New LP-Derived Features for Speaker Identification

Khaled T. Assaleh, *Member, IEEE* and Richard J. Mammone, *Senior Member, IEEE*

***Abstract*—A new set of features is introduced that has been found to improve the performance of automatic speaker identification systems. The new set of features is referred to as the adaptive component weighting (ACW) cepstral coefficients. The new features emphasize the formant structure of the speech spectrum while attenuating the broad-bandwidth spectral components. The attenuated components correspond to the variations in spectral tilt of transmission and recording environment, and other characteristics that are irrelevant to speaker identification. The resulting ACW spectrum introduces zeros into the usual all-pole linear prediction (LP) spectrum. This is equivalent to applying a finite impulse response (FIR) filter that *normalizes* the narrow-band modes of the spectrum. Unlike existing fixed cepstral weighting schemes, the ACW cepstrum provides an adaptively weighted version of the LP cepstrum. The adaptation results in deemphasizing the irrelevant variations of the LP cepstral coefficients on a frame-by-frame basis.**

The ACW features are evaluated for text-independent speaker identification and are shown to yield improved performance.

I. Introduction

THE purpose of a speaker identification (ID) system is to determine the identity of an individual from a sample of his or her voice. Speaker ID can be divided into two categories: closed set and open set. A closed-set speaker ID system identifies the speaker as one of those enrolled, even if he or she is not actually enrolled in the system. On the other hand, an open-set speaker ID system should be able to determine whether a speaker is enrolled or not and, if enrolled, determine his or her identity.

There is a further distinction that can be made between speaker recognition systems. They can be either text-dependent or text-independent. The system is said to be text-dependent if the same phrase is used for both training and testing. Text-independent systems usually impose no such constraints. The two main elements of a speaker ID system are feature extraction and classification. Fig. 1 shows a generic block diagram of a speaker identification system. This paper focuses on the feature extraction and preprocessing aspect of the problem of text-independent closed-set speaker ID.

Feature extraction is the process of deriving a compact set of parameters that are characteristic of a given speaker. Ideally, these parameters should efficiently preserve all the information relevant to the speaker's identity while eliminating any irrelevant information. That is, they should minimize the intraspeaker variance and at the same time maximize the interspeaker variances.

Manuscript received November 20, 1993; revised April 19, 1994. This work was supported by Contract No. F30602-91-C-0120 from Rome Laboratories.

The authors are with the CAIP Center, Rutgers University, Piscataway, NJ 08855 USA.

IEEE Log Number 9403973.

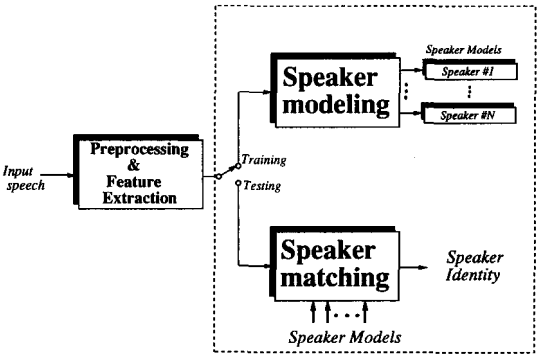


Fig. 1. Generic speaker recognition system.

The majority of speaker identification systems use some type of short-time spectral analysis. The methods used usually assume that speech is a short-time stationary process. Thus, the speech analysis is carried out using overlapping segments (frames) of 10–30 msec duration with an overlap of one-half to two-thirds of the segment length. The short-time spectrum is transformed into a sequence of feature vectors that compactly represents the underlying speech signal.

The most effective and widely used spectral analysis techniques for speech and speaker recognition applications are LP analysis [1], [2] and filter bank analysis [3]. In speech and speaker recognition applications, these two spectral analysis techniques provide comparable recognition rates [4], [6], [7]. This paper focuses on LP-derived features, although similar techniques would apply to filter bank analysis.

The short-time transfer function of the vocal tract filter obtained using the LP model is given by

$$H(z;m) = \frac{1}{A(z;m)} = \frac{1}{1+\sum_{i=1}^{P} a_i(m) z^{-i}} \qquad (1)$$

where m is the frame index, P is the order of the LP model, $a_i(m)$ is the set of prediction coefficients of the m^{th} frame, and $A(z;m)$ is the Z-transform of the inverse filter. Several sets of features can be derived from the transfer function. Atal [8] provided a comparison of the use of various LP parameters. The use of the impulse response, autocorrelation, vocal tract area function, and cepstral coefficients were investigated. The cepstrum was found to provide the best results for speaker recognition. Another comparison between the cepstrum and log area ratios (LAR's) [9] was performed for speaker verification. It was found that cepstral coefficients also outperformed the LAR's. For high quality speech, line spectral pairs (LSP's) were found to yield speaker identifica-

tion rates that are comparable to or possibly better than those of the cepstral coefficients [10]. However, for telephone quality speech, we have found that cepstral coefficients yield superior performance. Cepstral coefficients are the dominant features used for speaker recognition [9], [11], [12] and will be the focus of this paper.

The short-time LP cepstrum can be described by the Z-transform relationship

$$\ln H(z;m) = \sum_{n=1}^{\infty} c_n(m) z^{-n} \tag{2}$$

where $c_n(m)$ is the n^{th} cepstral coefficient of the m^{th} frame.

A simple and unique recursive relationship between $c_n(m)$ and the prediction coefficients $a_n(m)$ can be obtained by differentiating both sides of (2) with respect to z^{-1} and equating the coefficients of equal powers of z^{-1} [8]. An alternative view of the cepstral coefficients can be seen in terms of the poles, $z_i(m)$, of $H(z;m)$. It has been shown [13] that $c_n(m)$ can be interpreted as the power sum of $z_i(m)$ normalized by the cepstral index n

$$c_n(m) = \frac{1}{n} \sum_{i=1}^{P} z_i(m)^n. \tag{3}$$

Each pole, $z_i(m)$, is associated with a time varying center frequency $\omega_i(m)$ and bandwidth $B_i(m)$. Thus we write each pole in the form

$$z_i(m) = e^{-B_i(m) + j\omega_i(m)}. \tag{4}$$

Hence, $c_n(m)$ can be expressed as

$$c_n(m) = \frac{1}{n} \sum_{i=1}^{P} e^{-nB_i(m)} \cos(n\omega_i(m)). \tag{5}$$

Thus the n^{th} cepstral coefficient can be interpreted as a nonlinear transformation of the components center frequencies and bandwidths.

In most practical applications, speech is collected under different environments and possibly through different communication channels. This causes a mismatch among corresponding reference and testing data. The characteristics of the cepstral coefficients have been extensively studied for the purpose of minimizing such a mismatch. In this regard, two major postprocessing approaches have been introduced: intraframe processing known as cepstral weighting or liftering [4], [14]-[17], and interframe processing which exploits the time evolution of the cepstral coefficients [8], [18]-[20]. The ACW scheme introduced in this paper falls within the intraframe approach.

This paper is organized in the following way. Section II discusses several postprocessing techniques for cepstral features. The ACW scheme is also introduced in this section. Experimental results are presented in Section III. Section IV presents the summary and conclusions of the paper.

II. Cepstral Features Processing

Cepstral features are found to yield excellent performance for text-independent speaker identification when training and testing utterances are collected under relatively high-quality stationary environments. However, in practical applications, the speech waveform is subject to various sources of degradation such as background noise and communication channel variability. Such degradations often result in reduced recognition rates. This is due to the mismatch created among corresponding reference and testing data. The use of intraframe and interframe processing techniques to reduce this mismatch will be presented.

A. Intraframe Processing

Intraframe processing is also known as cepstral weighting or liftering. The rationale behind cepstral weighting is to account for the sensitivity of the low-order cepstral coefficients to the overall spectral slope and the sensitivity of the high-order cepstral coefficients to noise. In this regard several **fixed weighting** schemes have been recently introduced [16]. By "fixed weighting" it is meant that the applied weights are only a function of the cepstral index n. Therefore these weights are fixed with respect to the frame index m. Generally, the resulting weighted cepstrum is given by

$$\tilde{c}_n(m) = w_n c_n(m) \tag{6}$$

where w_n is the cepstral weighting window (also known as the lifter).

The simplest and most straightforward weighting sequence is the rectangular weights, which have the effect of truncating the infinite cepstral sequence. Other more sophisticated weighting schemes that take advantage of the statistical characteristic of the cepstral coefficients have been recently introduced. These included bandpass liftering (BPL) [14] and ramp liftering [15], [17].

It should be noted here that fixed cepstral weighting can be incorporated in the distance measure between two unweighted vectors. For example, a weighted L_2 distance measure between $c_n(i)$ and $c_n(j)$ is given by

$$d_{i,j} = \left(\sum_{n=1}^{L} w_n^2 (c_n(i) - c_n(j))^2 \right)^{\frac{1}{2}}. \tag{7}$$

The aforementioned fixed weighting schemes apply fixed weights to all the feature vectors extracted from an utterance assuming that all the frames undergo the same distortion. This assumption is not always applicable since in many practical cases distortions vary with time. In such cases an adaptive weighting scheme that is capable of adapting to the time-varying nature of the distortions is desired.

Subsequently, we shall introduce a new adaptive weighting scheme which results in a new set of cepstral features that show robustness to channel variations.

B. Adaptive Component Weighting (ACW)

The ACW scheme modifies the LP spectrum so as to emphasize the formant structure. This is achieved by operating

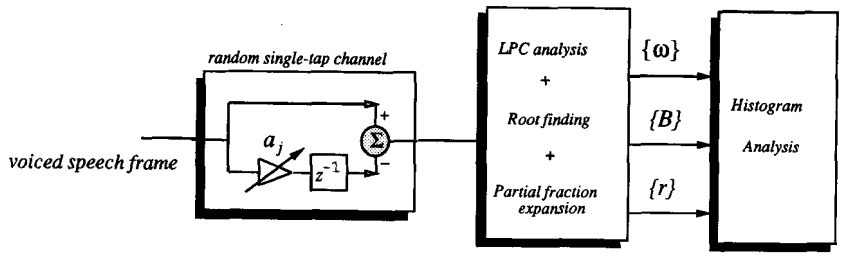


Fig. 2. Block diagram of the sensitivity of the LP spectral component parameters with respect to a random single-tap channel..

on the different components of the spectrum by amplifying the narrow-bandwidth components and attenuating the broad-bandwidth components.

The Concept of ACW: For a given speech frame, the all-pole model can be expressed in a parallel form by a partial fraction expansion

$$H(z) = \frac{1}{1 + \sum_{i=1}^{P} a_i z^{-i}} = \sum_{i=1}^{P} \frac{r_i}{(1 - z_i z^{-1})} \tag{8}$$

where r_i are the residues of the poles. Since each pole z_i represents the center frequency ω_i and the bandwidth B_i of the i^{th} component, each component can be represented by the set of the three parameters (ω_i, B_i, r_i). The formants are generally represented by narrowband components. The wider bandwidth components are generally due to channel and glottal characteristics.

It may be recalled that the pole sensitivity for an all–pole filter can be expressed by [5]

$$\frac{\partial z_i}{\partial a_k} = \frac{z_i^{P-k}}{\prod_{j=1, j\neq i}^{P}(z_i - z_j)} = r_i z_i^{P-k} \tag{9}$$

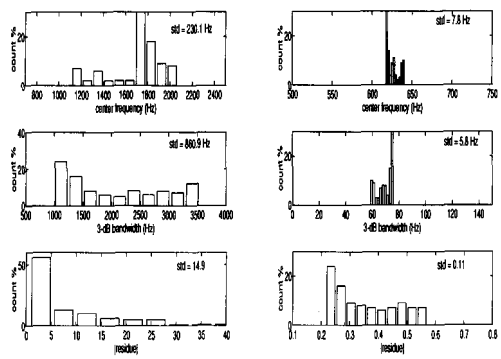
for $i = 1, 2, \ldots, P$ and $k = 1, 2, \ldots, P$. Thus the sensitivity of a pole to errors in the LP coefficients is proportional to the residues. Assuming that the distortion will affect the LP coefficients equally, the poles of the components with larger residues will be more strongly affected. This suggests the normalization of each component by its residue to reduce the effects of distortions on the feature set.

Another consideration in selecting a feature set is its sensitivity to channel variations. The sensitivity of each of the parameters (ω_i, B_i, r_i) with respect to channel variations has been experimentally evaluated by the the following experiment:

- A voiced frame of speech is processed through a random single-tap channel given by:

$$\Theta_j(z) = 1 - a_j z^{-1} \tag{10}$$

  where a_j is a random variable taken from a uniform distribution.
- Estimates of the parameters (ω_i, B_i, r_i) of all components are computed for each channel $\Theta_j(z)$, $j = 1, 2, \ldots, 1000$.
- Two sets of parameters (ω_i, B_i, r_i) are selected. One set is selected which represents a narrow-bandwidth component, and the other set is selected to represent a broad-bandwidth component.
- The sensitivity of the parameters of the selected narrow-bandwidth and broad-bandwidth components is evaluated by using a histogram analysis.



(a) (b)

Fig. 3. Histograms of the parameters of: (a) a broad-bandwidth component; and (b) a narrow-bandwidth component.

The block diagram of the experiment is shown in Fig. 2. By examining the resulting histograms of the parameters of the broad-bandwidth component shown in Fig. 3(a), one concludes that the three parameters (ω_i, B_i, r_i) associated with such components show large variances with respect to channel variations. Therefore, under channel variations, broad-bandwidth components show undesired variability that results in a mismatch condition between testing and training patterns.

Narrow-bandwidth components tend to have robust center frequencies and bandwidths. This characteristic is illustrated by the small variances of their parameters, as can be seen in the histograms in Fig. 3(b). However, the variance of their residues is relatively large, as shown in the histogram in Fig. 3(b).

These observations suggest guidelines to modify the LP spectrum so as to be robust to such variations. The modifications should include:

- Eliminating or reducing the effect of the residues r_i from the LP spectrum, and
- Attenuating the contribution of the broad-bandwidth components.

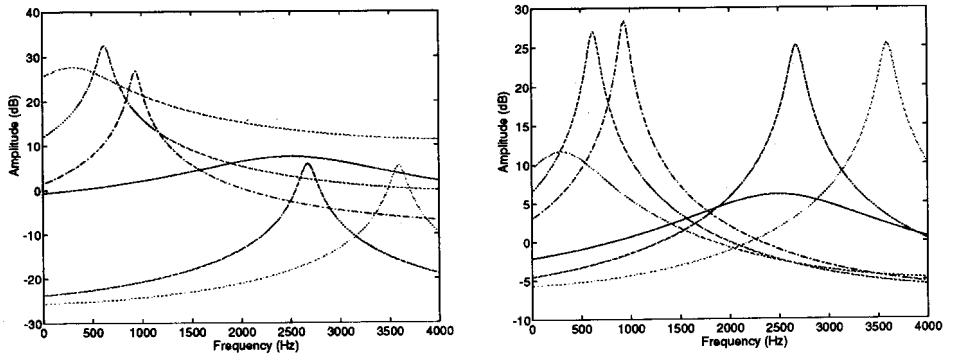


Fig. 4. Components of: (a) The LP spectrum; and (b) the ACW spectrum of a voiced speech frame.

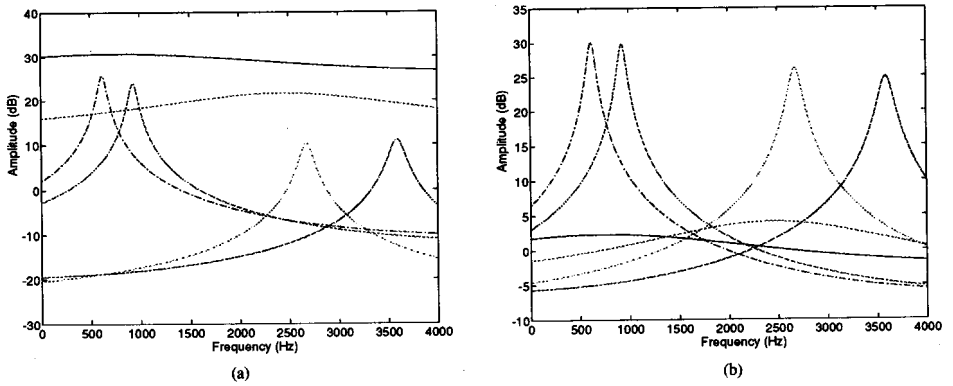


Fig. 5. Components of: (a) The LP spectrum; and (b) the ACW spectrum of a voiced speech frame processed through $(1 - 0.9z^{-1})$.

One way of achieving the suggested modifications is to normalize the components by the residues $\{r_i\}$. That is, by setting all residues to be equal to a given constant such as unity. This can be viewed as weighting the i^{th} component by $\frac{1}{r_i}$. Normalizing $\{r_i\}$ results in a modified spectrum which we refer to as the ACW spectrum. The ACW spectrum is given by

$$\hat{H}(z) = \sum_{i=1}^{P} \frac{1}{(1 - z_i z^{-1})} = \frac{N(z)}{1 + \sum_{i=1}^{P} a_i z^{-1}} \tag{11}$$

where

$$N(z) = \sum_{k=1}^{P} \prod_{i=1 \neq k}^{P} (1 - z_i z^{-1}) \tag{12}$$

which can be written in the form

$$N(z) = P\left(1 + \sum_{i=1}^{P-1} b_i z^{-i}\right). \tag{13}$$

This modification to the LP spectrum yields a peak-value of each component of

$$\frac{1}{(1 - z_i z^{-1})}\Big|_{z=e^{j\omega_i}} = \frac{1}{1 - |z_i|} \approx \frac{1}{B_i}. \tag{14}$$

Equation (14) shows that the ACW spectrum emphasizes the formant structure by weighting each component approximately by $\frac{1}{B_i}$. Thus, narrow-bandwidth components are amplified and broad-bandwidth components are attenuated.

The resulting transfer function, $\hat{H}(z)$, is no longer an all pole or autoregressive (AR) transfer function. There is a moving average (MA) term with $P - 1$ zeros. This MA filter introduced by normalizing the residues can be viewed as an FIR filter. The filter creates a spectrum whose components' peak values are inversely proportional to their bandwidths. This concept is illustrated in Fig. 4, where the components of the LP spectrum $H(z)$ and the ACW spectrum $\hat{H}(z)$ for a voiced speech frame are shown. Fig. 5 demonstrates the spectral mismatch created by a single-tap channel by showing the components of the LP spectrum of the same frame used in Fig. 4(a) after being processed through a single tap filter

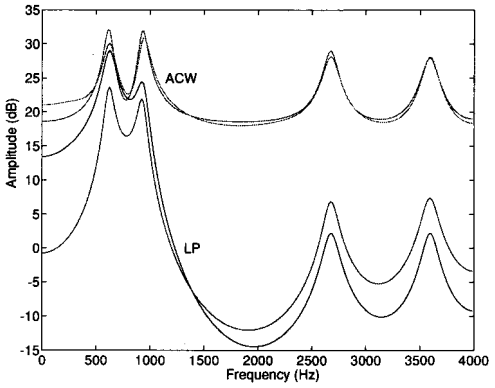


Fig. 6. The channel effect on the composite LP and ACW spectra.

$(1 - 0.9z^{-1})$. Fig. 5(b) demonstrates the robustness of the ACW spectrum. This can be seen by comparing Fig. 5(b) with 4(b). The channel effect on the composite LP and ACW spectra is shown in Fig. 6. It is clear that the mismatch between the LP spectra before and after processing through the channel is much larger than that between the corresponding ACW spectra.

ACW Cepstrum versus Conventional Cepstrum: The short time ACW cepstrum $\hat{c}_n(m)$ can be defined as the inverse Z-transform of the natural logarithm of the short-time ACW spectrum:

$$\ln \hat{H}(z;m) = \ln \sum_{i=1}^{P} \frac{1}{(1 - z_i(m)z^{-1})} = \sum_{n=1}^{\infty} \hat{c}_n(m) z^{-n}. \tag{15}$$

Whereas, the conventional LP-derived cepstrum, $c_n(m)$, is given by

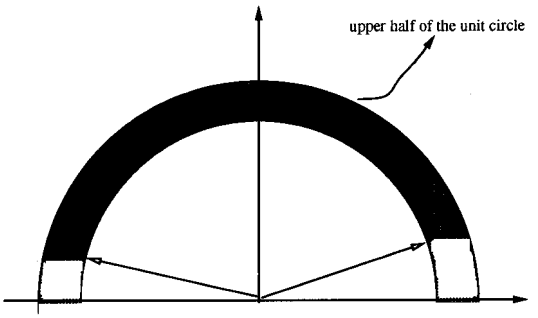
$$\ln H(z;m) = \sum_{i=1}^{P} \ln \frac{1}{(1 - z_i(m)z^{-1})} = \sum_{i=1}^{\infty} c_n(m) z^{-n}. \tag{16}$$

Note that the order of summation and logarithm are interchanged. Both methods normalize by the residue. However, the ACW cepstrum is less sensitive to the interference between components, since they are not non-linearly weighted before the summation.

ACW Computation In the cepstral domain, the introduction of the MA filter, $N(z)$, results in a subtractive component to the all-pole cepstrum. This component can be thought of as a frame-varying set of weights $w_n(m)$ applied to the LP cepstrum $c_n(m)$.

The subtractive cepstral component, $c_n^b(m)$, which is associated with $N(z)$, can be obtained by its recursive relation with $b_k(m)$:

$$c_1^b(m) = -b_1(m),$$



Frames that have a certain number of poles that lie within the shaded region are selected.

Fig. 7. Frame selection based on formant information.

$$c_n^b(m) = -b_n(m) + \sum_{k=1}^{n-1} \left(\frac{k}{n} - 1\right) b_k(m) c_{n-k}^b(m), \quad \text{for } 1 < n \le p-1,$$
$$c_n^b(m) = \sum_{k=1}^{n-1} \left(\frac{k}{n} - 1\right) b_k(m) c_{n-k}^b(m) \quad \text{for } n > P-1. \tag{17}$$

The ACW cepstrum is given by

$$\hat{c}_n(m) = c_n(m) - c_n^b(m). \tag{18}$$

Thus $\hat{c}_n(m)$ can be viewed as a frame-by-frame corrected cepstrum.

Frame Selection: The computation of the component information (center frequencies and bandwidths) is an intermediate step in obtaining the ACW features. This information can be used for selecting acceptable voiced frames for testing and training. This frame selection criterion is based on the following observations:

- Voiced speech frames provide most of the discriminative ability for speaker ID;
- Speech frames with formant-like spectra are more robust to noise, as shown by the sensitivity analysis.

Thus the formant information can be used to devise a criterion for frame selection. This criterion can be summarized as follows. Frames that have a certain number of components (usually three) that lie within a specified frequency range, and have bandwidths smaller than a specified threshold are selected. This concept is depicted in Fig. 7. This frame selection process had been found to be very important in obtaining improved recognition rates.

B. Interframe Processing

Although the contribution of this paper falls within intraframe processing methods, common interframe processing methods are discussed for completeness.

Unlike intraframe processing, interframe processing exploits the temporal variability of a sequence of feature vectors. The

rationale behind interframe processing can be summarized by the following:

- To emphasize the transitional information which is believed to provide orthogonal information to the instantaneous features obtained from the intraframe processing [18].
- To compensate for stationary and slowly varying linear channel effects that result in severe mismatch between training and testing data. This is achieved by removing time-invariant spectral information.

It has been shown [8] that the effect of any fixed frequency response distortion introduced by the recording apparatus or the transmission channel can be eliminated from a cepstral sequence simply by subtracting its long-term mean.

Transitional information is often referred to as dynamic features. Spectral dynamic features are often represented by the time differential information of the cepstral sequence, the most straightforward representation being the first difference. However, the first difference is susceptible to noise since it amplifies the high frequency components of the temporal trajectories of the cepstral coefficients. Therefore, the time derivative of $c_n(m)$ is approximated by a polynomial approximation [9]. This approximation has the effect of bandpass filtering the temporal trajectories of $c_n(m)$ instead of the highpass filtering effect of the first difference. The filtered coefficients are known as the delta-cepstral coefficients. Another intraframe processing technique is known as RASTA (RelATive SpecTrA) [19]. Similar to the delta-cepstrum, RASTA has the effect of bandpass filtering the temporal trajectories of the cepstral coefficients. However, the RASTA filter includes a first order autoregression which has the effect of recursively removing the temporal average of the cepstral sequence. It also results in smoother cepstral trajectories due to the lowpass nature of the first-order autoregression. To show the effect of intraframe processing on the short-time cepstral trajectories in the frequency domain, the frequency responses of the first difference, delta-cepstrum, and the RASTA filter are shown in Fig. 8 for a frame rate of 100 frames/sec. Experimental evaluation of the various intraframe and interframe cepstral processing techniques is given in the following section.

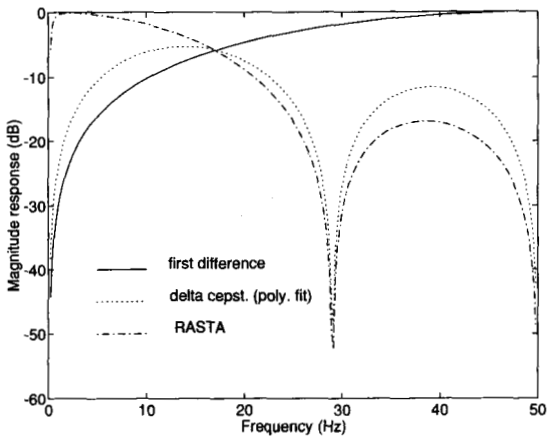


Fig. 8. Frequency responses of various interframe filters.

III. Experiments and Results

In this section we present several closed-set speaker identification experiments to evaluate different cepstral processing methods. A description of the database and preprocessing is also provided.

A. The Database

For these experiments the narrowband portion of the King database is used. Twenty-six subjects were chosen from the San Diego recordings. The first five sessions were recorded at nominal time intervals of one week, the second five were recorded at intervals of one month. In some cases, when scheduling conflicts existed or when subjects were not available for scheduled sessions, intervals between sessions were greater or less than the nominal interval. The speech material itself consisted of excerpts from conversations involving each subject and an interlocutor. The interlocutor's side of the conversation was not recorded. Conversational elicitations were designed and employed to obtain natural, extemporaneous speech from the subjects. Each session contains 20 to 30 seconds of speech excluding silence. Every conversation (session) used a different connection. Sessions 1 to 5 and sessions 6 to 10 were collected under different environments. As was observed in [21], The division of data results in serious degradations in speaker identification rates. This data is sampled at 8 kHz sampling frequency and quantized at 12 bits/sample.

For the results reported in this paper, the training is done on one session (session 1), while testing is done on each of the other nine. Due to the division of the data, testing on sessions 2 to 5 will be denoted by the "**within the great divide**" experiment, whereas testing on sessions among 6 to 10 will be denoted by the "**across the great divide**" experiment. Since each session includes 26 messages, the number of test messages are 104 for the "within the great divide" experiment, and 130 for the "across the great divide" experiment. It should be noted that the choice of the training session has been found to have no significant effect on the reported results.

B. Preprocessing

The King database consists of about 40% silence intervals. Such intervals have no speaker-dependent information, therefore, their removal is found to improve the performance of speaker identification [7].

For the following experiments, speech/silence discrimination is achieved by message-dependent energy thresholding. For each message, the energy threshold is decided by constructing the histogram of the frame energies. Only frames of energies higher than the decided threshold are kept for further processing. Following speech/silence discrimination, the speech is processed by a single-tap high frequency preemphasis filter, and partitioned into 30 ms Hamming windowed overlapping frames at a rate of 100 frames/sec.

TABLE I
BASELINE IDENTIFICATION RESULTS, "WITHIN THE GREAT DIVIDE" AND "ACROSS THE GREAT DIVIDE"

test session	identification rate
2	19/26
3	9/26
4	11/26
5	11/26
average	48.08%

"within the great divide"

test session	identification rate
6	4/26
7	2/26
8	1/26
9	3/26
10	2/26
average	9.23%

"across the great divide"

C. Classification

The classifier used here is a VQ classifier [22], [23]. During training, a codebook of 46 codewords is constructed to model each speaker. The choice of the number of codewords is related to the number of phonemes that span the feature space. Upon identifying an unknown speaker, each test vector is compared to the codebook of each speaker. The codebook entries which are closest to the test vectors are found using a full search, and the corresponding distances are recorded. The distances are accumulated for each codebook and the unknown speaker's identity is chosen as the one corresponding to the codebook associated with the minimum accumulated distance.

D. Intraframe Processing of Cepstral Features

This section demonstrates the effect of different intraframe processing techniques on the performance of speaker identification. These techniques include:

- variance normalization;
- bandpass liftering (BPL); and
- adaptive component weighting (ACW).

Table I shows the baseline speaker identification results using *unprocessed* 12^{th} order cepstral features obtained from 12^{th} order LP analysis. The division of the data is clearly seen in the "across the great divide" results. Also, due to the fact that different connections are used in different sessions, the performance of the "within the great divide" experiment shows relatively low identification rates.

Variance normalization can be either applied to the cepstral coefficients during feature extraction, or can be incorporated with the distance computation. Typically, the latter method is used. In this weighting scheme, the cepstral coefficients are weighted by the inverse of their standard deviations. This weighting scheme is found to provide no improvement in the "within the great divide" results, and slight improvement in the "across the great divide" results.

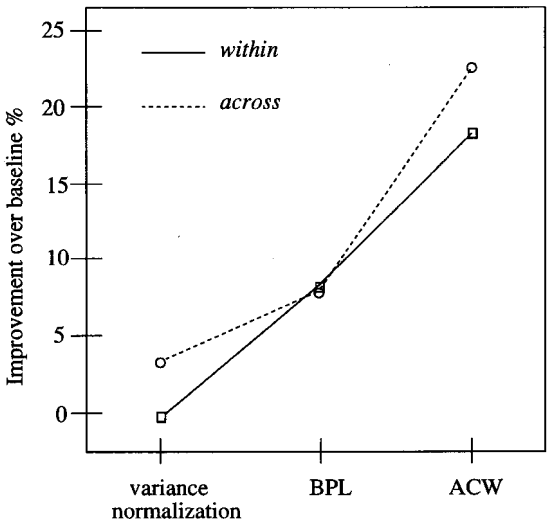


Fig. 9. Identification improvement over baseline rates using different cepstral intraframe processing techniques.

For the BPL weighting scheme, it is necessary to increase the cepstral dimension. It has been found [24] that a 20^{th} order cepstral vector derived from a 14^{th} LP analysis is a good choice. BPL is found to provide some improvement in both the "within and across the great divide results."

While variance normalization and BPL techniques apply fixed weights to the cepstral coefficients, the ACW applies adaptive weights which result in a significant improvement in the identification results for both the "within and across the great divide results."

Fig. 9 demonstrates the improvement in identification rates achieved by different intraframe processing methods. The ACW method provides the best improvement among the three compared techniques for both the "within and across the great divide" experiments. The BPL method performs better than the variance normalization technique.

E. Interframe Processing of Cepstral Features

Interframe processing exploits the temporal information of a sequence of feature vectors, therefore it is believed to provide orthogonal information to the intraframe processing.

In this section the following interframe processing techniques are evaluated:

- FIR filtering (delta-cepstrum);
- RASTA processing;
- long-term mean removal.

Fig. 10 shows the improvement in identification achieved by different interframe processing methods. It is clearly seen that the long-term mean removal method provides the best results for both the "within and across the great divide" experiments. RASTA and *delta* processing are found to provide no improvement over baseline rates for the "within the great divide" experiment. However, some improvement is obtained in the

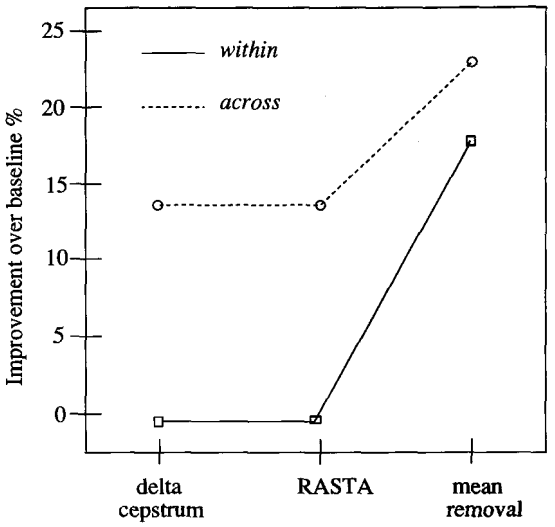


Fig. 10. Identification improvement over baseline rates using different cepstral interframe processing techniques.

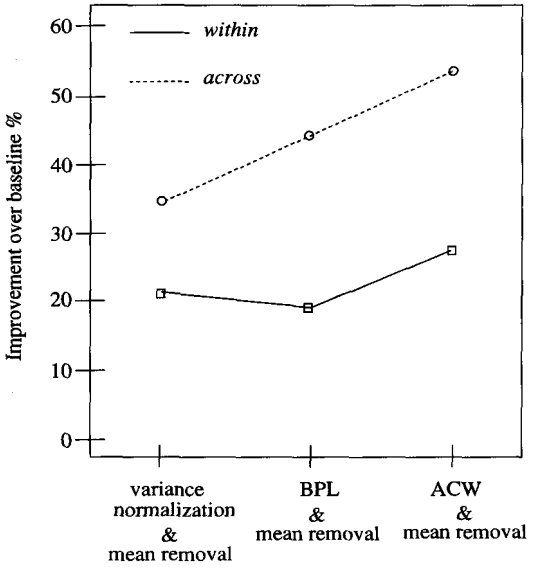


Fig. 11. Identification improvement over baseline rates using different cepstral intra-interframe processing techniques.

"across the great divide" results. It should be noted that long-term mean removal requires the availability of the feature vectors of the whole utterance, while delta or RASTA operate locally on a small number of feature vectors (in this case, five vectors).

F. Combining Intraframe and Interframe Processing

Since intraframe and interframe processing operate on a given sequence of cepstral features in two different domains, it is expected that their combination will result in further improvement in the identification rates. The combination is accomplished by cascading the two processing techniques. This will be referred to as intra-interframe processing.

Experiments are conducted with all the combinations of intra-interframe processing techniques. Among interframe processing techniques, long-term mean removal is found to provide the best performance.

This subsection demonstrates the effect of the following intra-interframe processing techniques:

- variance normalization followed by mean removal;
- BPL followed by mean removal;
- ACW followed by mean removal.

The ACW weighting scheme followed by long-term mean removal is found to provide the best performance for both the "within and across the great divide" experiments. Improvement rates over baseline performance of the three intra-interframe processing techniques are shown in Fig. 11.

The improvement due to the combined intraframe and interframe processing techniques is comparable to the sum of the improvements achieved by applying these techniques separately. Thus the intraframe and interframe enhancements appear to be relatively independent of each other.

IV. Summary and Conclusion

In this paper we presented a new LP-based feature set called the ACW cepstrum. The new features were applied to the problem of closed-set text-independent speaker identification.

The development of the ACW scheme is motivated by the characteristics of the parameters of the parallel form of the all-pole model. In the spectral domain, the ACW scheme has several effects on the components of the LP spectrum. It emphasizes the formant information, and attenuates the broad-bandwidth spectral components which are susceptible to noise and transmission over communications channels. In the cepstral domain, the ACW scheme can be viewed as an intraframe cepstral processing technique by which frame-dependent (adaptive) weights are applied to the LP cepstra. The performance of the ACW cepstrum is compared with that of other cepstral weighting schemes using the San Diego speakers of the narrowband portion of the King database. The ACW cepstrum is found to perform significantly better than the variance normalization and the BPL intraframe processing techniques. The experiments given in Section III emphasize the mismatch among the different sessions of the database, and to demonstrate the relative effectiveness of different cepstral processing techniques. Therefore, the relative improvement in the identification rates has more significance than the identification rates themselves.

Interframe cepstral processing exploits the temporal variations of the cepstral trajectories. The performance of several interframe techniques are evaluated. Long-term mean removal is shown to perform significantly better than delta-cepstrum and RASTA. Further improvement is achieved by combining intraframe and interframe processing. It was found that the improvement due to combined intraframe and interframe pro-

cessing is approximately equal to the sum of the improvements due to intraframe processing and interframe processing applied separately. The combination of ACW and long-term mean removal provided the highest identification rates.

ACKNOWLEDGMENT

The authors would like to thank J. Couples, J. Flanagan, J. Grieco for their useful comments and suggestions.

REFERENCES

[1] B. S. Atal. "Speech analysis/synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637–655, 1971.
[2] J. Makhoul, "Linear prediction: A Tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.
[3] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–28, pp. 357–366, 1980.
[4] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: 1993.
[5] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1989.
[6] D. Reynolds, "Evaluation of different features for speaker identification," presented at the Robust Speech Recognition Workshop, Rutgers University, August 1993.
[7] Yu-Hung Kao, "Robustness study of free-text speaker identification and verification," Ph.D. thesis, Univ. of Maryland, Dec. 1992.
[8] 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.
[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, pp. 254–272, Apr. 1981.
[10] J. P. Campbell, "Features and measures for speaker recognition." Pd.D. thesis, Oklahoma State Univ., Dec. 1992.
[11] F. K. Soong *et al.*, "A vector quantization approach to speaker recognition," *Proc. Int. Conf. ASSP*, 1985, pp. 387–390.
[12] G. Velius, "Variants of cepstrum based speaker identity verification," *Proc. Int. Conf. ASSP*, 1988, pp 583–586.
[13] M. R. Schroeder, "Direct (nonrecursive) relations between cepstrum and and predictor coefficients," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-29, pp. 297–301, Apr. 1981.
[14] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 947–954, July 1987.
[15] K. K. Paliwal, "On the performance of the frequency weighted cepstral coefficients in vowel recognition". *Speech Commun.*, vol. 1, pp. 151–154, 1982.
[16] Y. Tohkura, "A weighted cepstral measure for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 947–954, Oct. 1987.
[17] H. Wakita, "Spectral slope based distortion measure for all pole models of speech," *Proc. Int. Conf. ASSP*, 1986, pp. 757–760.
[18] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp., 871–879, June 1988.
[19] H. Hermansky *et al.*, "RASTA-PLP speech analysis technique," *Proc. Int. Conf. ASSP*, 1992, pp. I-121–124.
[20] S. Furui, "On the role of dynamic characteristics of speech spectra for syllable perception," *Fall Meeting Acoust. Soc. Japan*, Oct. 1984, pp. 1–1–2..
[21] H. Gish, "Robust discrimination in automatic speaker identification," *Proc. Int. Conf. ASSP*, 1990, pp. 289–292.
[22] R. Gray, "Vector quantization," *IEEE ASSP Mag.*, pp. 4–29, Jan. 1984.
[23] T. Kohonen, *Self-Organization and Associative Memory* New York: Springer-Verlag, 1989, 3rd ed.
[24] Y. Kao, J. Baras, and P. Rajasekaran, "Robustness study of free-text speaker identification and verification," *Proc. Int. Conf. ASSP*, pp. 379–382, 1993.

Khaled T. Assaleh (S'91-M'93) received the B.S. in electrical engineering from the University of Jordan, Amman, in 1988. He received the M.S. degree from Monmouth College, West Long Branch, NJ in 1990 and the Ph.D. degree from Rutgers University, Piscataway, NJ in 1993, both in electrical engineering.

He is currently a Research Professor at the CAIP Center of Rutgers University. During 1989-1990 he worked at Bell Communications Research on low bit rate video coding. His current research interests are in the areas of robust speech and speaker recognition.

Richard J. Mammone (S'75-M'81-SM'86) received the B.E.E., M.E.E., and the Ph.D. degrees from the City University of New York in 1975, 1977, and 1981, respectively.

He is currently Professor of Electrical and Computer Engineering at Rutgers University, Piscataway, NJ. His research and teaching interests are in the areas of image and speech processing and neural networks. He has published numerous articles, and edited three books,and two special issues of international journals.

Dr. Mammone was an Associate Editor of the journal *Pattern Recognition* and an Associate Editor of *IEEE Communications Magazine*. He is a frequent consultant to industry and government agencies. He is a member of OSA, SPIE, Eta Kappa Nu, and Sigma Xi.