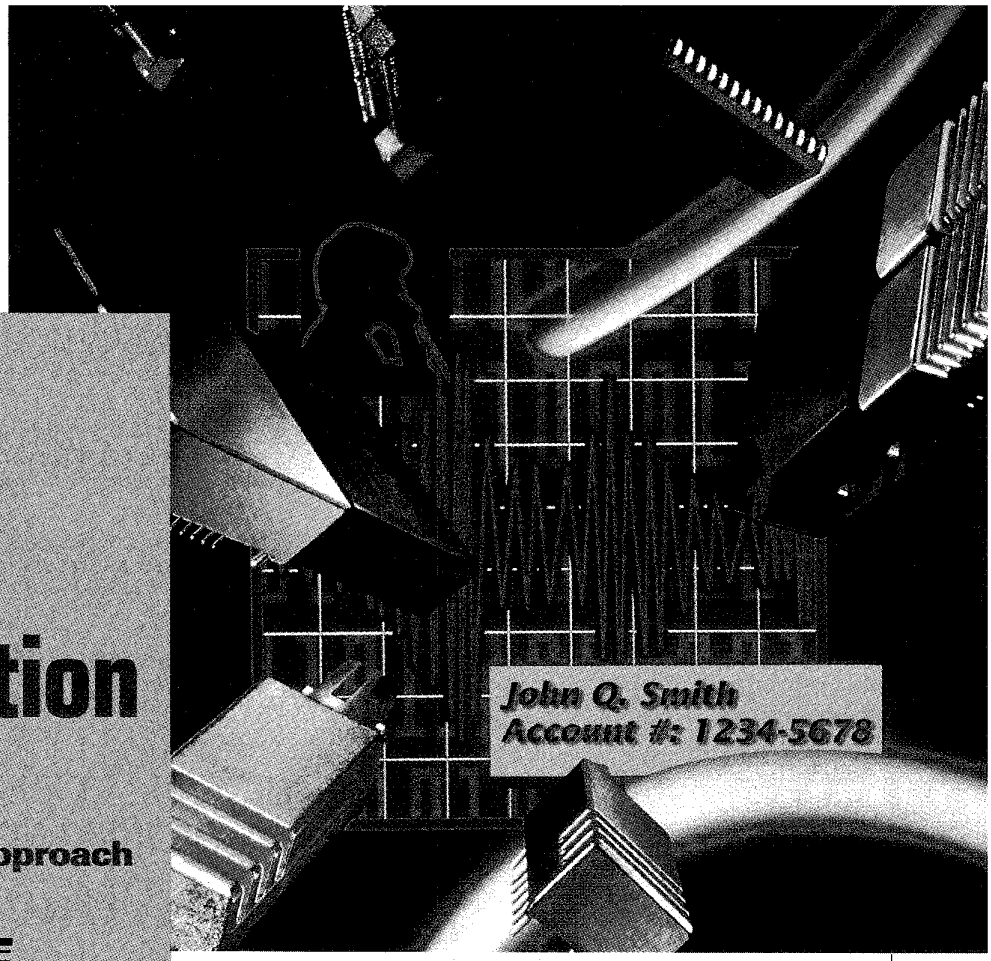
Robust Speaker Recognition

A Feature-based Approach

RICHARD J. MAMMONE
XIAOYU ZHANG
RAVI P. RAMACHANDRAN



The future commercialization of speaker- and speech-recognition technology is impeded by the large degradation in system performance due to environmental differences between training and testing conditions. This is known as the "mismatched condition." Studies have shown [1] that most contemporary systems achieve good recognition performance if the conditions during training are similar to those during operation (matched conditions). Frequently, mismatched conditions are present in which the performance is dramatically degraded as compared to the ideal matched conditions. A common example of this mismatch is when training is done on clean speech and testing is performed on noise- or channel-corrupted speech. Robust speech techniques [2] attempt to maintain the performance of a speech processing system under such diverse conditions of operation.

This article presents an overview of current speaker-recognition systems and the problems encountered in operation, and it focuses on the front-end feature extraction process of robust speech techniques as a method of improvement. Linear predictive (LP) analysis, the first step of feature extraction, is discussed, and various robust cepstral features derived from LP coefficients are described. Also described is the *affine transform,* which is a feature transformation approach that integrates mismatch to simultaneously combat both channel and noise distortion.

Overview

Speaker and Speech Recognition

A speaker-recognition system attempts to recognize a speaker by his/her voice. The system can be either text-dependent (constraint on what is spoken) or text-independent (no constraint on what is spoken). The idea is to identify the inherent differences in the articulatory organs (the structure of the vocal tract, the size of the nasal cavity, and vocal cord characteristics) and the manner of speaking. Speech recognition, on the other hand, is the task of understanding what is being said rather than who is speaking. First, the stream of sounds comprising the incoming speech must be recognized.

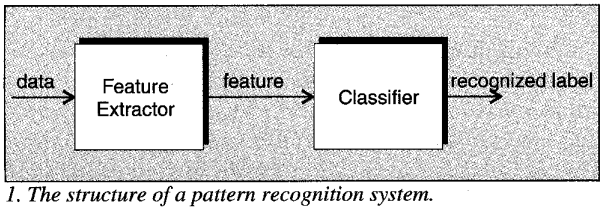
1053-5888/96/$5.00©1996IEEE

A language model can then be applied to the sequence of recognized sounds to improve performance through the use of contextual information.

Pattern Recognition

Speaker recognition and speech recognition are subsets of a more general area known as pattern recognition. Given the features that describe the properties of an object, a pattern-recognition system aims to recognize the object based on its previous knowledge of the object. Three stages are generally involved in building a pattern-recognition system: training, testing, and implementation. In the training stage, a set of parameters of the model is estimated so that in some sense the model "learns" the correspondence between the features and the labels of the objects. One such learning criterion is to minimize the overall estimation error. In the testing stage, the parameters of the model are then adjusted using a set of cross-validation data to achieve a good generalization of the performance of the system. The cross-validation data usually consists of a set of features and labels that are different from the training data. The task of recognition is carried out in the implementation stage, where the feature with an unknown label is passed through the system and assigned a label at the output.

A pattern-recognition system basically consists of a front-end feature extractor and a classifier, as shown in Fig. 1. The feature extractor normalizes the collected data and transforms them to the feature space. In feature space, the data are compressed and represented in such an effective way that objects from the same class behave similarly and a clear distinction among objects from different classes exists. The classifier takes the features computed by the feature extractor and performs either template matching or probabilistic-likelihood computation on the features, depending on the type of algorithm employed. Before it can be used for classification, the classifier has to be trained so that a mapping from the feature to the label of a particular class is established. Since an object is characterized in the classifier by a module or a part of an integrated model, training is also the stage of enrollment. Such an approach has been demonstrated to be efficient in performing a pattern-recognition task. However, the implicit assumption of this approach is that the training and testing conditions are comparable. Problems arise when there is a mismatch between the environments for training and testing, which is generally true in most applications. For example, assume that speaker recognition is carried out over the telephone network. It is very likely that a test speaker calls from a telephone other than the one used at the time of training (communication-channel mismatch).



1. The structure of a pattern recognition system.

Robust Speech Techniques

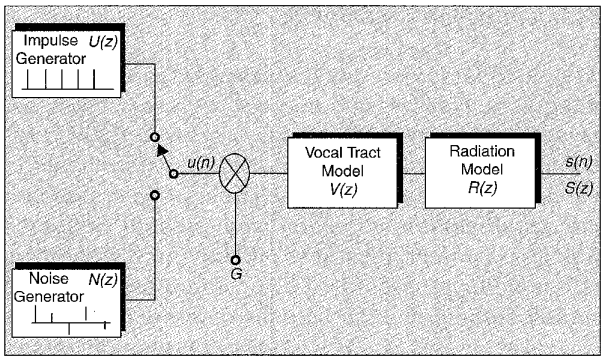
Two strategies of robust speech techniques have emerged to mitigate the problems that arise due to channel effects and noise. The first strategy is normally carried out in the front-end feature extractor before the feature vectors are passed to the classifier for comparison and labeling. One method is to enhance the speech by spectral subtraction [3] so that the features are more representative of clean speech in that noise effects are suppressed. Also, if the features are the cepstral vectors, mean removal [4] attempts to remove the transmission channel effect. Efforts are also made to find new features that are robust to noise and channel effects (one example is the *short-time modified coherence* [5]). The second strategy aims at making the classifier more robust by compensating for the distortions at the classification stage. Statistical approaches are usually adopted to obtain the probabilistic modeling of features so that a robust mapping from the testing data to the training data can be created. Methods such as probabilistic optimum filtering [6], Gaussian mixture model (GMM) [7], and hidden Markov model (HMM) adaptation [8] all fall into this category. Also, robust distance metrics such as the Itakura spectral distortion measure [9, 10] and the projection measure [11] will lead to more accurate labeling of the test data.

Linear Prediction of Speech

The general feature-extraction step of interest here can be divided into two parts. First, LP analysis of speech is carried out to produce a set of predictor coefficients. Second, the predictor coefficients are transformed into feature vectors. In this section, we discuss the rationale behind the use of LP analysis, give some interpretations, and point out the computational aspects of LP analysis.

Autoregressive Model

Speech sounds can be classified into three distinct classes: voiced sounds, fricative or unvoiced sounds, and plosive sounds. The speech waveform is an acoustic pressure wave that originates from voluntary physiological movements of anatomical structures such as the vocal cords, vocal tract, nasal cavity, tongue, and lips [12, 13]. The vocal tract is usually modeled as a concatenation of nonuniform lossless tubes of varying cross-sectional area that begins at the vocal cords and ends at the lips [13]. The opening of the vocal cords is called the glottis. Voiced sounds such as /i/ and /e/ are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxed oscillation and thereby excite the vocal tract with quasi-periodic pulses of air. The greater the tension, the higher the pitch or fundamental frequency of the voice. Unvoiced sounds are generated by voluntarily holding the vocal cords open, forming a constriction using the articulator, and forcing the air through the constriction at a high enough velocity to produce turbulence. The vocal tract is excited by a broad-band noise



2. The linear vocal-tract model for speech production.

source during the production of unvoiced sounds. Plosive sounds result from building up air pressure in the mouth and abruptly releasing it.

A linear model of speech production was developed by Fant in the late 1950s [14] where the glottal pulse, vocal tract, and radiation are individually modeled as linear filters. A complete model of speech production represented in the z-transform domain is shown in Fig. 2. The source is either a quasi-periodic impulse sequence for the voiced sounds or a random noise sequence for unvoiced sounds with a gain factor G set to control the intensity of the excitation. The transfer function $V(z)$ for the vocal tract relates volume velocity at the source to volume velocity at the lips. It is generally an all-pole model for most speech sounds. Each pole of $V(z)$ corresponds to a formant or resonance of the sound. For nasals and fricatives that require both resonances and anti-resonances (poles and zeros), an all-pole model is still preferred because the effect of a zero in the transfer function can be achieved by including more poles [15]. The radiation model $R(z)$ describes the air pressure at the lips, which can be reasonably approximated by a first-order backward difference. Combining the glottal pulse, vocal tract, and radiation yields a single all-pole transfer function [13, 14] given by

$$H(z) = G(z)V(z)R(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \quad . \tag{1}$$

With this transfer function, we get a difference equation for synthesizing the speech samples $s(n)$ as

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n). \tag{2}$$

It can be noted that $s(n)$ is predicted as a linear combination of the previous p samples. Therefore, the speech production model is often called the *linear prediction* (LP) model, or the *autoregressive model.*

Modulation Model Representation of Speech

The transfer function $H(z)$ in Eq. 1 can be rewritten as

$$H(z) = \frac{1}{A(z)} = \prod_{i=1}^{p} \frac{1}{1 - z_i z^{-1}} = \sum_{i=1}^{p} \frac{r_i}{1 - z_i z^{-1}} \tag{3}$$

where r_i represents the residues and z_i represents the poles of $H(z)$. The poles are expressed as

$$z_i = \sigma_i e^{j\omega_i}, \quad i = 1,2,\ldots,p, \tag{4}$$

where ω_i corresponds to the i^{th} center frequency. The magnitude of the poles are denoted σ_i, which falls into the range (0,1). The bandwidth of the i^{th} pole is defined as [16]

$$B_i = \frac{1}{\pi} \ln\left(\frac{1}{|z_i|}\right) = \frac{1}{\pi} \ln\left(\frac{1}{\sigma_i}\right). \tag{5}$$

Thus, the vocal-tract model corresponds to the causal impulse response given by

$$h(n) = \sum_{i=1}^{p} r_i z_i^n = \sum_{i=1}^{p} r_i \sigma_i^n e^{j\omega_i n} \tag{6}$$

which, in turn, is also the form of the homogeneous solution to Eq. 2. The speech signal $s(n)$ is a multicomponent signal expressed as a linear combination of amplitude- and phase-modulated exponentials that are specified by the autoregressive model. This is a special case of the more general modulation model for speech as discussed in [16]. For each component, there are three parameters, namely, r_i, σ_i, and, ω_i. The parameters r_i and σ_i specify the amplitude-modulated portion, while ω_i is the parameter for phase modulation. A pole close to the unit circle signifies a formant at ω_i with a relatively low bandwidth.

Computational Aspects

In practice, the *predictor coefficients* $\{a_i\}$ describing the autoregressive model must be computed from the speech signal. Since speech is time-varying in that the vocal-tract configuration changes over time, an accurate set of predictor coefficients is adaptively determined over short intervals (typically 10 ms to 30 ms) called frames, during which time-invariance is assumed. The gain G is usually ignored to allow the parameterization to be independent of the signal intensity. The *autocorrelation* method and the *covariance* method are two standard methods of solving for the predictor coefficients [12, 17]. Both approaches are based on minimizing the mean-square value of the estimation error $e(n)$ as given by

$$e(n) = s(n) - \sum_{i=1}^{p} a_i s(n-i) \tag{7}$$

The methods differ with respect to the details of implementation. The autocorrelation method is computationally simpler than the covariance approach and, unlike its covariance counterpart, assures that all the poles of $H(z)$ lie within the unit circle. Specifically consider the autocorrelation method. The mean-square error is minimized over a frame of

N samples. Moreover, it is assumed that the speech samples are identically zero outside the frame of interest. If the autocorrelation of the signal $s(n)$ is defined as

$$r_s(k) = \sum_{n=0}^{N-1-k} s(n)s(n+k), \tag{8}$$

then the predictor coefficients a_i can be obtained by solving the following set of equations

$$\begin{pmatrix} r_s(0) & r_s(1) & \cdots & r_s(p-1) \\ r_s(1) & r_s(0) & \cdots & r_s(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_s(p-1) & r_s(p-2) & \cdots & r_s(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} r_s(1) \\ r_s(2) \\ \vdots \\ r_s(p) \end{pmatrix}. \tag{9}$$

Denoting the $p \times p$ Toeplitz autocorrelation matrix on the left-hand side by $\mathbf{R_S}$, the predictor coefficient vector by $\mathbf{a}$, and the right-hand-side vector of autocorrelation coefficients by $\mathbf{r_S}$, we have

$$\mathbf{R_S a} = \mathbf{r_S} \tag{10}$$

Therefore,

$$\mathbf{a} = \mathbf{R_S}^{-1}\,\mathbf{r_S}. \tag{11}$$

Since the matrix $\mathbf{R_S}$ is Toeplitz, a computationally efficient algorithm known as the Levinson-Durbin recursion can be used to solve this system of equations [13]. Upon solving for $H(z)$, the magnitude response $|H(e^{j\omega})|$ represents the spectral envelope of the speech.

A robust solution technique will result in the vocal-tract information being captured by $H(z)$, whether speech is clean or corrupted by noise and/or channel effects. Then, the predictor coefficients would either be invariant or show very little variation when speech is corrupted. Subsequently, the features would be naturally robust. A comparative study of different approaches to find the predictor coefficients based on minimizing various objective functions was done in [18]. In addition to the standard LP autocorrelation and covariance approaches, methods based on the least-absolute-value criterion, the iterative weighted least-squares criterion, and the total least-squares criterion were considered. It was found that the best method depends on the type of noise present and, hence, a universally acceptable solution is not known [18].

Another attempt at representing the speech spectrum involves an approximation that gives more emphasis to those frequencies that have greater auditory prominence. This is known as perceptual linear prediction (PLP) [19]. The actual speech spectrum (obtained by a DFT of the speech samples) is modified based on the principles of critical-band auditory masking and the unequal sensitivity of human hearing at different frequencies [19]. This modified spectrum is approximated by an autoregressive model to obtain $H(z)$. The autocorrelation values are obtained as the inverse DFT of the modified spectrum and Eq. 10 is solved to obtain the predictor coefficients. It has been recently shown that the PLP technique is more robust to some mismatched environments than the standard LP autocorrelation method for large-vocabulary speech recognition [20]. In this article, we proceed by assuming that the standard LP autocorrelation method is used to find the predictor coefficients.

Robust Cepstral Analysis

The next step is to convert the predictor coefficients into feature vectors. Examples of such vectors include [15] the predictor coefficients themselves, cepstral coefficients and their derivatives, line spectral pairs (LSP), log area ratios (LAR), vocal-tract area functions, and the impulse response $h(n)$ of the filter $H(z)$. For speaker recognition, some of the above features were compared and the cepstral coefficients were found to provide the best results [21]. Also, the derivatives of the cepstral coefficients capture the temporal information in speech that is essential for text-dependent tasks. Although the line spectral pairs recently have been shown to have some promise [22], emphasis will be put on cepstrum-related features in this article. All the cepstrum-related features described are obtained after LP analysis, with the exception of the mel-warped cepstrum that is obtained from a filter-bank analysis [17] (discussed later).

Cepstrum

Consider a (not necessarily causal) signal $x(n)$ whose z-transform $X(z)$ exists and has a region of convergence that includes the unit circle. Suppose $C(z) = \log X(z)$ has a convergent power series expansion in which, again, the region of convergence includes the unit circle. The cepstrum is defined as the inverse z-transform of $C(z)$ in that [23]

$$C(z) = \sum_n c(n) z^{-n}. \tag{12}$$

Note that $c(n)$ is also not necessarily causal. Let us continue by assuming that $X(z)$ is a rational function of z that is completely described by its poles, zeros, and gain. Then, the cepstrum $C(z)$ will have the following properties [23]:

1. The sample $c(0)$ is the natural logarithm of the gain.
2. The poles and zeros of $X(z)$ inside the unit circle contribute only to the casual part of $c(n)$ starting at $n = 1$.
3. The poles and zeros of $X(z)$ outside the unit circle contribute only to the anticausal part of $c(n)$.
4. The cepstrum is causal if and only if $X(z)$ is minimum phase.
5. The cepstrum is anticausal if and only if $X(z)$ is maximum phase.
6. The cepstrum $c(n)$ decays as fast as $1/|n|$ as n approaches ∞ and $-\infty$.
7. The cepstrum has infinite duration whether $x(n)$ is of finite or infinite duration.
8. If $x(n)$ is real, $c(n)$ is real.

As a special case of the more general $X(z)$, consider the *minimum phase* all-pole LP filter $H(z)$ obtained by the autocorrelation method. Given that all the poles $z = z_i$ are inside

the unit circle and the gain is 1, the causal LP cepstrum $c_{lp}(n)$ of $H(z)$ is given by [17, 23, 24].

$$c_{lp}(n) = \begin{cases} \frac{1}{n}\sum_{i=1}^{p} z_i^n & n > 0 \\ 0 & n \le 0 \end{cases} . \tag{13}$$

A recursive relation between the LP cepstrum and the predictor coefficients is given as [17]

$$c_{lp}(n) = a_n + \sum_{i=1}^{n-1} (\frac{i}{n}) c_{lp}(i) a_{n-i} . \tag{14}$$

The use of this recursion allows for an efficient computation of $c_{lp}(n)$ and avoids polynomial factorization. Since $c_{lp}(n)$ is of infinite duration, the feature vector of dimension p consists of the components $c_{lp}(1)$ to $c_{lp}(p)$, which are the most significant due to the decay of the sequence with increasing n. Even with this truncation, the mean-square difference between two LP cepstral vectors is approximately equal to the mean-square difference between the log spectra of the corresponding all-pole LP filters [17]. Hence, this provides a good measure of the difference in the spectral envelope of the speech frames that the cepstral vectors were derived from.

Cepstral Derivatives

The LP cepstrum represents the local spectral properties of a given frame of speech. However, it does not characterize the temporal or transitional information in a sequence of speech frames. For text-related applications such as speech recognition and text-dependent speaker recognition, improved performance has been found by introducing cepstral derivatives into the feature space because the cepstral derivatives capture the transitional information in the speech. The first derivative of the cepstrum (also known as the delta cepstrum) is defined as [17]

$$\frac{\partial c_{lp}(n,t)}{\partial t} = \Delta c_{lp}(n,t) \approx \mu \sum_{k=-K}^{K} k c_{lp}(n,t+k) , \tag{15}$$

where $c_{lp}(n,t)$ denotes the n^{th} LP cepstral coefficients at time t, μ is an appropriate normalization constant, and $(2K + 1)$ is the number of frames over which the computation is performed. The LP cepstrum and the delta cepstrum together have been used to improve speaker recognition performance [25].

Cepstral Weighting

The basic idea behind cepstral weighting is to account for the sensitivity of the low-order cepstral coefficients to overall spectral slope and the sensitivity of the high-order cepstral coefficients to noise [17]. Weighting is accomplished by multiplying $c_{lp}(n)$ by a window $w(n)$ and using the weighted cepstrum as the feature vector. This weighting operation is also known as liftering. The first consequence of liftering is in extracting a finite dimensional feature vector from an infinite duration $c_{lp}(n)$. Also, careful choices of $w(n)$ enhance robustness. There are several schemes of weighting that differ in the type of cepstral window $w(n)$ that is used. The simplest one is the rectangular window as given by

$$w(n) = \begin{cases} 1 & n = 1,2,\ldots,L \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

where L is the size of the window. The first L samples, which are the most significant due to the decaying property, are kept. Other forms of $w(n)$ include *quefrency liftering* (or linear weighting) where

$$w(n) = \begin{cases} n & n = 1,2,\ldots,L \\ 0 & \text{otherwise} \end{cases} , \tag{17}$$

and *bandpass liftering* (BPL) [17, 26] where

$$w(n) = \begin{cases} 1 + \frac{L}{2}\sin(\frac{n\pi}{L}) & n = 1,2,\ldots,L \\ 0 & \text{otherwise} \end{cases} . \tag{18}$$

The quefrency liftering weights each individual cepstral component by its index n, thereby downplaying the lower-order components. The BPL weights a cepstral sequence by a raised sinusoidal function so that the lower- and higher-order components are de-emphasized. Note that the weighting schemes described are fixed in the sense that the weights are only a function of the cepstral index and have no explicit bearing on the instantaneous variations in the cepstrum that are introduced by different environmental conditions (like noise and channel effects).

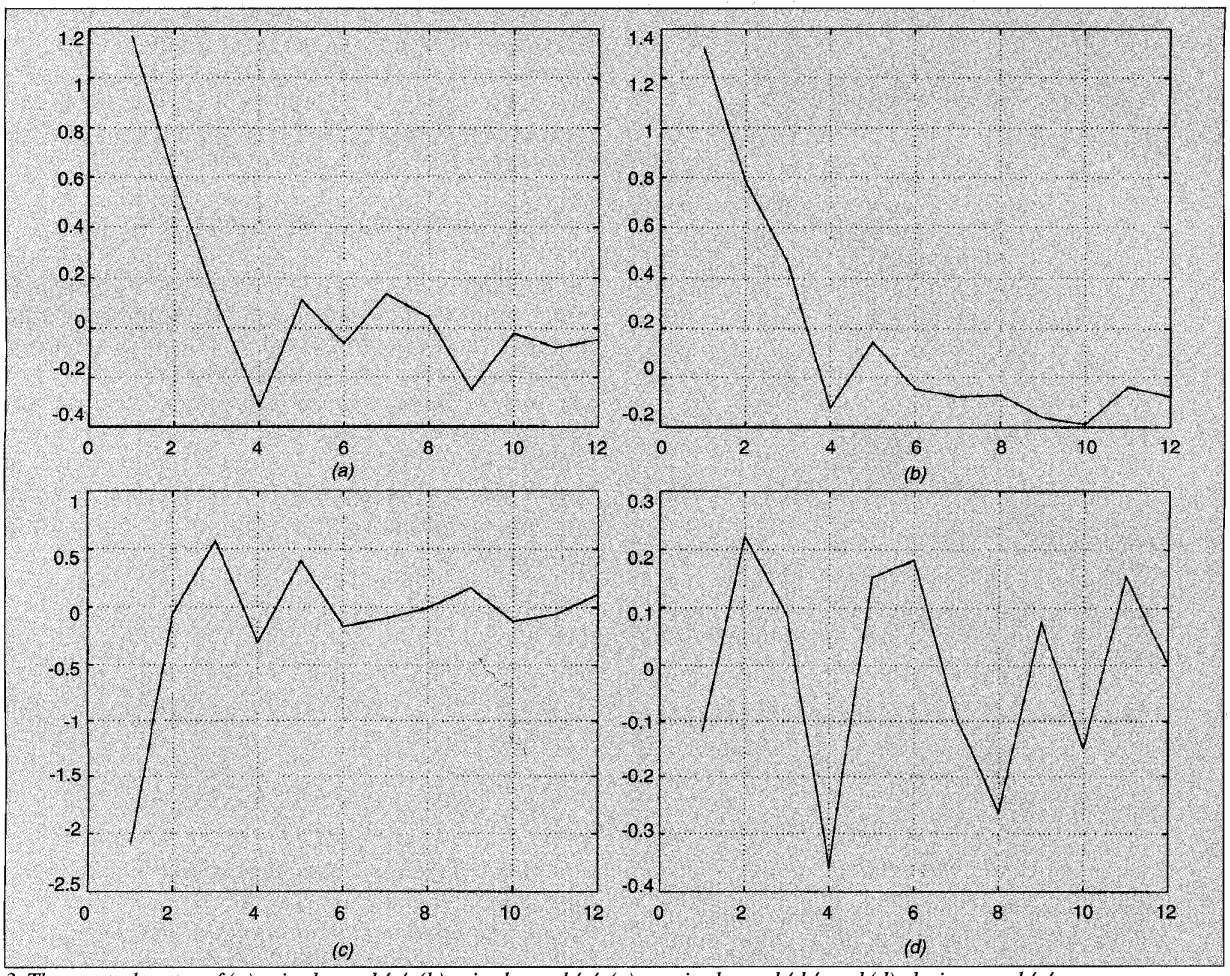
Cepstral Mean Subtraction (CMS)

A speech signal transmitted over a telephone network often encounters a linear distortion due to the filtering effect of the channel. This is simply expressed as $T(z) = S(z)G(z)$ where $S(z)$ corresponds to the original clean speech, $G(z)$ corresponds to the telephone channel, and $T(z)$ corresponds to the filtered speech. In the log domain,

$$\log T(z) = \log S(z) + \log G(z) . \tag{19}$$

Assuming that the speech and channel spectra are well approximated by the all-pole LP model, it is observed that a channel influence on the speech leads to an additive component on the LP cepstrum of the clean speech $S(z)$. By further assuming that the mean of the LP cepstrum of the clean speech is zero, the estimate of the channel cepstrum is merely the mean of the LP cepstrum of the filtered speech $T(z)$. To compensate for the channel effect, the channel estimate is removed by way of cepstral mean subtraction (CMS) [4, 21, 27]. The feature vector is

$$c_{cms}(n) = c_{lp}(n) - E[c_{lp}(n)] \tag{20}$$



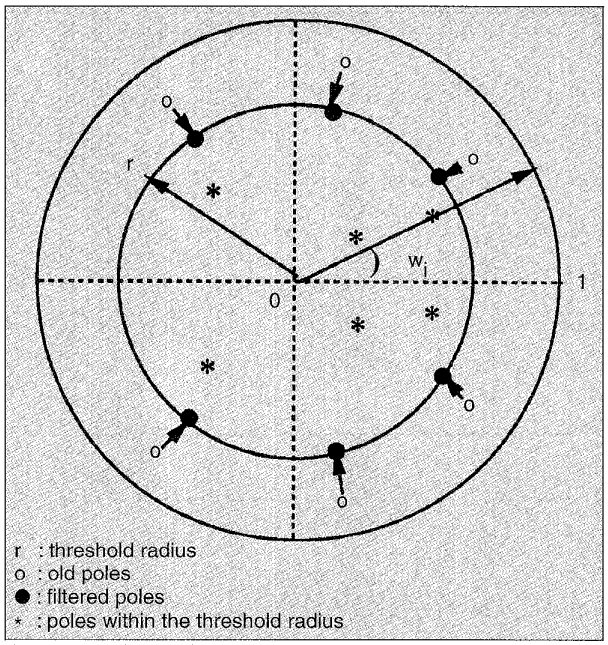
3. The cepstral vector of (a) voiced sound /a/; (b) voiced sound /u/; (c) unvoiced sound /sh/; and (d) plosive sound /p/.

where the expectation is taken over a number of frames of channel-corrupted speech. Note that rectangular weighting is implicitly assumed.

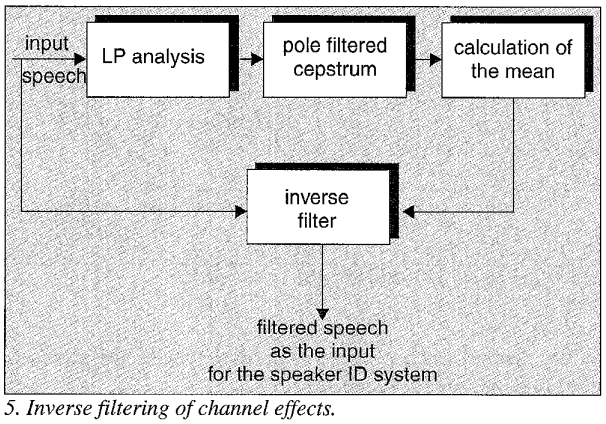
It has been shown that mean subtraction significantly improves the performance of a system in which training is done on one channel condition while testing is done on another channel condition. However, considerable loss of recognition accuracy is experienced when CMS is used for speaker recognition in which training and testing are done on the same channel. This is due to the implicit assumption of CMS that, in order to represent the channel cepstrum by the long-term cepstral mean of the channel-corrupted speech, the long-term cepstral mean of the clean speech has to be zero. Also, the assumption is only true when the speech segment is phonetically balanced in that the speech segment includes approximately the same amount of voiced, unvoiced, and plosive sounds. This is because the trajectories of cepstral coefficients for different sounds deviate from each other by a significant amount, but behave similar to that of sounds in the same category. This phenomenon can be observed from plots in Fig. 3. A more accurate estimate of the channel cepstrum could be obtained if the cepstral mean, solely due to the clean speech prior to convolution with the channel, could be maximally reduced.

Pole-filtered Cepstral Mean Subtraction (PFCMS)

The CMS concept is based on getting a channel estimate given by $E[c_{lp}(n)]$. The LP poles with narrow bandwidths that lie close to the unit circle usually represent the formants and are less sensitive to channel and noise effects. Hence, these poles do not contribute to the channel estimate as they contain much speech information. In contrast, the broad bandwidth poles model the spectral tilt, sub-glottal variation, and the channel effects. These poles offer a better estimate of the channel. A new concept, known as pole filtering, modifies the LP poles so as to broaden the bandwidth of the formant poles [28]. Bandwidth broadening is accomplished by moving the formant poles radially away from the unit circle. The pole frequency is left intact. Figure 4 illustrates the concept of pole filtering. The cepstrum formed from these filtered or modified poles (denoted as $c_{mlp}(n)$) has less speech information and more channel information than $c_{lp}(n)$, due to the de-emphasis of the formant poles. The channel estimate is given as $E[c_{mlp}(n)]$ and the feature vector is



4. Concept of pole filtering.



5. Inverse filtering of channel effects.

$$c_{pfcms}(n) = c_{lp}(n) - E[c_{mlp}(n)] \tag{21}$$

Note that rectangular weighting is implicitly assumed.

The technique of forming the feature vector is known as pole-filtered cepstral mean subtraction (PFCMS) and the details are given below.

- Select a threshold radius r_{th}
- For each frame of speech:
  —Calculate LP poles z_i for $i = 1$ to p
  —For each pole z_i:
  * If $|z_i| > r_{th}$, modify z_i such that its magnitude is r_{th} and its angle is unaltered.
  —Calculate $c_{mlp}(n)$ based on the modified or filtered poles
- Find the channel estimate $E[c_{mlp}(n)]$ over all speech frames
- Find $c_{pfcms}(n)$

It has been shown that PFCMS outperforms CMS in speaker-identification experiments [28].

Note that since $E[c_{mlp}(n)]$ is a good channel estimate, it can be converted to an all-pole filter representing the channel and, hence, to an inverse finite impulse response (FIR) filter. When this FIR filter is applied to the speech, the speech is enhanced as the channel effects are alleviated. This enhancement of the speech prior to feature extraction also improves speaker-identification performance [29]. Figure 5 shows a block diagram.

Adaptive Component Weighted Cepstrum

The techniques of CMS and PFCMS are examples of interframe processing techniques in that information over more than one frame is used to develop the feature vector. In this and the next subsection, two intraframe (information taken only over the frame of concern) approaches are described. The first is known as adaptive component weighting (ACW) [30]. Consider the LP transfer function $H(z)$ as parameterized by the residues r_k and the poles z_k, which are in turn further described by σ_k and ω_k. First we can write the impulse response $h(n)$ as

$$\begin{pmatrix} h(1) \\ h(2) \\ \vdots \\ h(p) \end{pmatrix} = \begin{pmatrix} \sigma_1 e^{j\omega_1} & \sigma_2 e^{j\omega_2} & \cdots & \sigma_p e^{j\omega_p} \\ \sigma_1^2 e^{j2\omega_1} & \sigma_2^2 e^{j2\omega_2} & \cdots & \sigma_p^2 e^{j2\omega_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1^p e^{jp\omega_1} & \sigma_2^p e^{jp\omega_2} & \cdots & \sigma_p^p e^{jp\omega_p} \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{pmatrix} \tag{22}$$

Formants of the speech signal are weighted by the residues r_i individually. It was observed in [30] that the residues show considerable variation when speech is passed through a channel. This is equivalent to saying that the amplitudes r_k of the individual eigenmodes in the modulation model representation (see Eq. 6) are most perturbed by the channel among the three parameters r_k, σ_k, and ω_k. The ACW cepstrum removes the variations caused by channel variability by normalizing the residues so that the narrow-band components corresponding to formants are emphasized and the broad-band components are suppressed. Hence, we get a pole-zero system function of the form

$$H_{acw}(z) = \frac{N(z)}{A(z)} = \sum_{k=1}^{p} \frac{1}{1 - z_k z^{-1}} \tag{23}$$

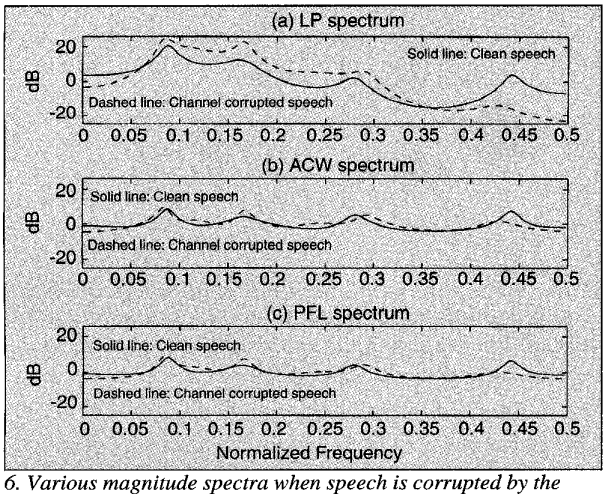
where

$$N(z) = \sum_{k=1}^{p} \prod_{i=1 \neq k}^{p} (1 - z_i z^{-1}) \tag{24}$$

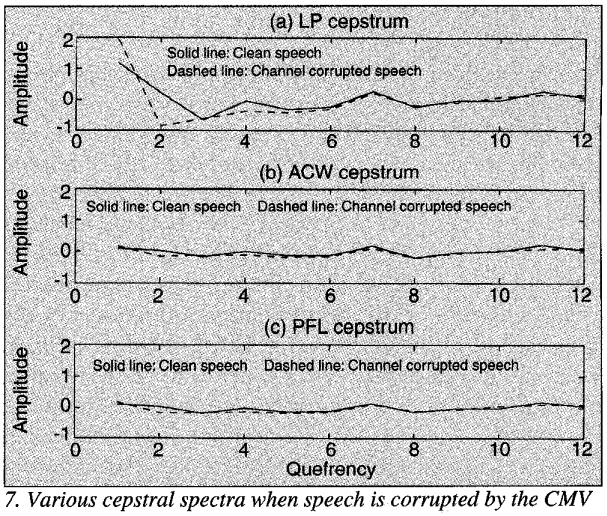
which can be further written as

$$N(z) = p(1 - \sum_{k=1}^{p-1} b_k z^{-1}) . \tag{25}$$

It can be shown that $N(z)$ is minimum phase [31]. Therefore, the ACW cepstrum is causal and given by $c_{acw}(0) = \log p$ and



6. Various magnitude spectra when speech is corrupted by the CMV channel (clean spech = solid line, channel-corrupted speech = dashed line). (a) Magnitude response of 1/A(z), (b) Magnitude response of $H_{acw}(z)$, (c) Magnitude response of $H_{pfl}(z)$ (α=1, β=0.9).



7. Various cepstral spectra when speech is corrupted by the CMV channel (clean spech = solid line, channel-corrupted speech = dashed line). (a) LP cepstrum $c_{lp}(n)$, (b) ACW cepstrum $c_{acw}(n)$, (c) PFL cepstrum $c_{pfl}(n)$ (α=1, β=0.9).

$$c_{acw}(n) = c_{lp}(n) - c_{nn}(n) \tag{26}$$

for $n > 0$ where $c_{nn}(n)$ can be found by a recursion involving the coefficients b_k (same type of recursion as in Eq. 14). Moreover, since b_k is simply expressed as [31]

$$b_k = \frac{p-k}{p} a_k \tag{27}$$

for $1 \le k \le p-1$, the computation of the ACW cepstrum is very simple. The subtractive component $c_{nn}(n)$ serves as an estimate of the channel. Unlike CMS and PFCMS, this component is adaptive on a frame-by-frame basis. In practice, rectangular weighting is applied so that the feature vector consists of the components of $c_{acw}(n)$ for $n = 1$ to p. The performance of speaker-identification systems is definitely better if the ACW cepstrum is used as opposed to the LP cepstrum [30, 32].

Postfilter Cepstrum

The concept of a postfilter was introduced in [33] to enhance noisy speech. The philosophy in developing a postfilter relies on the fact that more noise can be perceptually tolerated in the formant regions (spectral peaks) than in the spectral valleys. The postfilter is obtained from $A(z)$ and its transfer function is given by

$$H_{pfl}(z) = \frac{A(z/\beta)}{A(z/\alpha)} \qquad 0 < \beta < \alpha \le 1 \tag{28}$$

If $A(z)$ is minimum phase, $H_{pfl}(z)$ is guaranteed to be minimum phase. Therefore, the postfilter cepstrum (referred to as the PFL cepstrum) [32] is causal and given by $c_{pfl}(0) = 0$ and

$$c_{pfl}(n) = c_{lp}(n)[\alpha^n - \beta^n] \tag{29}$$

for $n > 0$. The PFL cepstrum is merely a weighting or liftering of the LP cepstrum and is very robust to channel and noise effects [32]. Like other ways of liftering the LP cepstrum, namely, bandpass liftering [26], quefrency liftering [34], and inverse variance liftering [35], the lower-indexed cepstral coefficients are de-emphasized. If $\alpha = 1$, $c_{pfl}(n) = c_{lp}(n) - \beta^n c_{lp}(n)$. There is a subtractive component that serves as a channel estimate and that is adaptive on a frame-by-frame basis. In practice, rectangular weighting is applied so that the feature vector consists of the components of $c_{pfl}(n)$ for $n = 1$ to p.

Figure 6 shows the magnitude responses of $H(z)$, $H_{acw}(z)$, and $H_{pfl}(z)$ for a frame of clean speech and for the same frame of speech corrupted by the continental mid voice (CMV) channel [36], which is a typical bandpass channel encountered on the telephone network. Due to the channel-filtering effect, there is a glaring mismatch in the spectra of $1/A(z)$ as revealed in Fig. 6(a). This mismatch is alleviated considerably by introducing $H_{acw}(z)$ and $H_{pfl}(z)$. As can be seen in Fig. 6(b) and (c), the mismatch in the magnitude spectrum for the ACW and PFL methods is reduced over that of $1/A(z)$. The ACW spectrum and the PFL spectrum are similar in that they both emphasize the formant peaks that are more crucial for speaker identification. Also, there is no apparent spectral tilt. The PFL spectrum is sensitive to changes in α and β. A decrease of α causes formant bandwidth broadening while a change in β affects the spectral tilt. As β decreases, the spectral tilt becomes more apparent.

Figure 7 shows the corresponding cepstral coefficients of $c_{lp}(n)$, $c_{acw}(n)$, and $c_{pfl}(n)$ for a frame of clean speech and for the same frame of speech corrupted by the CMV channel [36]. There is much less mismatch in $c_{acw}(n)$ and $c_{pfl}(n)$ as compared with $c_{lp}(n)$.

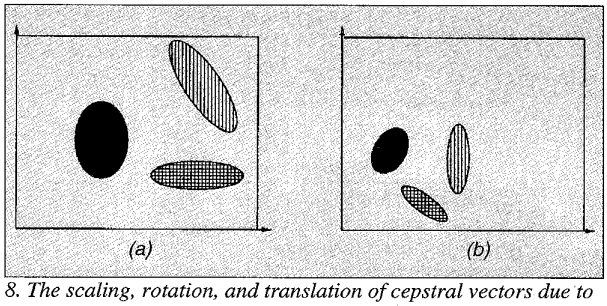
Mel-warped Cepstrum

The mel-warped cepstrum differs from the LP cepstrum in that the mel-warped cepstrum is calculated using a filter-bank approach in which the set of filters are of equal bandwidth with respect to the mel-scale of frequencies. This is because human perception of the frequency content of sounds does not follow a linear scale. Instead, they are approximately linear with logarithmic frequency beyond about 1000 Hz. In addition, the critical band is a constant with the logarithmic frequency below 1000 Hz and linear with respect to the logarithmic frequency beyond 1000 Hz. The critical band refers to the bandwidth within which subjective responses such as loudness remain constant until the noise bandwidth exceeds the width of the critical band. The mel-scale is defined in such a way that the 1000 Hz in the linear frequency domain is 1000 mels, and the other values are obtained by adjusting the frequency of a tone such that the human perceived frequency is half or twice the perceived frequency of a reference point with a known mel frequency. The mel-scale spectrum is simulated using a filter bank spaced uniformly on a mel-scale, where the output energy from each filter band approximates the modified spectrum. If we denote the output energy of the k^{th} filter by $\tilde{S}_k$, the mel-warped cepstrum $c_{mel}(n)$ is obtained by taking the shifted discrete cosine transform (DCT) of the mel-scale spectrum as

$$c_{mel}(n) = \sum_{k=1}^{K} \log(\tilde{S}_k) \cos(n(k-0.5)\frac{\pi}{K}) \tag{30}$$

Other Robust Cepstral Techniques

Methods such as noise-cancelling microphones, preprocessor noise suppression, and internal modification of the processing algorithms to explicitly compensate for signal contamination have been proposed to reduce the background noise acoustically added to speech. Noise suppression [30] using spectral normalization enhances the auditory quality of speech. However, it rarely improves the performance of a recognition system. The autoregressive moving average model (ARMA) for getting a robust estimate of the linear predictor coefficients suggests adding zeros to the vocal-tract model to take into account the effect due to uncorrelated additive noise. However, it is not feasible because the computation is intensive and the solution is not guaranteed to converge to the global minimum of the highly nonlinear cost function. The relative spectral (RASTA) technique [37] takes advantage of the fact that the rate of change of nonlinguistic components in speech often lies outside the typical rate of change of the vocal-tract shape. Therefore, it suppresses the spectral components that change more slowly or quickly than the typical rate of change of speech. The RASTA approach can be combined with the PLP method to get the LP transfer function $H(z)$ [37]. The critical-band speech spectrum is found just like in the PLP method. The time trajectories of the spectral components are filtered to suppress the nonlinguistic components in the spectrum. This filtered spectrum is approximated by an autoregressive model. The delta cepstrum, which also reflects spectral changes, has been shown to be a special case of RASTA processing [37]. Also, unlike cepstral mean subtraction, which removes the dc component of the short-term log spectrum, RASTA processing influences the speech spectrum in a more complex manner and emphasizes spectral transitions. The use of RASTA processing has been shown to improve speech-recognition performance under mismatched environments [37]. Alternatively, RASTA processing can be directly applied to cepstral coefficients found from the LP transfer function obtained by the conventional autocorrelation method. A bandpass filter of the form [38]



8. The scaling, rotation, and translation of cepstral vectors due to noise and channel effects: (a) the spatial distribution of the clean cepstral vectors; (b) the spatial distribution of the cepstral vectors of noise and/or contaminated speech.

$$B(z) = \frac{a_0 + a_1 z^{-1} + a_3 z^{-3} + a_4 z^{-4}}{(1 - b_1 z^{-1}) z^{-4}} \tag{31}$$

is applied to the cepstral coefficients. This bandpass operation, combined with BPL filtering, has been shown to improve speaker-recognition performance under mismatched conditions [38]. Other techniques that attempt to internally modify the algorithm to adapt the model set-up from a clean environment to that of a noisy environment can be found in [6, 39, 40].

Distortion Correction Using Affine Transformation

When the speech signal is corrupted by a channel and noise from the environment, the cepstral vectors are found to be rescaled, rotated, and translated. This can be seen in Fig. 8.

In practice, CMS, cepstral liftering, and other techniques previously mentioned have been found to offer enhanced robustness of speaker-recognition systems. Specifically, if we consider doing CMS to normalize the channel, and performing cepstral liftering to reduce the noise effect, then the corrected cepstral coefficients $\hat{c}_i$ can be represented as

$$\hat{c}_i = w_i c_i' - \overline{c_i'} \tag{32}$$

where c_i' are the cepstral coefficients of the clean speech. The generalization of the relationship can be written as an affine transform given by

$$\mathbf{c}' = \mathbf{A}\mathbf{c} + \mathbf{b} \tag{33}$$

where $\mathbf{c}'$ is the cepstrum of the degraded speech and $\mathbf{c}$ is the cepstrum of the original clean speech. This becomes a similarity transform when the matrix A is diagonal and the vector b is zero.

The concept of using an affine transform to correct the distortions of the cepstral coefficients caused by the channel and noise interferences has been proposed in [41]. Its underlying idea is that the predictor coefficients are affinely transformed when the speech signal is contaminated by environmental perturbations, resulting in an affine transform of the cepstral coefficients. The transformations are dependent on the spectral properties of the sounds. A degraded, spectrally similar set of cepstral vectors would undergo the same transformation. In the following, we will briefly review the analysis and examine the effects of additive noise and the linear channel individually. The autocorrelation-based solution of the predictor coefficients as given by $\mathbf{a} = \mathbf{R}_s^{-1}\,\mathbf{r}_s$ will be used throughout the following analysis.

Additive Noise

The random noise arising from the background and the fluctuation of the transmission channel is generally assumed to be additive white noise (AWN). The noisy observation of the original speech signal is then given by

$$s'(n) = s(n) + q(n) \tag{34}$$

where the noise $q(n)$ is such that

$$E[q(n)] = 0 \quad \text{and} \quad E[q^2(n)] = \sigma^2 \tag{35}$$

It can be shown that the predictor coefficients of the noise-corrupted speech as given by Eq. 11 are [41]

$$\mathbf{a}' = \mathbf{R}_{s'}^{-1}\mathbf{r}_{s'} = (\mathbf{R}_s + \sigma^2\mathbf{I})^{-1}\mathbf{r}_s = (\mathbf{R}_s + \sigma^2\mathbf{I})^{-1}\mathbf{R}_s\mathbf{a}. \tag{36}$$

It can be seen from Eq. 11 that the addition of AWN noise to the speech is equivalent to taking a linear transformation of the predictor coefficients. The linear transformation depends on the autocorrelation of the speech and thus, in a spectrum-based model, all the spectrally similar predictors will be mapped by a similar linear transform.

The singular value decomposition (SVD) of the transformation in Eq. 36 will help gain some insight into the interaction of noise and the predictor coefficients. Assume that the Toeplitz autocorrelation matrix of the original speech signal $\mathbf{R}_s$ is decomposed as

$$\mathbf{R}_s = \mathbf{U}\Lambda\mathbf{U}^T \tag{37}$$

where $\mathbf{U}$ is a unitary matrix and Λ is a diagonal matrix whose diagonal elements are the eigenvalues of the matrix $\mathbf{R}_s$. It can be shown that Eq. 36 can be rewritten as [41]

$$\mathbf{a}' = \left[\mathbf{U}(\Lambda + \sigma^2\mathbf{I})\mathbf{U}^T\right]^{-1}(\mathbf{U}\Lambda\mathbf{U}^T)\mathbf{a} = \mathbf{U}[(\Lambda + \sigma^2\mathbf{I})^{-1}\Lambda]\mathbf{U}^T\mathbf{a} \tag{38}$$

$$= \mathbf{U}\begin{pmatrix} \frac{\lambda_1^2}{\lambda_1^2+\sigma^2} & & & \\ & \frac{\lambda_2^2}{\lambda_2^2+\sigma^2} & & \\ & & \ddots & \\ & & & \frac{\lambda_n^2}{\lambda_n^2+\sigma^2} \end{pmatrix}\mathbf{U}^T\mathbf{a}$$

From the above equation we can see that the norm of the predictor coefficients is reduced when the speech is perturbed by white noise. When speech is corrupted by additive white noise, the predictor coefficient vector maintains its general orientation but undergoes a shrinkage that brings it closer to the origin.

Linear Channel

When a sample sequence is passed through a convolutional channel of impulse response $s'(n)$, the filtered signal obtained at the output of the channel is

$$s'(n) = p(n) \otimes s(n) \tag{39}$$

If the power spectra of the signals $s(n)$ and $s'(n)$ are denoted $S_s(\omega)$ and $S_{s'}(\omega)$, respectively, then

$$S_{s'}(\omega) = |P(\omega)|^2 S_s(\omega). \tag{40}$$

Therefore in the time domain,

$$r_{s'}(k) = [p(n) \otimes p(-n)] \otimes r_s(k) = r_p(k) \otimes r_s(k) \tag{41}$$

where $r_s(k)$ and $r_{s'}(k)$ are the autocorrelation of the input and output signals, respectively, and $\otimes$ is the convolution operator. By using these relations, it can be shown that the predictor coefficients of the output signal $s'(n)$ is given by [41]

$$\mathbf{a}_{s'} = \mathbf{A}\mathbf{a}. \tag{42}$$

Therefore, the predictor coefficients of a speech signal filtered by a convolutional channel can be obtained by taking a linear transformation of the predictor coefficients of the input speech signal. No translation of the predictor coefficients results. Note that the transformation in Eq. 42 is sound dependent, as the estimates of the autocorrelation matrices assume stationarity.

Co-channel Interference

The co-channel interference due to a second speaker can also be interpreted as an affine transformation. In the case of interference due to another speaker talking on the same channel, the observed signal s_T is

$$s_T = s_1 + s_2 \tag{43}$$

Accordingly,

$$\mathbf{R}_{s_T} = \mathbf{R}_{s_1} + 2\mathbf{R}_{s_1 s_2} + \mathbf{R}_{s_2} \tag{44}$$

and

$$\mathbf{r}_{s_T} = \mathbf{r}_{s_1} + 2\mathbf{r}_{s_1 s_2} + \mathbf{r}_{s_2} \tag{45}$$

Therefore, the linear prediction coefficients for signal s_T is

$$\begin{aligned} \mathbf{a}_T &= \mathbf{R}_{s_T}^{-1}\mathbf{r}_{s_T} \\ &= (\mathbf{R}_{s_1} + 2\mathbf{R}_{s_1 s_2} + \mathbf{R}_{s_2})^{-1}\mathbf{R}_{s_1}\mathbf{R}_{s_1}^{-1}(\mathbf{r}_{s_1} + 2\mathbf{r}_{s_1 s_2} + \mathbf{r}_{s_2}) \\ &= \mathbf{A}\mathbf{a}_1 + \mathbf{b} \end{aligned} \tag{46}$$

where

$$\mathbf{A} = (\mathbf{R}_{s_1} + 2\mathbf{R}_{s_1 s_2} + \mathbf{R}_{s_2})^{-1}\mathbf{R}_{s_1} \tag{47}$$

and

$$\mathbf{b} = (\mathbf{R}_{s_1} + 2\mathbf{R}_{s_1 s_2} + \mathbf{R}_{s_2})^{-1}\mathbf{R}_{s_1}(2\mathbf{r}_{s_1 s_2} + \mathbf{r}_{s_2}) \tag{48}$$

Again, the co-channel interference carries out an affine transformation on the predictor coefficients.

The above derivations show that mismatches due to additive noise, linear channel, and co-channel interference are individually either a linear or an affine transformation on the linear predictor coefficients. In general, due to the transitivity of the affine transform, a sequence of distortions resulting from the noise and channel interferences are also equivalent to an affine transformation of the form

$$\mathbf{a}' = \mathbf{A}\mathbf{a} + \mathbf{b} \tag{49}$$

Affine Transform of Cepstrum

Empirically, cepstral features have been found to be the most robust to various sources of degradation. In this section, we will see that for the LP cepstrum, the effect of channel and noise can also be modeled as an affine transformation.

The LP cepstrum is by definition

$$c_{lp}(n) = \mathcal{Z}^{-1}[\log(H(z))] = \mathcal{Z}^{-1}[\log(\frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}})] \ . \tag{50}$$

The first-order partial derivative of $c_{lp}(n)$ with respect to a_i is given by

$$\begin{aligned} \frac{\partial c_{lp}(n)}{\partial a_i} &= \frac{\partial \mathcal{Z}^{-1}[\log(\frac{1}{1-\sum_{k=1}^{p} a_i z^{-1}})}{\partial a_i} \\ &= \mathcal{Z}^{-1}[\frac{\partial \log(\frac{1}{1-\sum_{k=1}^{p} a_i z^{-k}})}{\partial a_i} \\ &= h(n-i) \end{aligned} \tag{51}$$

where $h(n)$ is the causal and stable impulse response associated with the LP transfer function $H(z)$. Therefore, if $\mathbf{c}$ is the vector of the first p LP cepstral coefficients, then

$$dc = \mathbf{H}d\mathbf{a} \tag{52}$$

where

$$\mathbf{H} = \begin{pmatrix} h(0) & 0 & \cdots & 0 \\ h(1) & h(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h(p-1) & h(p-2) & \cdots & h(0) \end{pmatrix} \tag{53}$$

Note that the impulse response matrix $\mathbf{H}$ would be the same for a group of spectrally similar cepstral vectors. The relationship between a differential LP cepstral vector and a differential predictor coefficient vector is given by Eq. 52.

Suppose the speech signal undergoes distortion due to a channel and/or noise. A new LP transfer function $H'(z)$ results with the corresponding predictor coefficient vector given by $\mathbf{a}'$ and the LP cepstrum vector given by $\mathbf{c}'$. For a set of spectrally similar cepstral vectors, the same transformation can be written as

$$dc' = \mathbf{H}'d\mathbf{a}' \tag{54}$$

Since the predictor coefficients are affinely transformed in the presence of interference in that $\mathbf{a}' = \mathbf{A}\mathbf{a} + \mathbf{b}$, differentiating both sides yields

$$d\mathbf{a}' = \mathbf{A}d\mathbf{a} \tag{55}$$

Then, we have

$$\frac{d\mathbf{c}'}{d\mathbf{c}} = (\frac{d\mathbf{c}'}{d\mathbf{a}'})(\frac{d\mathbf{a}'}{d\mathbf{a}})(\frac{d\mathbf{a}}{d\mathbf{c}}) = \mathbf{H}'\mathbf{A}\mathbf{H}^{-1} \tag{56}$$

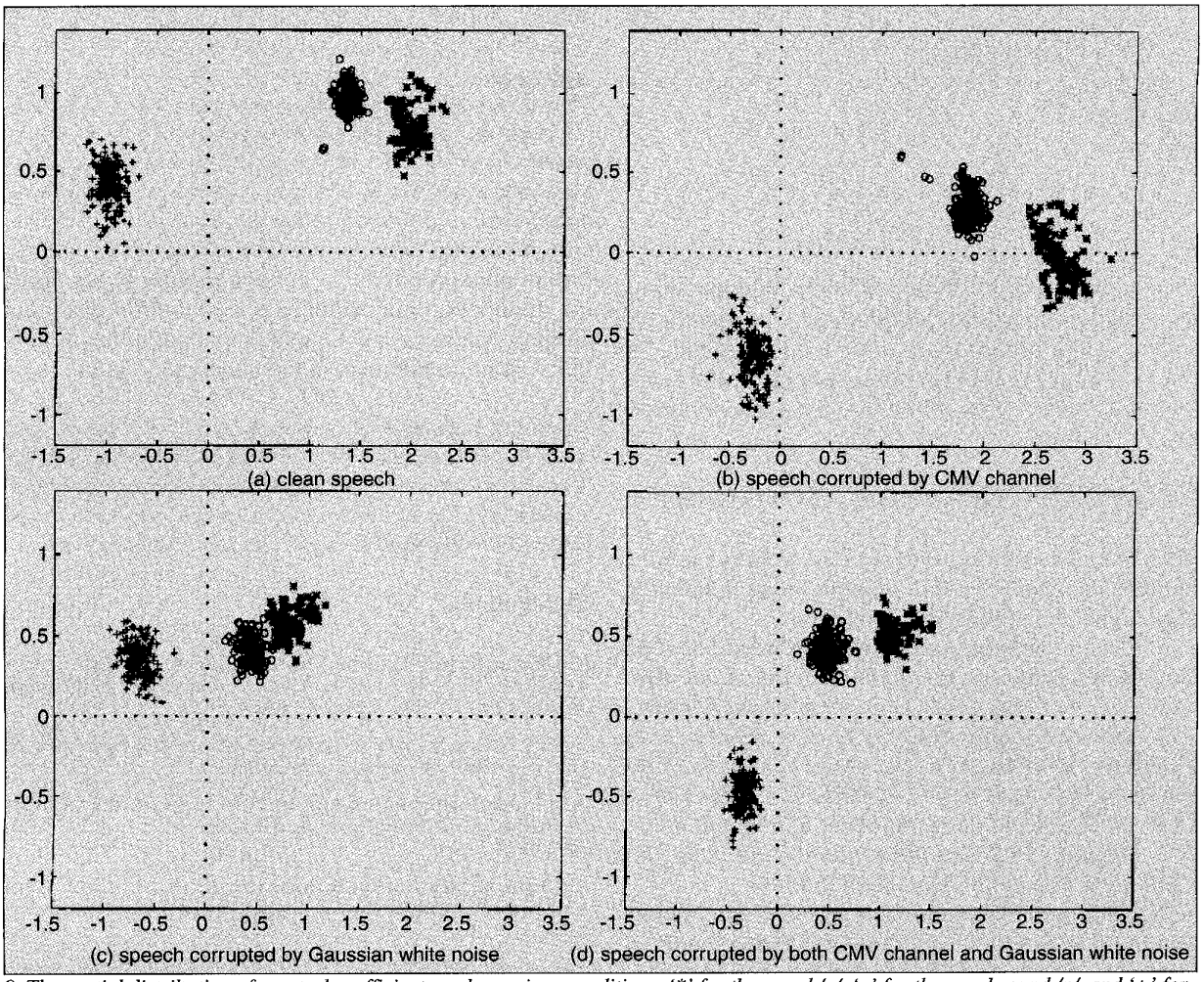
By performing integration, we see that the LP cepstrum corresponding to the distorted speech is given by

$$\mathbf{c}' = \mathbf{H}'\mathbf{A}\mathbf{H}^{-1}\mathbf{c} + \mathbf{b} \tag{57}$$

which is an affine transformation. At this point, we observe that the LP cepstral coefficients are affinely mapped by mismatches in the noise and channel conditions of the acoustical environments. The parameters of the affine mapping are spectrally dependent.

Computation of Affine Transform Parameters

Assume that the correspondence between the cepstral vectors for the training condition $\mathbf{c}_i = [c_{i1}c_{i2}\cdots c_{ip}]^T$ and the cepstral



9. The spatial distribution of cepstral coefficients under various conditions, '' for the vowel /a/, 'o' for the nasal sound /n/, and '+' for the sound /sh/. (a) Cepstrum of the clean speech; (b) Cepstrum of signals filtered by continental U.S. voice mid channel (CMV); (c) Cepstrum of signals with 15 dB SNR. The noise type is additive white Gaussian (AWG); (d) Cepstrum of speech corrupted by both CMV channel and AWG noise of 15 dB SNR.*

vectors for the testing condition $\mathbf{c}_i = [c'_{i1}c'_{i2}\cdots c'_{ip}]^T$ are known for $i = 1$ to N where N is the number of cepstral vectors. The affine transform relating the vectors $\mathbf{c}_i$ and $\mathbf{c}'_i$ is given by

$$\mathbf{c}'_i = \mathbf{A}\mathbf{c}_i + \mathbf{b} \tag{58}$$

which is expanded as

$$\begin{pmatrix} c'_{i1} \\ \vdots \\ c'_{ip} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix} \begin{pmatrix} c_{i1} \\ \vdots \\ c_{ip} \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} \tag{59}$$

for $i = 1$ to N. Each individual row of the matrix $\mathbf{A}$ (the elements a_{jk} for $k = 1$ to p) and the corresponding element of the vector $\mathbf{b}$ (element b_j) are determined separately. To determine the jth row of $\mathbf{A}$ and b_j, we gather the jth component of each of the cepstral vectors of the testing condition and get the following system of equations

$$\begin{pmatrix} c'_{1j} \\ \vdots \\ c'_{Nj} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1p} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ c_{N1} & \cdots & c_{Np} & 1 \end{pmatrix} \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \\ b_j \end{pmatrix} \tag{60}$$

$$= \begin{pmatrix} \mathbf{c}_1 & 1 \\ \vdots & \vdots \\ \mathbf{c}_N & 1 \end{pmatrix} \begin{pmatrix} a_{j1} \\ \vdots \\ a_{jp} \\ b_j \end{pmatrix}$$

for $j = 1$ to p. This is almost always an overdetermined system of equations and, hence, a least-squares solution is obtained as [42]

$$\begin{pmatrix} a_{ji} \\ \vdots \\ a_{jp} \\ b_j \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} \mathbf{c}_i \mathbf{c}_i^T & \sum_{i=1}^{N} \mathbf{c}_i \\ \left(\sum_{i=1}^{N} \mathbf{c}_i\right)^T & N \end{pmatrix} \begin{pmatrix} c'_{1j} \\ \vdots \\ c'_{Nj} \end{pmatrix} \quad (61)$$

$$(\mathbf{c}_1 \cdots \mathbf{c}_N \;\; 1)$$

for $j = 1$ to p. With the affine transform as presented above, the vectors for the training condition can be mapped into the space occupied by the vectors of the testing condition. The reverse mapping is also possible by solving for the vectors for the training condition in terms of the vectors for the testing condition.

Geometric Interpretation of Affine Transform

The affine transform can take care of a wide variety of mismatched conditions and subsumes other robust techniques. Consider the following examples.

1. If the training and testing conditions are matched in that the corresponding cepstral vectors are identical, the affine transform parameters will be found to be $\mathbf{A} = \mathbf{I}$ and $\mathbf{b} = 0$.

2. Channel distortion will lead to $\mathbf{A}$ being close to the identity matrix and $\mathbf{b}$ representing a translation that is equivalent to the techniques of CMS or PFCMS. The methods of CMS and PFCMS are subsumed by the affine transform.

3. Liftering, which does offer enhanced robustness [17], is subsumed in that $\mathbf{A}$ is diagonal and $\mathbf{b} = 0$.

4. Additive noise causes the magnitude of the cepstral vectors to shrink without significantly changing their orientation [11]. This type of distortion can be taken care of by a diagonal $\mathbf{A}$ in which the diagonal elements are generally different and have a magnitude less than 1. Also, $\mathbf{b} = 0$. Note that this is also a special case of liftering.

5. Composite effects of channel and noise are also modeled as an affine transform with $\mathbf{A}$ primarily responsible for the noise distortion and $\mathbf{b}$ primarily responsible for the channel distortion. Fig. 9 shows the change of the spatial clustering of the cepstral coefficients due to interferences of the linear channel, white noise, and the composite effect of both the linear channel and white noise.

The use of the affine transform has been shown to dramatically improve the performance of text-dependent speaker recognition systems [41]. The composite effect of channel and noise has been studied.

Summary

This article has presented a review of some of the techniques used in robust speaker recognition with an emphasis on feature extraction and enhancement steps. Most of the features described are based on the LP model of speech. The classical autocorrelation method for finding the LP coefficients is not by itself very robust to a very wide variety of environmental conditions. However, a better model that is robust and computationally tractable has yet to be realized. The LP coefficients are converted into different types of cepstral features. In particular, the ACW cepstrum and the PFL cepstrum are robust to channel and noise effects. But, more effort is needed in finding features for achieving very high recognition performance (especially under severe channel conditions and very low signal-to-noise ratios). The affine transform is a very recent and promising technique for mapping the feature space from one region to another to correct for deviations caused by the corruption of the speech signal by channel and noise. With the affine transform, relatively better performance at low signal-to-noise ratios is achieved, and further research on the affine transform is encouraged.

Richard J. Mammone is a professor and Ravi P. Ramachandran is a research assistant professor at Rutgers University's Center for Computer Aids for Industrial Productivity in Piscataway, NJ. Xiaoyu Zhang is a former graduate student who currently works at SPEAKEZ as a senior speech scientist.

References

1. A.E. Rosenberg, F.K. Soong. "Recent Research in Automatic Speaker Recognition." In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*, 701-738, Marcel Dekker, 1991.

2. J.H.L. Hansen, R.J.Mammone, S.Young, editors. *IEEE Transactions on Speech and Audio Processing*, October, 1994.

3. S.F. Boll. "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, 27:113-120, April, 1979.

4. S. Furui. "Cepstral Analysis Techniques for Automatic Speaker Verification," *IEEE Trans. Acoust., Speech, Signal Processing*, 29:254-272, April, 1981.

5. D. Mansour and B.H. Juang. "The short-time modified coherence representation and noisy speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, 37:795-804, June, 1989.

6. L. Newneyer and M. Weintraub. "Probabilistic Optimum Filtering for Robust Speech Recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1:417-420, 1994.

7. D.A. Reynolds and R.C. Rose. "Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. Speech, Audio Processing*, 3:72-83, January, 1995.

8. J.A. Nolazco Flores, S.J. Young. "Continuous Speech Recognition in Noise Using Spectral Subtraction and HMM Adaptation," Proc. *IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1:409-412, 1994.

9. F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoust., Speech, Signal Processing, 23:67-72, Feb., 1975.

10. F.K. Soong, M.M. Sondhi. "A Frequency-weighted Itakura Spectral Distortion Measure and its Application to Speech Recognition in Noise," *IEEE Trans. Acoust., Speech, Signal Processing*, 36:41-48, Jan., 1988.

11. D. Mansour, B.H. Juang. "A Family of Distortion Measures Based Upon Projection Operation for Robust Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, 37:1659-1671, November, 1989.

12. J.D. Markel, A.H. Gray Jr. *Linear Prediction of Speech*, Springer-Verlag, Berlin Heidelberg New York, 1976.

13. L.R. Rabiner, R.W. Schafer. *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, NJ, 1978.

14. G. Fant, *Acoustic Theory of Speech Production*, Mouton and Co., Gravenhage, The Netherland, 1960.

15. J.R. Deller, J.G. Proakis, J.H. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan, New York NY, 1993.

16. K.T. Assaleh, R. J. Mammone, M.G. Rahim, J.L. Flanagan. "Speech Recognition Using The Modulation Model." *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2, 664-667, April, 1993.

17. L.R. Rabiner, B.H. Juang. *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993

18. R.P. Ramachandran, M.S. Zilovic, R.J. Mammone. "A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification," *IEEE Trans. Speech, Audio Processing*, 3:117-125, March, 1995.

19. H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoust. Society of Amer.*, 87:1738-1752, April, 1990.

20. P.C. Woodland, M.J.F. Gales, D. Pye, V. Valtchev. "The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task," *ARPA Speech Recognition Workshop*, February, 1996.

21. B.S. Atal. "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *J. Acoust. Soc. Am.*, 55:1304-1312, 1974.

22. C.-S. Liu, M.-T. Lin, W.-J. Wang, H.-C. Wang. "Study of Line Spectrum Pair Frequencies for Speaker Recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 277-280, 1990.

23. A.V. Oppenheim, R.W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

24. M.R. Schroeder. "Direct (Nonrecursive) Relations Between Cepstrum and Predictor Coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, 29:297-301, April, 1981.

25. F.K. Soong, A.E. Rosenberg. "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, 36:871-879, June, 1988.

26. B.H. Juang, L.R. Rabiner. J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, 35:947-954, July, 1987.

27. B.S. Atal. "Automatic Recognition of Speakers from their Voices," *Proc. IEEE*, 64:460-475, April, 1976.

28. D. Naik. "Pole-filtered Cepstral Mean Subtraction," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1:157-160, 1995

29. D. Naik, K.T. Assaleh, R.J. Mammone. "Robust Speaker Identification Using Pole Filtering," *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1994.

30. K.T. Assaleh and R.J. Mammone, New LP-derived features for speaker identification, *IEEE Trans. Speech, Audio Processing*, 2:630-638, October, 1994

31. M.S. Zilovic, R.P. Ramachandran, R.J. Mammone. "A Fast Algorithm for Finding the Adaptive Component Weighted Cepstrum for Speaker Recognition," *Submitted to IEEE Trans. Speech, Audio Processing*, March, 1995.

32. M.S. Zilovic, R.P. Ramachandran, R.J. Mammone. "Speaker Identification Based on the Use of Robust Cepstral Features Obtained from Pole-zero Transfer Functions," *Submitted to IEEE Trans. Speech, Audio Processing*, March, 1995.

33. V. Ramamoorthy, N.S. Jayant, R.V. Cox, M.M. Sondhi. "Enhancement of ADPCM Speech Coding with Backward Adaptive Algorithms for Postfiltering and Noise Feedback," *IEEE Jour. on Select. Areas in Commun*, 6:364-382, February, 1988.

34. K.K. Paliwal. "On the Performance of the Frequency-weighted Cepstral Coefficients in Vowel Recognition," *Speech Communication*, 1:151-154, May, 1982.

35. Y. Tohkura. "A Weighted Cepstral Distance Measure for Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, 35:1414-1422, Oct., 1987.

36. J. Kupin. "A Wireless Simulator (Software)," *CCR-P*, April, 1993.

37. H. Hermansky, N. Morgan. "RASTA Processing of Speech," *IEEE Trans. Speech, Audio Processing*, 2:578-589, October, 1994.

38. J.S. Baras, P.K. Rajasekaran. "Robustness Study of Free-text Speaker Identification and Verification," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2:379-382, 1993.

39. A. Nadas, D. Nahamoo, M.A. Picheny, "Adaptive Labeling: Normalization of Speech by Adaptive Transformation Based on Vector Quantization," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 521-524, 1988.

40. H. Gish, K. Ng and J.R. Rohlicek. "Robust Mapping of Noisy Speech Parameters for HMM Word Spotting," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2:109-112, 1992.

41. Xiaoyu Zhang, R.J. Mammone. "The Affine Transformed Cepstrum for Robust Speaker Identification," *Submitted to IEEE Trans. Speech, Audio Processing*, May, 1995.