estimating the HMM parameters from an independent (and preferably large) speech data during the training phase. In order to capture the sequential nature of speech, the left-to-right HMMs are used to model individual words. For a large vocabulary continuous speech recognition system, it is not possible to have one HMM for each word, so we seek smaller units (subword units) to characterize these probabilities. Examples of these subword units are phonemes, demisyllables and syllables. If there are M phonemes in the (English) language, we can have M HMMs, each estimated from the training data belonging to the particular phoneme. These are called context-independent models. For a large vocabulary speech recognition system, these models are not adequate and one requires context-dependent modeling to get good recognition performance. Current recognition systems use HMMs for all possible left and right contexts for each phoneme (triphone models). Once the acoustic models (in the form of HMMs) are available for individual subword units (e.g., triphones) from the training phase, the word models are constructed from the subword models according to the transcription of the words (in terms of subword units) contained in the lexicon.

- **Language model:** The language model is used to compute the probability $\Pr(W)$. Note that $\Pr(W)$ is independent of the observed feature vector sequence $\mathbf{Y}$. Like the acoustic models, the language model is estimated from an independent and large corpus of training data. Among different language models proposed in the literature, the N-gram model (where N is typically 2, 3 or 4) is perhaps the most popular and simple model to represent the syntactic, semantic and pragmatic sources of knowledge. In this model, the probability of the current word depends on N preceding words. Thus, this model is very effective in capturing local dependencies between words. In an N-gram model, the probability $\Pr(w_k|w_1, w_2, \ldots, w_{k-1})$ is approximated by $\Pr(w_k|w_{k-1}, w_{k-2}, \ldots, w_{k-N+1})$. As an example, we show below the procedure for calculating the probability $\Pr(W)$ using the trigram model ($N = 3$).

$$\begin{aligned} \Pr(W) &= \Pr(w_1, w_2, \ldots, w_K) \\ &= \prod_{k=1}^{K} \Pr(w_k|w_1, w_2, \ldots, w_{k-1}) \\ &= \prod_{k=1}^{K} \Pr(w_k|w_{k-1}, w_{k-2}). \end{aligned} \tag{3}$$

Thus, we compute the acoustic and language models in the training phase from the data available for training the speech recognizer. Let us denote the set of acoustic models (i.e.; subword HMMs) by $\Lambda_{\mathbf{X}}$ and the set of N-gram models by Υ_W. Then, the MAP decision rule (Eq. (2)) can be written in terms of these models as follows:

$$\hat{W} = \operatorname*{argmax}_{W} \Pr(\mathbf{Y}|W, \Lambda_{\mathbf{X}}) \cdot \Pr(W|\Upsilon_W). \tag{4}$$

During the test phase, the recognizer uses the acoustic and language models to compute the probabilities $\Pr(\mathbf{Y}|W, \Lambda_{\mathbf{X}})$ and $\Pr(W|\Upsilon_W)$, and carry out the recognition of the input utterance according to the MAP decision rule given by Eq. (4).

3. VARIABILITY IN THE SPEECH SIGNAL

Robust speech recognition deals with the problem caused by mismatch between the training and testing conditions. Most of the mismatch occurs due to the variability in the speech signal resulting from various sources; some of which are listed below.

- **Background noise:** When speech is recorded in a given acoustic environment, the resulting signal is a sum of speech produced by a speaker and the background (or ambient) noise. This additive noise has generally a colored spectrum, whose shape depends on the source that generates it. In an office environment, the background noise results from sources such as computers, printers, typewriters, air-conditioners, ringing of telephone, fans, persons talking in the background, etc. In a moving car environment, it can be due to engine, wind, tires, road, etc. Similarly, sources of background noise for other environments (such as telephone booth, industrial plant, plane cockpit, etc.) can be easily identified. Depending on the source, the background noise can be stationary (e.g. fans, air-conditioners) or non-stationary (e.g., moving car).

Though the spectral shape as well as the level of background noise cause a degradation in recognition performance, the later affects the recognition performance more. When the background noise has a relatively high level, it produces Lombard effect [24], which changes even the characteristics of the speech signal produced by the given speaker.

- **Room reverberation:** Room reverberation causes a convolutional distortion; i.e., the reverberated speech signal can be modeled as convolution of the intended speech signal with the impulse response characterizing the distortion. The amount of reverberation distortion is determined by the room acoustics and the position of speaker and microphone within the room. When a speaker is at a relatively large distance from the microphone, the reverberation distortion becomes serious and can affect the speech recognition performance significantly [25].

- **Microphone characteristics:** Microphone acts on the speech signal as a linear filter and causes convolution-type of distortion. Since different types of microphones have different frequency responses, their mismatch during training and test conditions causes severe degradation in recognition performance [10].

- **Transmission channel:** When a speech recognizer is accessed through a telephone or a mobile phone, the transmission channel one gets is totally unknown and unpredictable. This causes mismatch between training and testing; and the speech recognition performance suffers because of this. Transmission channel acts like a linear filter on the speech signal and causes convolution-type of distortion. Mobile telephony introduces another distortion resulting from speech coders which also affects the recognition performance adversely [26].

- **Intra-speaker variability:** When a person speaks the same word twice at different times of a day, the resulting speech utterances show different acoustic characteristics. This intra-speaker variability is mainly caused by changes in the health and emotional states of the speaker.

- **Inter-speaker variability:** Inter-speaker variability is one of the main sources of introducing the mismatch between training and testing conditions and is a major cause of affecting the performance of a speech recognizer adversely. Differences in the length and shape of the vocal tract, dialect, pronunciation and articulatory habits are some of the sources of this variability.

Most of the sources of speech variability discussed above produce the additive-type distortion (e.g., background noise) and/or the convolution-type of distortion (e.g., microphone mismatch) in the speech signal. A common model that describes these distortions and helps in understanding the robust speech recognition techniques (discussed in the next section) is as follows:

$$y(m,n) = x(m,n) \star h(m,n) + w(m,n) \qquad (5)$$

where the symbol $\star$ denotes the convolution operation and m is the frame index, n the time index within the m-th frame, $x(m,n)$ the clean speech signal, $y(m,n)$ the distorted signal, $h(m,n)$ the impulse response describing the convolution-type distortion and $w(m,n)$ the additive noise signal. In this equation, both of the distortions are non-stationary in nature. If we assume these distortions to be stationary, Eq. (5) becomes

$$y(m,n) = x(m,n) \star h(n) + w(n) \qquad (6)$$

If $x(m,n)$, $h(n)$ and $w(n)$ are uncorrelated, it can be easily shown that

$$P_{yy}(m,f) = P_{xx}(m,f)|H(f)|^2 + P_{ww}(f) \qquad (7)$$

where f is the frequency variable, $H(f)$ the Fourier transform of $h(n)$, and $P_{yy}(m,f)$, $P_{xx}(m,f)$ and $P_{ww}(f)$ are the power spectra of $y(m,n)$, $x(m,n)$ and $w(n)$, respectively. If there is only the convolution-type of distortion present in the signal, Eq. (7) can be written as

$$\log P_{yy}(m,f) = \log P_{xx}(m,f) + 2\log|H(f)| \qquad (8)$$

When the signal is corrupted by the additive noise distortion (i.e., there is no convolutional distortion), then Eq. (7) can be written as

$$P_{yy}(m,f) = P_{xx}(m,f) + P_{ww}(f) \qquad (9)$$

Equations (7), (8) and (9) form the basis of a number of the robust speech recognition techniques discussed in the next section.

4. ROBUST SPEECH RECOGNITION TECHNIQUES

As mentioned earlier, robust speech recognition tries to deal with the problem resulting from the mismatch between training and testing. A speech recognizer is called robust if it (approximately) maintains its good recognition performance even if there is a mismatch between the training and testing conditions.

Some researchers believe that one can solve the mismatch problem by increasing the size of the training data set and including all possible speech variation into it. Though it solves the problem to some extent, this belief is not completely true. The models computed from this large training data will be diffused and diluted; i.e., they will have large variance. As a result, their performance will be relatively poor for all the test conditions. By increasing the size of training data set and including all possible speech variations into it, one is only improving the generalization capability at the cost of recognition performance.

In order to really solve the robust speech recognition problem, one has to understand the basic characteristics of the speech signal and the effect of different sources of distortion and variability, and then capture this knowledge during the feature extraction and acoustic modeling stages. If the mismatch still remains, then use small amount of adaptation data prior to testing the recognition system for fine tuning. A number of techniques have been reported in the literature on these lines for robust speech recognition. Some of these are briefly described below. We assume here that the clean (undistorted) speech $\{x(n)\}$ is used for training the recognition system and the distorted speech $\{y(n)\}$ for testing.

4.1. Speech Enhancement Techniques

The aim of a speech enhancement system is to suppress noise from the noisy speech signal. For robust speech recognition, it is used as a preprocessor to a speech recognizer. Since it produces a clean speech signal, it is not necessary to make any changes in the recognition system to make it robust. A number of speech enhancement techniques have been reported in the literature [27]. These include spectral suppression [28, 29, 27], Wiener and Kalman filtering [30], comb filtering [27] and singular value decomposition (SVD) [31, 32].

These enhancement techniques were originally developed with the aim of improving the intelligibility of noisy speech in mind. However, they can be used for robust speech recognition as well. The technique that has been used most for this purpose is the spectral subtraction method [33, 34]. In this method, the power spectrum of clean speech $P_{xx}(m, f)$ is estimated by explicitly subtracting the noise power spectrum $P_{ww}(f)$ from the noisy speech power spectrum $P_{yy}(f)$ using Eq. (9). This method requires the information about noise power spectrum. This can be estimated from the non-speech frames. However, it is not always possible to detect the non-speech frames correctly. This affects the estimation of the noise power spectrum and may result in poor speech enhancement performance. When used in combination with other noise compensation method, the spectral subtraction method has been found to be very useful for robust speech recognition [34]. Recently, this method has been applied in wavelet domain with good results [35].

Another technique of speech enhancement based on singular value decomposition (SVD) has been recently used for robust speech recognition [36]. In this method, SVD is applied to an over-determined, over-extended data matrix formed from the noisy speech signal and a noise-free, low rank approximation is obtained by retaining a specific number of singular values. This technique is found to improve the recognition performance significantly for SNRs less than 15 dB.

4.2. Special Transducer Arrangement

A speech recognizer can be made robust to adverse acoustic environments if the transducers can be arranged in a favorable fashion. For example, if we use a unidirectional microphone and place it near the mouth, we can reduce the distortion due to background noise and reverberation. This will help in improving the SNR of the recorded speech. If we use two microphones, one to capture the noise signal and the other to pick up the noisy speech signal, we can apply the adaptive filtering algorithms such as the least mean squares (LMS) algorithm to achieve speech enhancement. This helps in canceling both stationary and nonstationary noises and improves the recognition performance in the presence of noise [37, 38].

For hands-free speech recognition applications (e.g., teleconferencing, speech recognition in a car, etc.), it is not possible to use a close-talking microphone. In these applications, an array of microphones is used to improve the SNR of speech as well as to cope up with the speaker mobility [39, 40, 41]. The SNR of speech is increased under both stationary and nonstationary acoustic environments (background noise and reverberation) by using an adaptive beam-forming procedure and the speaker ability is managed by using a source location procedure working jointly with the beamformer. Good recognition results are reported in the literature for mismatched acoustic environments using a microphone array [42].

4.3. Robust Feature Selection and Extraction Methods

Selection of proper acoustic features is perhaps the most important task in the design of a robust speech recognition system. It directly affects the performance of the recognition system. These features should be selected with the following criteria in mind:

1. They should contain maximum information necessary for speech recognition.

2. They should be insensitive to speaker characteristics, manner of speaking, background noise, channel distortion, etc.

3. We should be able to estimate them accurately and reliably.

4. It should be possible to estimate them through a computationally efficient procedure.

5. They should have a physical meaning (preferably consistent with the human auditory perception process).

Obviously, it is very difficult to select a set of acoustic features which satisfy all these requirements, and a great deal of research has been done to identify these features (see [43] and references given therein for different front-ends).

Once these features are selected, the task of the acoustic front-end is to extract these features from the speech signal. For this, it divides the speech signal into overlapping time frames and computes the values of these features for each frame. The complexity of the acoustic front-end depends on the type of features selected. These features may be as simple as the energy and zero-crossing rate of the waveform during each frame. A better, but more complex, method for feature analysis is based on the source/system model of the speech production system. It is generally considered that the system part of this model represents the vocal tract response and it contains most of the linguistic information necessary for speech recognition. The power spectrum of each speech frame contains information about the source part (in the form of fine structure) and vocal tract system part (in the form smooth spectral envelope). The task of the acoustic front-end is to compute the smooth spectral envelope from the power spectrum by removing the fine structure. Once the smooth spectral envelope is estimated, it can be represented in terms of a few parameters (such as cepstral coefficients). These parameters are used as acoustic features in a speech recognition system.

Traditionally, the power spectrum of a speech frame is computed either by fast Fourier transform (FFT) algorithm or through filter-bank analysis technique. The smooth spectral envelope is computed from this power spectrum by using one of the following two signal processing techniques: 1) linear prediction (LP) analysis and 2) homomorphic analysis. In the LP analysis technique, the smooth spectral envelope is modeled by an all-pole filter and parameters of this filter are estimated through a least-squares procedure. In the homomorphic analysis technique, a logarithmic function is used on the power spectrum which makes the source and system components additive in the log power spectrum. This allows a simple linear filter to remove the source component (fine structure) from the log power spectrum. This is done by computing an inverse Fourier transform of the log power spectrum where a first few terms (called cepstral coefficients) represent the smooth spectral envelope.

Most of the speech recognizers reported in the literature use cepstral features which are derived from the FFT power spectrum by using the LP analysis technique. These linear prediction cepstral coefficients (LPCCs) are known to be very sensitive to additive noise and channel mismatch distortions which are very common in practice. As a result, the performance of these recognition systems deteriorates drastically in the presence of these distortions. Human listeners, on the contrary, can recognize speech even in the presence of large amount of noise and channel distortions. Therefore, it is argued that the acoustic front-end can be made more robust to these distortions by utilizing the properties of human auditory system. We call these front-ends as auditory front-ends.

A number of auditory front-ends have been proposed in the literature. These front-ends employ some property of human auditory system to modify the power spectrum and then use either the LP analysis technique or the homomorphic analysis technique to get the smooth spectral envelope which, in turn, is represented in terms of a few cepstral features. Some examples of popular auditory front-ends are Mel filter-bank analysis [44], perceptual linear prediction analysis [46], ensemble interval histogram (EIH) analysis [47], etc.

The Mel filter-bank analysis procedure [44] is based on the fact that the frequency sensitivity of the human ear is higher at low frequencies than at higher frequencies. Therefore, this method computes the power spectrum of a given speech frame by using a nonuniform filter bank where filter bandwidth increases logarithmically with filter frequency (according to Mel scale). The Mel frequency cepstral coefficients (MFCCs) representing the smooth spectral envelope are computed from the power spectrum using the homomorphic analysis technique. The MFCC features have been found to be more robust to additive noise and channel mismatch distortions than the LPCC features [45].

The PLP analysis technique [46] uses more detailed properties of the human auditory system than the Mel filter-bank analysis technique to compute the power spectrum. In addition to nonuniform filter-bank (where filters are spaced according to Bark scale), it uses equal loudness curve and the intensity-loudness power law to model the auditory system better. The cepstral features are estimated from the resulting power spectrum by using the LP analysis technique. The EIH analysis technique [47] uses a measure of the spatial (tonotopic) extent of coherent neural activity across the stimulated auditory nerve to compute the power spectrum. The cepstral features are computed from this power spectrum using the LP analysis technique.

Though the cepstral features (LPCCs or MFCCs) provide a reasonable recognition performance, one of the major problems they have is that they are very sensitive to additive noise distortion. Recently, a number of techniques for robust extraction of cepstral features have been reported. These techniques exploit some special properties of the speech and interfering noise for their operation. For example, the cumulant-based LP analysis method [48] assumes that the speech signal is *non-Gaussian* and the additive noise signal is *Gaussian*. Since Gaussian processes have identically zero cumulants of all orders greater than two, the cumulants of the noisy speech signal will be insensitive to noise. The cepstral features are estimated from these robust cumulants using the all-pole modeling.

Cyclic autocorrelation-based LP analysis provides another method for robust extraction of cepstral features [49]. This method assumes that the speech signal is cyclostationary and uses cyclic autocorrelation function for computing LP parameters. Since the cyclic autocorrelation function of a stationary random signal is zero, independent of its statistical description, this analysis is robust to additive noise, white or colored.

If the additive background noise can be assumed to be a white stationary random noise process, then its 0-th lag autocorrelation coefficient has a finite positive value, all other autocorrelation coefficients are zero. If we compute the LP parameters by solving the higher-order Yule-Walker equations (which do not include the 0-th lag autocorrelation coefficient), the resulting estimate of the cepstral features will be robust to the noise. This concept has been used in the past in a number of robust feature extraction techniques [50, 51, 52, 53].

Auditory masking properties have been successfully used in the past for improving the performance of speech and audio coders [54, 55, 56, 57]. These properties have been used recently for robust extraction of cepstral features [58]. Using these properties, auditory masking threshold as a function of frequency is computed for a given speech frame from its power spectrum. All those portions of the power spectrum which are below the auditory threshold are not heard by the human auditory system due to masking effects and, hence, are discarded. These portions are replaced by the corresponding portions in the masking threshold spectrum. This modified power spectrum is processed by the LP analysis procedure to derive cepstral features for the speech frame.

Though these robust extraction techniques help in reducing the sensitivity of the cepstral features to additive noise distortion, the problem still remains. In addition, the cepstral features have another problem that they do not have any physical meaning. This makes it difficult to incorporate noise masking properties in cepstral domain during the recognition process, though one can go in a round about fashion to achieve it [59]. Because of these problems, some efforts are currently being made to investigate alternate features for robust speech recognition.

Addition of white noise to the speech signal affects the speech power spectrum at all the frequencies, but the effect is less noticeable in the higher amplitude (formant) portions of the spectrum (i.e., signal-to-noise ratio is more in the formant regions than in the non-formant regions). Since cepstrum features use formant as well as non-formant regions of the power spectrum in their computation, they become very sensitive to additive white noise. This problem can be overcome by using formant frequencies as features, as the formant locations are not disturbed by the additive noise distortion. In addition to this robustness to noise, formants have many other advantages. For example, they provide most parsimonious representation of the spectral envelope and have physical interpretation as vocal tract resonances. Because of these advantages, the formant frequencies were used as recognition features in the sixties. But they have been lately abandoned mainly due to the problems associated with their estimation from the speech signal. These problems arise due to merging of peaks in the spectrum and appearance of spurious peaks in the spectrum. These problems cause gross errors in formant extraction. If we can overcome these problems or devise features which have properties similar to formant frequencies, we can improve the speech recognition performance. Recently, the spectral subband centroids have been proposed as an alternative to formant features [60]. Preliminary results indicate the usefulness of these features for speech recognition. However, they have to be investigated more rigorously for robust speech recognition.

Since the filter-bank energies (FBEs) also have physical meaning, they can be used as recognition features. The problem with FBEs is that they are highly correlated and, therefore, provide poor recognition performance in comparison to the cepstral features. Recently, some studies have been reported which provide methods to decorrelate FBEs [61, 62]. When decorrelated FBEs are used as features, they perform better than the cepstral features for noisy speech.

Currently, the MFCCs and their first and second time derivatives are the most popular features used for speech recognition. Inclusion of time derivatives in the feature set improves the recognition performance in matched as well as mismatched acoustic conditions [63].

4.4. Feature Enhancement Techniques

Feature enhancement techniques try to suppress the effect of distorting interferences (such as additive background noise and channel mismatch) and are used after the feature extraction stage to achieve robust speech recognition. These techniques utilize certain properties of the speech signal and interfering sources to achieve feature enhancement. Cepstral liftering is one such technique which uses the property that the interfering signals show less variation in the log-power spectrum than the speech signal. Since cepstrum is obtained as an inverse Fourier transform of the log-power spectrum, the interfering signals affect the lower-quefrency cepstral coefficients more. In cepstral liftering, the lower-quefrency cepstral coefficients are weighted down with respect to the higher-quefrency cepstral coefficients [64, 65]. This technique improves the recognition performance for the dynamic time warping (DTW) based speech recognizers under the matched as well as mismatched conditions [65, 62]. But, the cepstral liftering does not offer this improvement when used with a continuous Gaussian density HMM-based speech recognizer [62]. However, when the FBE features are used with liftering effect, they provide similar improvement even for the HMM-based systems [61, 62].

Another feature enhancement technique is the cepstral mean normalization [66]. It assumes that the interfering distortion is stationary and convolutional (see Eq. (8)), and suppresses it by subtracting the long-term cepstral mean vector (over the input utterance) from the current cepstral vector. This technique is currently very popular for overcoming the channel mismatch distortion. When the channel is slowly varying with time, its effect can be removed by highpass filtering (e.g., RASTA) the sequence of cepstral feature vectors [67]-[70]. Though this technique should theoretically improve the speech recognition performance for mismatched channel conditions, it provides in practice robustness to both additive background noise and channel mismatch distortions [71].

4.3. Robust Distortion Measures

In the preceding subsection, the knowledge about the effect of additive noise on the features of clean speech has been used to devise feature enhancement techniques. This knowledge can also be used to define robust distortion (or similarity) measure. For example, it has been observed [72] that the presence of additive white noise in the speech signal causes a reduction in the norm of its cepstral vector. This observation, when cast in the perspective of a Euclidean vector space, leads to a projection distance measure. This distance measure has been found to be robust with respect additive white noise distortion when applied to a DTW-based recognizer [72]. The concept of cepstral norm reduction has also been incorporated in the HMM-based speech recognition system and good performance for noisy speech has been reported [9].

Discriminative similarity measures which are designed through a discriminative training algorithm have also been found to be robust to additive noise distortion. For example, the multi-layer perceptron (MLP) type of neural network classifier is trained through the back-propagation algorithm and provides this type of similarity measure. When used in a single-frame based vowel recognition task with cepstral features, the MLP classifier offers much better performance for noisy speech than the maximum likelihood and K-nearest classifiers [73]. Similar results have been reported for the phonetic classification task [74].

4.7. Feature and Model Compensation Techniques

Given the acoustic and language models $\Lambda_{\mathbf{X}}$ and Υ_W (computed from the training data), the task of a speech recognizer is to transcribe the unknown input utterance represented by the feature vector sequence $\mathbf{Y}$ using the MAP decision rule (Eq. (4)). In this paper, we are concentrating on the mismatch between the training and testing conditions resulting from the variability in the speech signal. This means that we have to handle the mismatch between the acoustic models $\Lambda_{\mathbf{X}}$ and the observation sequence $\mathbf{Y}$, and do not have to worry about the language model Υ_W. We can reduce this mismatch either by modifying the feature sequence $\mathbf{Y}$ (feature compensation) or the acoustic models $\Lambda_{\mathbf{X}}$ (model compensation). Some of the feature and model compensation techniques are described below.

4.7.1. Training-based compensation: As mentioned earlier, one way to achieve robust speech recognition is to train the recognizer afresh (from the scratch) every time the test condition changes. This is practically impossible because every time we have to collect a large amount of training data, and train the system which is computationally very expensive. Another way is to capture some information about the mismatch during the training phase and use it to perform robust speech recognition.

One way to do it is to collect a small set of stereo training data [75]; i.e., a pair of stereo training sets, one corresponding to speech utterances recorded in clean (undistorted) condition and the other corresponding to same utterances recorded simultaneously in presence of distortion. This stereo training data can then be used to derive a mapping in the feature space which can be used to clean the distorted feature vectors in the test phase. Several methods have been reported in the literature to derive this mapping. For example, the probabilistic optimum filtering method [76] finds the mapping in the form of a piecewise linear transformation, estimated by quantizing the feature space into a set of distinct regions and computing a set of filters optimum in the mean square error sense. A similar approach has been used to derive a family of environment dependent cepstral compensation methods [75].

It is not always possible to have stereo training data (e.g., when the mismatch is due to inter-speaker variability). In such cases, one uses a pair of training utterances representing the same text, but recorded separately under two mismatched conditions. The two training utterances are aligned in time using the DTW algorithm prior to their use in deriving the transformation (or mapping) [77, 78].

In the parallel model combination method [79, 80], a statistical model (e.g., HMM) of the background noise is constructed during the training phase. During the testing phase, the clean HMMs $\Lambda_{\mathbf{X}}$ are combined with this noise HMM to recognize the sequence $\mathbf{Y}$. This is done by carrying out Viterbi search in the combined state-space of two models. Since this method provides a framework for incorporating independent concurrent signals (speech and noise), it can be used to handle non-stationary interfering signals. For example, a multi-state HMM can used to capture the changing statistical characteristics of the non-stationary noise signal. This method is closely related to the HMM decomposition method [81]. The main difference between the two methods is that clean speech features are estimated from the noisy features during the recognition phase in the HMM decomposition method. Methods reported in [82, 83] are also related to the HMM decomposition method.

4.7.2. Adaptation-based compensation: The methods using adaptation-based compensation require a small amount of data collected at the testing stage for feature or model adaptation. These methods have been originally developed for speaker adaptation. However, they are equally useful for handling other sources of mismatch (background noise, microphone and channel mismatch distortion). Here, we describe these methods in the context of speaker adaptation.

In a typical speaker adaptation scenario, speaker independent (SI) HMMs are computed during the training phase from a large collection of data coming from a number of speakers. Adaptation is carried out for a new speaker (who is going to use the system during the testing phase) either in feature space or in model space using a small amount of speaker-specific adaptation data.

Let us denote this adaptation data by $\mathcal{Y}$. Assume that the transcription of this adaptation data is available. Let us denote this transcription by $\mathcal{W}$. This data is utilized to design a transformation in the model space

$$\Lambda_{\mathbf{Y}} = G_{\eta}(\Lambda_{\mathbf{X}}), \tag{10}$$

where G is the transformation whose functional form can be assumed to be known from our prior knowledge of the source of mismatch and η are the associated parameters. These parameters are estimated in such a way so as to provide best match between the transformed models $\Lambda_{\mathbf{Y}}$ and the adaptation data $\mathcal{Y}$. In the maximum likelihood formulation, their estimates $\hat{\eta}$ are found as follows:

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \Pr(\mathcal{Y}|\eta, \mathcal{W}, \Lambda_{\mathbf{X}}) \cdot Pr(\mathcal{W}|\Upsilon_W). \tag{11}$$

The maximization in this equation can be carried out using the expectation-maximization algorithm [84].

We have described above an adaptation procedure using a transformation in the model space. A similar procedure can be developed for a feature space based transformation.

Several functional forms of the transformation in Eq. (10) have been tried out in the literature and procedures for estimating their parameters have been developed. Some of these functional forms include a simple cepstral bias [84, 85, 86], linear affine transformation [87, 88, 89] and nonlinear transformation realized through an MLP [90]. The linear affine transformation is currently the most popular choice and the resulting formulation (given by Eq. (11)) is called the maximum likelihood linear regression (MLLR) method [88, 89].

So far, we have described methods which use the adaptation speech with its transcription in a batch mode. It is possible to carry out speaker adaptation with an unlabelled adaptation data (unsupervised mode). We can also update the transformation parameters as new adaptation data becomes available (incremental mode).

Instead of using the transformation based adaptation, one can adapt the SI HMMs directly using the MAP algorithm [91, 92]. This algorithm incorporates the prior knowledge of SI HMM parameters to get the MAP estimate for the new speaker using his/her adaptation data. Though the MAP algorithm provides an optimal solution, it has the problem that it converges slowly and requires a relatively large amount of adaptation data. The MAP algorithm can be combined with the transformation-based methods to get better and fast adaptation performance [93, 94, 95, 96]. Recently there has been an interest in fast adaptation techniques (such as the cluster adaptive training [97] and eigenvoice techniques [98]), which use a very small amount of adaptation data to adapt to a new speaker or invironment.

4.7.3. Self-adaptation based compensation: In self-adaptation based compensation, no adaptation data is available. Here, only information available to the recognition system is the observed feature sequence $\mathbf{Y}$ and the models ($\Lambda_{\mathbf{X}}$ and Υ_W). The system has to do the adaptation of the models ($\Lambda_{\mathbf{X}}$) using the observed feature sequence $\mathbf{Y}$ and at the same time perform the recognition task. Assuming a model-space based transformation (Eq. (10), the dual task of adaptation and recognition can be achieved through the following equation:

$$(\hat{\eta}, \hat{W}) = \underset{(\eta, W)}{\operatorname{argmax}} \Pr(\mathbf{Y}|\eta, W, \Lambda_{\mathbf{X}}) \cdot Pr(W|\Upsilon_W). \tag{12}$$

This joint optimization over (η, W) is a very difficult task. An iterative suboptimal procedure is generally used for this purpose where the maximization is carried out sequentially in two steps [84, 85, 71]. In the first step, the transformation parameters η are assumed to be known from the previous iteration and are used to carry out maximization over W as follows:

$$\hat{W} = \underset{W}{\operatorname{argmax}} \Pr(\mathbf{Y}|\eta, W, \Lambda_{\mathbf{X}}) \cdot Pr(W|\Upsilon_W). \quad (13)$$

In the second step, the estimated word sequence $\hat{W}$ is used to carry out maximization over η as follows:

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \Pr(\mathbf{Y}|\eta, \hat{W}, \Lambda_{\mathbf{X}}) \cdot Pr(\hat{W}|\Upsilon_W). \quad (14)$$

Thus, the first step does recognition and the second step performs adaptation. These steps are iterated until the convergence is reached.

SUMMARY

In this paper, we have addressed the topic of robust speech recognition. Mismatch between the training and testing conditions causes a severe degradation in the speech recognition performance. The aim of a robust speech recognition is to overcome this mismatch problem and provide a moderate and graceful degradation in the recognition performance. We have concentrated here on the mismatch problem resulting from the variability of the speech signal. The sources for this variability include additive background noise, channel and microphone mismatches, speaker mismatch, and different accents, stress types and speaking styles. A number of robust speech recognition techniques have been briefly described.

REFERENCES

[1] K.H. Davis, R. Biddulph and S. Balashek, "Automatic recognition of spoken digits", *J. Acoust. Soc. Am.*, Vol. 24, p. 637, 1952.

[2] H. Dudley and S. Balashek, "Automatic recognition of phonetic patterns in speech", *J. Acoust. Soc. Am.*, Vol. 30, pp. 721-732, 1958.

[3] C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), *Automatic Speech Recognition: Advanced Topics*, Kluwer Academic Publishers, Boston, 1996.

[4] S. Young, "A review of large-vocabulary continuous-speech recognition", *IEEE Signal Processing Magazine*, pp. 45-57, Sept. 1996.

[5] S. Das, R. Bakis, A. Nadas, D. Nahamoo and M. Picheny, "Influence of background noise and microphone on the performance of the IBM Tangora speech recognition system", *Proc. ICASSP*, pp. 71-74, 1993.

[6] R. Cardin, Y. Normandin and E. Millie, "Inter-word coarticulation modeling and MMIE training for improved connected digit recognition", *Proc. ICASSP*, pp. 243-246, 1993.

[7] L.R. Rabiner, "Applications of voice processing to communications", *Proc. IEEE*, Vol. 82, No. 2, pp. 199-228, Feb. 1994.

[8] J.C. Junqua, "Impact of the unknown communication channel on automatic speech recognition: A review", *Proc. EUROSPEECH*, pp. KN29-KN32, 1997.

[9] B.H. Juang and K.K. Paliwal, "Hidden Markov models with first-order equalization for noisy speech recognition", *IEEE Trans. Signal Processing*, Vol. 40, pp. 2136-2143, Sept. 1992.

[10] A. Acero and R.M. Stern, "Environmental robustness in automatic speech recognition", *Proc. ICASSP* pp. 849-952, 1990.

[11] J.C Junqua and J.P. Haton (Eds.), *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, 1996.

[12] B.H. Juang, "Speech recognition in adverse environments", *Computer Speech and Language*, Vol. 5, pp. 275-294, 1991.

[13] Y. Gong, "Speech recognition in noisy environments: A survey", *Computer Speech and Language*, Vol. 16, pp. 261-291, 1995.

[14] S. Furui, "Recent advances in robust speech recognition", *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, April 1997, pp. 11-20.

[15] J.R. Bellegarda, "Statistical techniques for robust ASR: Review and Perspectives", *Proc. EUROSPEECH*, pp. KN33-KN36, 1997.

[16] C.H. Lee, "Adaptive compensation for robust speech recognition", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 357-364, 1997.

[17] C.H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition", *Speech Communication*, Vol. 25, pp. 29-47, Aug. 1998.

[18] H. Ney and X. Aubert, "Dynamic programming search strategies: From digit strings to large vocabulary word graphs", in *Automatic Speech Recognition: Advanced Topics*, C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), Kluwer Academic Publishers, Boston, 1996, pp. 384-411.

[19] P.S. Gopalakrishnan and L.R. Bahl, "Fast search techniques", in *Automatic Speech Recognition: Advanced Topics*, C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), Kluwer Academic Publishers, Boston, 1996, pp. 413-428.

[20] R. Schwartz, L. Nguyen and J. Makhoul, "Multiple-pass search strategies", in *Automatic Speech Recognition: Advanced Topics*, C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), Kluwer Academic Publishers, Boston, 1996, pp. 429-456.

[21] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973.

[22] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. IEEE*, Vol. 77, pp. 257-286, Feb. 1989.

[23] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[24] J.C. Junqua and Y. Anglade, "Acoustic and perceptual studies of Lombard speech: Application to isolated-words automatic speech recognition", *Proc. ICASSP*, pp. 841-844, 1990.

[25] S. Nakamura, T. Takiguchi and K. Shikano, "Noise and room acoustics distorted speech recognition by HMM composition", *Proc. ICASSP*, pp. 69-72, 1996.

[26] B.T. Lilly and K.K. Paliwal, "Effect of speech coders on speech recognition performance", *Proc. ICSLP*, pp. 2344-2347, Oct. 1996.

[27] J.S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth compression of noisy speech", *Proc. IEEE*, Vol. 67, pp. 1586-1604, Dec. 1979.

[28] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. ASSP-27, pp. 113-120, 1979.

[29] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by additive noise", *Proc. ICASSP*, pp. 208-211, 1979.

[30] K.K. Paliwal and A. Basu, "A speech enhancement based on Kalman filtering", *Proc. ICASSP*, pp. 177-180, 1987.

[31] M. Dendrinos, S. Bakamidis and G. Carayannis, "Speech enhancement from noise: A regenerative approach", *Speech Communication*, Vol. 10, No. 2, pp 45-57, Feb. 1991.

[32] S. Jensen, P. Hansen, S. Hansen and J. Sorensen, "Reduction of Broad-Band Noise in Speech by Truncated QSVD", *IEEE Trans. Speech and Audio Processing*, Vol. 3, No. 6, pp. 439-448, Nov. 1995.

[33] D. Van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system", *Computer Speech and Language*, Vol. 3, pp. 151-167, 1989.

[34] P. Lockwood, and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection for robust speech recognition in cars", *Speech Communication*, Vol. 11, pp. 215-228, 1992.

[35] E. Bernstein and W. Evans, "Wavelet based noise reduction for speech recognition", *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, April 1997, pp. 111-114.

[36] B.T. Lilly and K.K. Paliwal, "Robust speech recognition using singular value decomposition based speech enhancement", *Proc. IEEE Region 10 Conf. on Speech and Image Technologies for Computing and Communications*, Brisbane, Australia, pp. 257-260, Dec. 1997.

[37] G. Powell, P. Darlington and P. Wheeler, "Practical adaptive noise reduction in the aircraft cockpit environment", *Proc. ICASSP*, pp. 173-176, 1987.

[38] Y. Nakadai and N. Sugamura, "A speech recognition method for noise environments using dual inputs", *Proc. ICSLP*, pp. 1141-1144, 1990.

[39] J.L. Flanagan, J.D. Johnston, R. Zahn and G.W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms", *J. Acoust. Soc. Amer.*, Vol. 78, pp. 1508-1518, 1985.

[40] H.F. Silverman and S.E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone data", *Computer Speech and Language*, Vol. 6, pp. 129-152, 1992.

[41] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverbrant environment using CSP analysis", *Proc. ICASSP*, pp. 921-924, 1996.

[42] E. Lleida, J. Fernandez and E. Masgrau, "Robust continuous speech recognition system based on a microphone array", *Proc. ICASSP*, pp. 241-244, 1998.

[43] J.W. Picone, "Signal Modeling techniques in speech recognition", *Proc. IEEE*, Vol. 81, No. 9, pp. 1215-1247, Sept. 1993.

[44] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech and Signal Processing*, Vol. 28, No. 4, pp. 357-366, Aug. 1980.

[45] K.K Paliwal and B.S. Atal, "A comparative study of feature representations for robust

speech recognition in adverse environments", *Proc. ICSLP*, pp. 1015-1018, 1994.

[46] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, Vol. 87, No. 4, pp. 1738-1752, Apr. 1990.

[47] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environments", *Computer Language and Speech*, Vol. 1, pp. 109-130, 1986.

[48] K.K. Paliwal and M.M. Sondhi, "Recognition of noisy speech using cumulant-based linear prediction analysis", *Proc. ICASSP*, pp. 429-432, May 1991.

[49] K.K. Paliwal and Y. Sagisaka, "Cyclic autocorrelation-based linear prediction analysis of speech", *Proc. EUROSPEECH*, pp. 279-282, 1997.

[50] Y.T. Chan and R.P. Langford, "Spectral estimation via the high-order Yule-Walker equations", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-30, pp. 689-698, Oct. 1982.

[51] K.K. Paliwal, "A noise-compensated long correlation matching method for AR spectral estimation of noisy signals", *Proc. ICASSP*, pp. 1369-1372, 1986.

[52] D. Mansour and B.H. Juang, "The short-time modified coherence representation and its application for noisy speech recognition", *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-37, pp. 795-804, 1989.

[53] J. Hernando, C. Nadeu and E. Lleida, "On the AR modeling of the one-sided autocorrelation sequence for noisy speech recognition", *Proc. ICSLP*, pp. 1593-1596, 1992.

[54] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP): High quality speech at very low bit rates", *Proc.ICASSP*, pp. 937-940, 1985.

[55] D. Sen and W.H. Holmes, "Perceptual enhancement of CELP speech coders", *Proc. ICASSP*, Vol. II, pp. 105-108. 1994.

[56] G. Theile, G. Stoll and M. Link, "Low bit-rate coding of high quality audio signals: An introduction to the MASCAM system", *EBU Review – Technical*, No. 230, Aug. 1988.

[57] J.D. Johnston and K. Brandenburg, "Wideband coding – Perceptual considerations for speech and music", in *Advances in Speech Signal Processing*, S. Furui and M.M. Sondhi (Eds.), Marcel Dekker, New York, pp. 109-140, 1992.

[58] B.T. Lilly and K.K. Paliwal, "Auditory masking based acoustic front-end for robust speech recognition", *Proc. IEEE Region 10 Conf. on Speech and Image Technologies for Computing and Communications*, Brisbane, Australia, pp. 165-168, Dec. 1997.

[59] A. Varga and R. Moore, "Hidden Markov model decomposition of speech and noise", *Proc. ICASSP*, pp. 845-848, 1990.

[60] K.K. Paliwal, "Spectral subband centroids as features for speech recognition", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 124-131, 1997.

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

[62] K.K. Paliwal, "Decorrelated and liftered filter-bank energies for robust speech recognition", *Proc. EUROSPEECH*, 1999.

[63] B.A. Hanson, T.H. Appelbaum and J.C. Junqua, "Spectral dynamics for speech recognition under adverse conditions", in *Automatic Speech Recognition: Advanced Topics*, C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), Kluwer Academic Publishers, Boston, 1996, pp. 331-356.

[64] K.K. Paliwal, "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition", *Speech Communication*, pp. 151-154, May 1982.

[65] B.A. Hanson and H. Wakita, "Spectral slope distance measures with linear prediction analysis for word recognition in noise", *IEEE Trans. Acoust., Speech and Signal Processing* Vol. ASSP-35, No. 7, pp. 968-973, July 1987.

[66] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Amer.*, Vol. 55, pp. 1304-1312, June 1974.

[67] D. Geller, R.H. Umbach and H. Ney, "Improvements in speech recognition for voice dialing in car environment", *Proc. ECSA Workshop on Speech Processing in Adverse Conditions*, pp. 203-206, 1992.

[68] H. Murveit, J. Butzberger and M. Weintraub, "Reduced channel dependence for speech recognition", *Proc. Speech and Natural Language Workshop (DARPA)*, pp. 280-284, 1992.

[69] H. Hermansky and N. Morgan, "RASTA processing of speech", *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589, Oct. 1994.

[70] K. Aikawa, H. Singer, H. Kawahara and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition", *Proc. ICASSP* Vol. 2, pp. 668-671, 1993.

[71] K.K Paliwal, "A maximum likelihood equalization technique for robust speech recognition in adverse environments", *Proc. EUROSPEECH*, pp. 1521-1524, Sept. 1995.

[72] D. Mansour and B.H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition", *IEEE Trans. Acoust., Speech and Signal Processing* Vol. ASSP-37, pp. 1659-1671, 1989.

[73] K.K. Paliwal, "Neural net classifiers for robust speech recognition under noisy environments", *Proc. ICASSP*, pp. 429-432, 1990.

[74] H. Leung, B. Chigier and J. Glass, "A comparative study of signal representations and classification techniques for speech recognition", *Proc. ICASSP*, Vol. II, pp. 680-683, 1993.

[75] R.M. Stern, A. acero, F.H. Liu and Y. Ohshima, "Signal processing for robust speech recognition", in *Automatic Speech Recognition: Advanced Topics*, C.H. Lee, F.K. Soong and K.K. Paliwal (Eds.), Kluwer Academic Publishers, Boston, 1996, pp. 357-384.

[76] L. Neumeyer and M. Weintraub, "Probabilistic optimum filtering for robust speech recognition", *Proc. ICASSP*, Vol. I, pp. 417-420, 1994.

[77] K. Choukri, G. Chollet and Y. Grenier, "Spectral transformations through canonical correlation analysis for speaker adaptation in ASR", *Proc. ICASSP*, pp. 2659-2662, 1986.

[78] S. Nakamura and K. Shikano, "A comparative study of spectral mapping for speaker adaptation", *Proc. ICASSP*, pp. 157-160, 1990.

[79] M.J.F. Gales and S.J. Young, "An improved approach to the hidden Markov model decomposition of speech and noise", *Proc. ICASSP*, pp. 233-236, 1992.

[80] F. Martin, K. Shikano and Y. Minami, "Recognition of noisy speech by composition of hidden Markov models", *Proc. EUROSPEECH*, pp. 1031-1034, 1993.

[81] A.P. Varga and R.K. Moore, "Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition", *Proc. EUROSPEECH*, pp. 1175-1178, 1991.

[82] R. Rose, E. Hofstetter and D. Reynolds, "Integrated models of speech and background with application to speaker identification in noise", *IEEE Trans. Speech and Audio Processing*, Vol. 2, pp. 245-257, 1994.

[83] Y. Ephraim, "Gain adapted hidden Markov models for recognition of clean and noisy speech", *IEEE Trans. Signal Processing*, Vol. 40, pp. 1303-1316, 1992.

[84] A. Sankar and C.H. Lee, "A maximum likelihood approach to stochastic matching for robust speech recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 4, pp. 190-202, 1996.

[85] M.G. Rahim and B.H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 4, pp. 19-30, 1996.

[86] Y. Zhao, "An acoustic-phonetic based speaker adaptation technique improving speaker independent continuous speech recognition", *IEEE Trans. Speech and Audio Processing*, Vol. 2, pp. 380-394, 1994.

[87] V.V. Digalakis, D. Rtischev and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures", *IEEE Trans. Speech and Audio Processing*, Vol. 3, pp. 357-366, 1995.

[88] C.J. Leggetter and P.C. Woodland, "Flexible speaker adaptation for large vocabulary speech recognition", *Computer Speech and Language*, Vol. 9, pp. 171-186, 1995.

[89] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework", *Computer Speech and Language*, Vol. 10, pp. 249-264, 1996.

[90] V. Abrash, A. Sankar, H. Franco and M. Cohen, "Acoustic adaptation using nonlinear transformations of HMM parameters", *Proc. ICASSP*, pp. 729-732, 1996.

[91] C.H. Lee, C.H. Lin and B.H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models", *IEEE Trans. Signal Processing*, Vol. 39, pp. 806-814, 1991.

[92] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE*

Trans. Speech and Audio Processing, Vol. 2, pp. 291-298, 1994.

[93] V.V. Digalakis and L.G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods", *IEEE Trans. Speech and Audio Processing*, Vol. 4, pp. 294-300, 1996.

[94] G. Zavaliagkos, R. Schwartz and J. McDonough, "Maximum a posteriori adaptation for large scale HMM recognizers", *Proc. ICASSP*, pp. 725-728, 1996.

[95] V. Nagesh and L. Gillick, "Studies in transformation-based adaptation", *Proc. ICASSP*, pp. 1031-1034, 1997.

[96] J. Ishii and M. Tonomura, "Speaker normalization and adaptation based on linear transformation", *Proc. ICASSP*, pp. 1055-1058, 1997.

[97] M.J.F. Gales, "Cluster adaptive training for speech recognition", *Proc. ICSLP*, pp. 1783-1786, 1998.

[98] R. Kuhn, P. Nguyen, J.C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field and M. Contolini, "Eigenvoices for speaker adaptation", *Proc. ICSLP*, pp. 1771-1774, 1998.