Speaker Identification Based on the Use of Robust Cepstral Features Obtained from Pole-Zero Transfer Functions

Mihailo S. Zilovic, Ravi P. Ramachandran, *Member, IEEE,* and Richard J. Mammone, *Senior Member, IEEE*

***Abstract*—A common problem in speaker identification systems is that a mismatch in the training and testing conditions sacrifices much performance. We attempt to alleviate this problem by proposing new features that show less variation when speech is corrupted by convolutional noise (channel) and/or additive noise. The conventional feature used is the linear predictive (LP) cepstrum that is derived from an all-pole transfer function which, in turn, achieves a good approximation to the spectral envelope of the speech. Recently, a new cepstral feature based on a pole-zero function (called the *adaptive component weighted* or ACW cepstrum) was introduced. We propose four additional new cepstral features based on pole-zero transfer functions. One is an alternative way of doing adaptive component weighting and is called the ACW2 cepstrum. Two others (known as the PFL1 cepstrum and the PFL2 cepstrum) are based on a pole-zero postfilter used in speech enhancement. Finally, an autoregressive moving-average (ARMA) analysis of speech results in a pole-zero transfer function describing the spectral envelope. The cepstrum of this transfer function is the feature. Experiments involving a closed set, text-independent and vector quantizer based speaker identification system are done to compare the various features. The TIMIT and King databases are used. The ACW and PFL1 features are the preferred features, since they do as well or better than the LP cepstrum for all the test conditions. The corresponding spectra show a clear emphasis of the formants and no spectral tilt. To enhance robustness, it is important to emphasize the formants. An accurate description of the spectral envelope is not required.**

***Index Terms*— Cepstrum, channel, linear prediction, noise, pole-zero transfer function, speaker identification.**

I. Introduction

SPEAKER recognition is the task of identifying a speaker by his or her voice. Systems performing speaker recognition operate in different modes. A closed set mode is the situation of identifying a particular speaker as one in a finite set of reference speakers [1]. In an open set system, a speaker is either identified as belonging to a finite set or is deemed not to be a member of the set [1]. For speaker verification, the claim of a speaker to be one in a finite set is either accepted or rejected [2]. Speaker recognition can either be done as a text-dependent or text-independent task. The difference is that in the former case, the speaker is constrained as to what must be said while in the latter case, no constraints are imposed.

The overall system that we consider will have three components:

1) linear predictive (LP) analysis for parameterizing the spectral envelope;
2) feature extraction for ensuring speaker discrimination;
3) classifier for making a decision.

The input to the system will be a speech signal possibly corrupted by noise and possibly influenced by other environmental conditions (like channel effects). The output will be a decision regarding the identity of the speaker. A robust system performs the recognition task successfully even when the speech is corrupted by noise and/or communication channel effects. The ideal situation is to achieve a high performance in terms of recognition accuracy given any type of speech material. The concentration of the work will be on the development of robust LP derived features in a closed set, text-independent mode. Note that existing methods will be used for the first and third components of the system.

After LP analysis of speech [3] is carried out, various equivalent representations of the LP parameters exist. A comparison of these parameters in terms of speaker recognition accuracy revealed that the LP cepstrum is the best when training and testing is done on clean speech [4]. The problem with the LP cepstrum is that a mismatch in training and testing conditions sacrifices much performance, thereby diminishing the robustness. The LP cepstrum is derived from an all-pole transfer function that describes the spectral envelope of the speech. This in particular gives information about the formants that is crucial for speaker recognition to be successful. Our attempt in finding more robust features is to first transform the all-pole transfer function derived from LP analysis into a pole-zero transfer function that gives more emphasis to the formants. The cepstrum of the pole-zero transfer function is the feature. Various new approaches that convert an all-pole function into a pole-zero function are formulated and compared. The question of why a two-step route that goes from the speech to a pole-zero transfer function emerges. We also consider a pole-zero model obtained by a direct autoregressive moving average (ARMA) analysis of the speech as the first component of the system. However, as revealed later, the performance obtained by an ARMA approach is inferior to

Manuscript received March 25, 1995; revised August 8, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joseph Campbell.

M. S. Zilovic is with Bell Communications Research, Red Bank, NJ 07701 USA.

R. P. Ramachandran is with the Department of Electrical Engineering, Rowan University, Glassboro, NJ 08028 USA (e-mail: ravi@rowan.edu).

R. J. Mammone is with the Computer Aids for Industrial Productivity Center, Rutgers University, Piscataway, NJ 08854 USA.

Publisher Item Identifier S 1063-6676(98)02901-0.

that of using a pole-zero transfer function derived after LP analysis.

II. Parameterization of Spectral Envelope

The first component of the system transforms the speech signal into a compact representation of its spectral envelope. A linear predictive (LP) analysis [3] is used for this purpose. An LP analysis of a speech signal, based on the model that a speech sample is a weighted linear combination of p previous samples, results in a set of weights a_k. The fundamental equation governing this model is

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n) \tag{1}$$

where $s(n)$ is the speech signal and $e(n)$ is the error or LP residual. These weights correspond to the direct form coefficients of a nonrecursive filter

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = \prod_{k=1}^{p} (1 - f_k z^{-1}) \tag{2}$$

where f_k for $1 \leq k \leq p$ represent the zeros of $A(z)$. Passing the speech signal through the filter $A(z)$ results in the LP residual $e(n)$ that is free of near-sample redundancies. The determination of the LP coefficients a_k is usually based on minimizing the weighted mean squared-error E_{mse} over a segment of speech consisting of N samples. In the minimization of E_{mse} using the autocorrelation approach [3], the coefficients a_k are found by solving a system of linear equations. Moreover, $A(z)$ is guaranteed to be minimum phase. The magnitude spectrum of $1/A(z)$ describes the spectral envelope of the speech. Since $1/A(z)$ is completely specified by its poles f_k, the LP analysis is based on an all-pole model.

An ARMA analysis leads to a transfer function $U(z)/V(z)$ that approximates the spectral envelope. We use Shanks method [5] to determine the coefficients of $U(z)$ and $V(z)$. In this approach, a minimum phase $V(z)$ is first determined by LP analysis and is equal to $A(z)$. The impulse response of $1/V(z)$ is $h(n)$, which is truncated to N samples as the segment of speech being analyzed consists of N samples. The error is $s(n) - h(n) \star u(n)$ where $u(n)$ is the finite impulse response of $U(z)$. Upon minimization of the mean-square error, the coefficients of $U(z)$ are found by solving a system of linear equations. Although $U(z)$ is not guaranteed to be minimum phase, this property can be forced by reflecting the zeros of $U(z)$ outside the unit circle to lie inside. The order of $U(z)$ is determined empirically so as to achieve an acceptable approximation of the spectral envelope.

III. Feature Extraction

The first component either gives an all-pole or pole-zero transfer function. The feature extractor generally performs a transformation of the function and then computes the cepstrum as the feature vector. Suppose a pole-zero transfer function

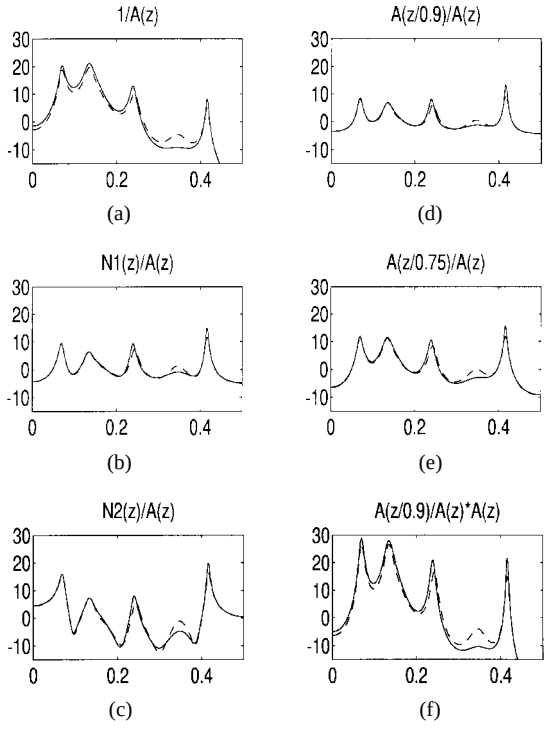

Fig. 1. Various spectra when speech is corrupted by additive white Gaussian noise (SNR of 20 dB). Clean speech, solid line; noisy speech, dotted line. (a) Magnitude response of LP filter. (b) Magnitude response of ACW transfer function. (c) Magnitude response of ACW2 transfer function. (d) Magnitude response of postfilter $H_{pf}(z)$ ($\alpha = 1$, $\beta = 0.9$). (e) magnitude response of postfilter $H_{pf}(z)$ ($\alpha = 1$, $\beta = 0.75$). (f) Spectral envelope of postfiltered speech $T(z)$ ($\alpha = 1$, $\beta = 0.9$).

$P(z)$ is given by

$$P(z) = \frac{U(z)}{V(z)} = \frac{\prod_{k=1}^{u}(1 - u_k z^{-1})}{\prod_{k=1}^{v}(1 - v_k z^{-1})}. \tag{3}$$

If $P(z)$ is minimum phase, the cepstrum $c_p(n)$ can be obtained either by a computationally efficient recursion based on the polynomial coefficients or by considering the polynomial roots u_k and v_k as given [6] by

$$c_p(n) = \frac{1}{n}\sum_{k=1}^{v} v_k^n - \frac{1}{n}\sum_{k=1}^{u} u_k^n \tag{4}$$

for $n > 0$.

The first feature we consider is the conventional LP cepstrum of the all-pole LP filter $1/A(z)$. This serves as a benchmark to which we compare our proposed features. For the next four features, the all-pole LP transfer function $1/A(z)$ is transformed into a pole-zero function. It is known that the mean-square difference between two cepstral vectors is directly related to the mean-square difference in the magnitude spectra of the transfer functions from which the cepstral vectors were derived from [6]. The magnitude spectra of $1/A(z)$ obtained from clean and corrupted speech shows a degree of dissimilarity even around the formant regions [see Figs. 1(a), 2(a), and 3(a)]. This is manifested as a clear difference in

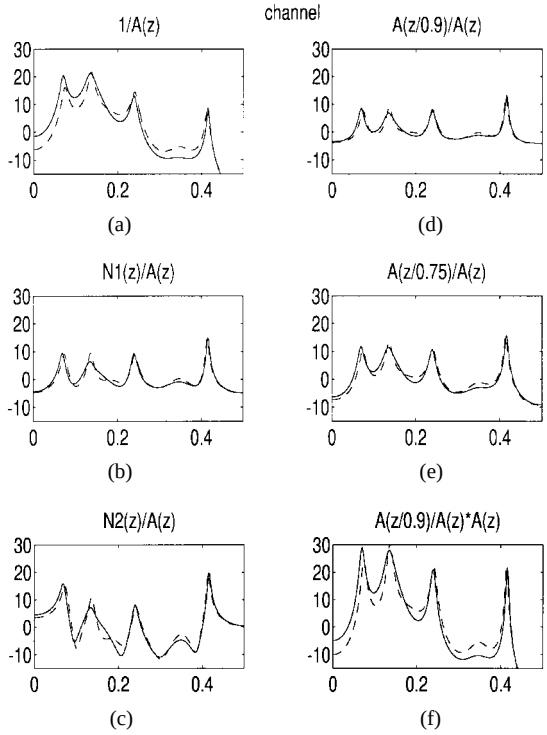

Fig. 2. Various spectra when speech is passed through the IRS filter. Clean speech, solid line; corrupted speech, dotted line. (a) Magnitude response of LP filter. (b) Magnitude response of ACW transfer function. (c) Magnitude response of ACW2 transfer function. (d) Magnitude response of postfilter $H_{pf}(z)$ ($\alpha = 1$, $\beta = 0.9$). (e) Magnitude response of postfilter $H_{pf}(z)$ ($\alpha = 1$, $\beta = 0.75$). (f) Spectral envelope of postfiltered speech $T(z)$ ($\alpha = 1$, $\beta = 0.9$).

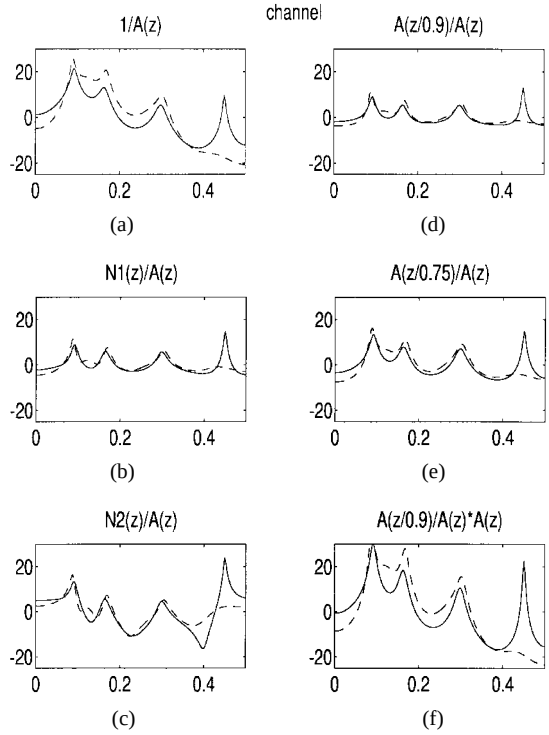

Fig. 3. Various spectra when speech is passed through the CMV filter. Clean speech, solid line; corrupted speech, dotted line. (a) Magnitude response of LP filter. (b) Magnitude response of ACW transfer function. (c) Magnitude response of ACW2 transfer function. (d) Magnitude response of postfilter $H_{pf}(z)$ ($\alpha = 1$, $\beta = 0.9$). (e) Magnitude response of postfilter $H_{pf}(z)$ ($\alpha = 1$, $\beta = 0.75$). (f) Spectral envelope of postfiltered speech $T(z)$ ($\alpha = 1$, $\beta = 0.9$).

the cepstral vectors which causes a performance degradation. Our objective is to transform the all-pole transfer function into a pole-zero transfer function such that the difference in the magnitude spectra decreases when noise is added to the speech and/or the speech is passed through a channel. We use a recently introduced approach [7] for comparison purposes and formulate three novel approaches.

The existing approach as developed in [7] is to first perform a partial fraction expansion of $1/A(z)$ to get

$$\frac{1}{A(z)} = \sum_{k=1}^{p} \frac{\lim_{z \to f_k} [(1 - f_k z^{-1})/A(z)]}{1 - f_k z^{-1}} = \sum_{k=1}^{p} \frac{r_k}{1 - f_k z^{-1}}. \tag{5}$$

The experiments in [7] reveal that the residues r_k show considerable variations especially for nonformant poles when the speech is degraded. Therefore, the variations in r_k were removed by forcing $r_k = 1$ for every k. Hence, the transfer function is a pole-zero type of the form

$$\frac{N(z)}{A(z)} = \sum_{k=1}^{p} \frac{1}{1 - f_k z^{-1}} = \frac{1}{A(z)} \sum_{k=1}^{p} \prod_{i=1 \neq k}^{p} (1 - f_i z^{-1})$$
$$= p \frac{1 - \sum_{k=1}^{p-1} b_k z^{-k}}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{6}$$

It has been shown in [8] that $N(z)$ is the derivative of $A(z)$ with respect to z and hence, the coefficients b_k are easily found from a_k as $b_k = (p-k)a_k/p$ for $k = 1$ to $p-1$. The mismatch in the magnitude spectra of $N(z)/A(z)$ for clean and corrupted speech is reduced over that of $1/A(z)$ [see Figs. 1(b), 2(b), and 3(b)]. The numerator polynomial $N(z)$ is guaranteed to be minimum phase [8]. The cepstrum of $N(z)/A(z)$ is used as the feature vector and can be obtained by an efficient recursion based on the polynomial coefficients. This method is known as *adaptive component weighting* (ACW) and is primarily used for mitigating channel effects [7].

Our first new approach is an alternative to the ACW method. From the perspective of system analysis, the LP filter $1/A(z)$ can be viewed as the cascade connection of p first order filters having a transfer function $1/(1 - f_k z^{-1})$. Connecting these p first-order sections in parallel results in the overall pole-zero transfer function for the ACW method [see (6)]. Using a similar reasoning, $1/A(z)$ can be interpreted as a cascade connection of second-order sections (pairs of first-order sections). The parallel combination of these $p/2$ second-order sections gives rise to another overall pole-zero transfer function. We refer to this as the ACW2 approach. For the initial cascade connection, the question of which first-order sections to pair up emerges. We choose to pair up the first order sections specified by the complex conjugate poles of $1/A(z)$. Any remaining real poles are also paired up. Suppose that among the p poles f_k, there are c complex poles and $p-c$ real poles. The complex poles are arranged as f_1, $f_1^\star$, f_2, $f_2^\star$,

$\cdots$, $f_{c/2}$, $f^{\star}_{c/2}$ where $f^{\star}_{k}$ is the complex conjugate of f_k. The remaining real poles are arranged as f_{c+1}, f_{c+2}, $\cdots$, f_{p-1}, f_p. In this case, the pole-zero transfer function is given as

$$\frac{N(z)}{A(z)} = \sum_{k=1}^{c/2} \frac{1}{(1-f_k z^{-1})(1-f_k^{\star} z^{-1})} + \sum_{k=0}^{(p-c)/2-1} \frac{1}{(1-f_{c+2k+1} z^{-1})(1-f_{c+2(k+1)} z^{-1})}. \tag{7}$$

In practice, we have observed that if real poles are present, there are only two of them for the case when $p = 12$ assuming 8 kHz sampled speech. Therefore, the optimal real pole pairing is not a practical issue. The motivation of pairing up complex conjugate pairs is based on the fact that the impulse response of a second-order section specified by a complex conjugate pole pair is a damped sinusoid. This provides for a more natural pole-zero model of the speech signal, representing it as a superposition of amplitude modulated sinusoids. We conjecture that $N(z)$ is minimum phase since no instance of a nonminimum phase $N(z)$ was found in practice. In a real system, any roots of $N(z)$ outside the unit circle should be reflected inside. Again, the cepstrum of $N(z)/A(z)$ is used as the feature vector.

The other family of pole-zero transfer functions that we formulate is based on the concept of a postfilter that was introduced in [9] 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

$$H_{pf}(z) = \frac{A(z/\beta)}{A(z/\alpha)} \qquad 0 < \beta < \alpha \leq 1. \tag{8}$$

The spectrum of $H_{pf}(z)$ emphasizes the formant peaks. The spectral envelope of the postfiltered speech is determined as the magnitude response of

$$T(z) = \frac{A(z/\beta)}{A(z)A(z/\alpha)}. \tag{9}$$

If $A(z)$ is minimum phase, both $H_{pf}(z)$ and $T(z)$ are guaranteed to be minimum phase. The cepstrum of both the pole-zero transfer functions $H_{pf}(z)$ and $T(z)$ are used as the feature vectors. The cepstrum of $H_{pf}(z)$ can be shown to be equivalent to weighting the LP cepstrum by a factor $\alpha^n - \beta^n$. The cepstrum of $T(z)$ can be shown to be equivalent to weighting the LP cepstrum by a factor $1 + \alpha^n - \beta^n$. Other different ways of weighting the LP cepstrum (like frequency weighting, inverse variance weighting and bandpass weighting) have been considered in [10]–[12]. The weightings we propose have an interpretation in terms of transfer functions. Also, like the weightings in [10], [11], the lower indexed cepstral coefficients are deemphasized. We will examine the effect of these weightings on the spectrum and on the speaker identification performance.

Fig. 1 shows the magnitude responses of the various transfer functions for a frame of clean speech and for the same frame of speech corrupted by additive white Gaussian noise. The signal to noise ratio (SNR) is 20 dB. There is a certain mismatch in the spectra of $1/A(z)$ as mentioned earlier and revealed in Fig. 1(a). We attempt to alleviate this mismatch by introducing the various pole-zero transfer functions. As can be seen in Fig. 1(b) and (c), the mismatch in the magnitude spectrum for the ACW and ACW2 methods is reduced over that of $1/A(z)$. It should be pointed out that the ACW2 spectrum shows very sharp peak values. Also, the amplitudes of the valleys are more equal for the ACW spectrum than the ACW2 spectrum. In analyzing the magnitude response of $H_{pf}(z)$ as shown in Fig. 1(d) and (e), note the similarity between it and the ACW spectrum. The formant amplitudes are emphasized without causing any spectral tilt. The response of the postfilter 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. By comparing Fig. 1(d) and (e), it can be seen that as β decreases, the spectral tilt becomes more apparent. The spectrum of the postfiltered speech [see Fig. 1(f)] shows some spectral tilt but reflects the spectral envelope of the enhanced speech, which is desired to be more like that of clean speech. The formant peaks are amplified and the valleys are depressed.

Fig. 2 shows the magnitude responses of the LP filter and of the pole-zero transfer functions when speech is passed through the intermediate reference mask (IRS) channel. A similar figure (Fig. 3) shows the responses when speech is passed through the continental mid voice (CMV) channel [13], [14]. Both the IRS and CMV channels are representative of telephone channels. Again, it is observed that the pole-zero transfer functions lower the spectral mismatch over that of the all-pole LP filter.

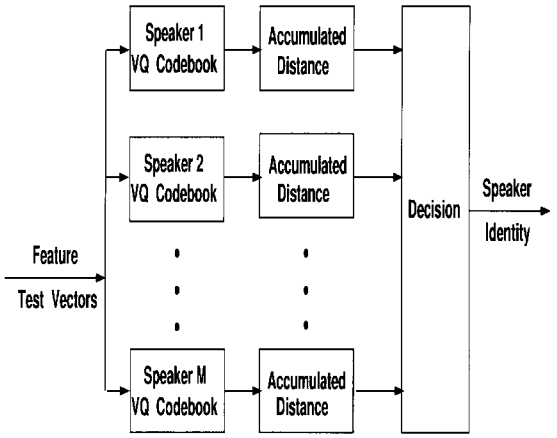

Fig. 4. Block diagram of VQ based speaker identification system.

IV. Vector Quantizer Classifier

A vector quantizer (VQ) classifier [15], [16] is used to render a decision as to the identity of a speaker. Note that we are not restricted to this type of classifier for the features we propose. A VQ classifier is used since it is known to perform very well and will make our results extremely reliable. The system is shown in Fig. 4. For each speaker, a training set of feature vectors is used to design a VQ codebook based on

the Linde–Buzo–Gray (LBG) algorithm [17]. There will be M codebooks, one pertaining to each of the M speakers.

To test the system, a test utterance from one of the M speakers is converted to a set of test feature vectors. Consider a particular test feature vector. This is quantized by each of the M codebooks. The quantized vector is that which is closest according to some distance measure to the test feature vector. We use the squared Euclidean distance as the measure. Hence, M different distances are recorded, one for each codebook. This process is repeated for every test feature vector. The distances are accumulated over the entire set of feature vectors. The codebook that renders the smallest accumulated distance identifies the speaker. When many utterances are tested, the success rate is the number of utterances for which the speaker is identified correctly divided by the total number of utterances tested.

The VQ codebooks will be trained for one particular condition, namely, for clean speech. Different test conditions corresponding to clean and corrupted speech will be used to provide a definitive and quantitative evaluation of robustness. If a feature is robust, a mismatch between the testing and training conditions should cause slight degradation in performance or success rate.

V. Experimental Protocol and Results

The experimental approach is described below. Prior to any analysis, the speech is preemphasized by using a nonrecursive filter $1 - 0.95z^{-1}$. For the LP analysis, the autocorrelation method [3] is used to get a 12th-order LP polynomial $A(z)$. For the ARMA analysis using Shanks method [5], the denominator polynomial is the LP polynomial. A sixth-order numerator polynomial is then computed. Both types of analyses are done over frames of 30 ms duration. The overlap between frames is 20 ms. The all-pole function $1/A(z)$ is converted into the conventional LP cepstrum of dimension 12. For the other four features described above, the all-pole function is first transformed into a pole-zero transfer function. The 12-dimensional (12-D) cepstrum of the pole-zero function is the feature vector. Similarly, the pole-zero transfer function derived from an ARMA analysis is converted into a 12-D cepstrum, which we denote as the ARMA cepstrum. The feature vectors are computed only in voiced frames. The voiced frames are selected based on energy thresholding and by the presence of at least three LP poles in an annular region close to the unit circle (formant poles). The latter concept of considering LP poles for frame selection was introduced in [7]. The VQ classifier [15], [16] (as described earlier) is trained using the 12-D feature vectors. A separate classifier is used for each feature. The distance measure is the squared Euclidean distance. The codebooks for each speaker are designed using the LBG algorithm [17]. The test speech material corresponds to various conditions. The performance of the features under mismatched training and testing conditions is a good indicator of robustness. The performance measure is the speaker identification success rate.

Two data bases are used in the experiments. For the TIMIT data base that comprises only clean speech, 20 speakers from the New England dialect are considered. The speech is downsampled from 16 to 8 kHz. For each speaker, there are ten sentences. The first five are used for training the VQ classifier. Therefore, the classifier is trained on clean speech only. The remaining five sentences are individually used for testing. One of the test conditions corresponds to clean speech for which there are 100 test utterances over which the speaker identification success rate is computed. Various other test conditions are simulated by adding different types of noise and passing the speech through different channels. For each channel test condition, there are again 100 test utterances. For each of the noise conditions, the ability to use different seeds to generate random noise permits 300 trials.

TABLE I
Identification Success Rate as a Percent for Clean Speech (TIMIT Data Base). The Three Success Rates Correspond to Codebook Sizes of 16, 32, and 64

Feature	Identification Success Rate
LP Cepstrum	91 96 94
ACW	92 93 91
ACW2	90 96 93
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	92 92 95
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	89 94 95
ARMA cepstrum	65 66 66

The King data base consisting of 26 San Diego and 25 Nutley speakers is also used. The speech is recorded over long distance telephone lines and sampled at 8 kHz. There are ten recording sessions, each having one utterance per speaker. The data is divided such that there is a big mismatch in the conditions between sessions 1 to 5 and sessions 6 to 10. This mismatch is due to a change in the recording equipment, which translates to a significantly changed environment [18]–[20]. Training is done on session 1. Testing "within the great divide" corresponds to the utterances in sessions 2 to 5 in which there is some mismatch with session 1. Testing "across the great divide" corresponds to the utterances in sessions 6 to 10, which in turn provide a big mismatch. Additional results are obtained as follows. Training is done on session 2 while the remaining nine sessions are used for testing. For the experiments, the total number of test utterances "within the great divide" is 208 for the San Diego portion and 200 for the Nutley portion. The total number of test utterances "across the great divide" is 260 for the San Diego portion and 250 for the Nutley portion.

A. Testing on Clean Speech

The first experiment involves the testing of clean speech, which is performed by using the TIMIT data base. Table I shows the results. The performance does not always monotonically increase as the codebook size gets bigger. Therefore, merely using a large codebook size does not benefit in terms of performance and imposes a cost in terms of memory and search complexity. In the limit as the codebook size equals the number of vectors in the training set, a nearest neighbor classifier is obtained. Experiments have shown that the nearest neighbor classifier is inferior to the VQ technique using modest size codebooks [21]. This is because overlearning of the training data has taken place. For a codebook size of 32 (which is practically very feasible), the cepstrum and the ACW2

TABLE II
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR SPEECH DEGRADED BY ADDITIVE WHITE GAUSSIAN NOISE (TIMIT DATA BASE). THE THREE SUCCESS RATES CORRESPOND TO CODEBOOK SIZES OF 16, 32, AND 64

Feature	Test Condition		
	Noisy speech 30 dB SNR	Noisy speech 20 dB SNR	Noisy speech 10 dB SNR
LP Cepstrum	79 85.3 86.3	47 56.3 61.3	18.7 24.7 21
ACW	82.3 84.7 87	57 64.7 64	26.3 26.7 23.3
ACW2	84.3 88.3 91.3	50.7 63.3 60.7	19.3 23.7 23.3
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	87 83.3 86	63 67 68	27 28.3 22.7
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	82.3 85 88.7	52.7 62.7 63.3	22.3 24 23
ARMA Cepstrum	60 60.3 60.7	45.7 43.3 46.7	21 20.3 22.3

TABLE III
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR SPEECH DEGRADED BY COLORED NOISE (TIMIT DATA BASE). THE THREE SUCCESS RATES CORRESPOND TO CODEBOOK SIZES OF 16, 32, AND 64

Feature	Test Condition		
	Noisy speech 30 dB SNR	Noisy speech 20 dB SNR	Noisy speech 10 dB SNR
LP Cepstrum	88 94.3 92.3	84.7 86 86	38 43 44.3
ACW	92.3 93 94.3	81.7 84.3 82.7	37 36.7 39.3
ACW2	88.3 94.3 94.7	74.7 78.7 79.7	38.7 36 40.7
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	91.3 90.7 93.7	84 84.3 87.3	42 44.3 43.7
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	84.7 92 96.3	84.7 88 92.3	44.3 46.3 50.3

TABLE IV
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR SPEECH DEGRADED BY BABBLE NOISE (TIMIT DATA BASE). THE THREE SUCCESS RATES CORRESPOND TO CODEBOOK SIZES OF 16, 32, AND 64

Feature	Test Condition		
	Noisy speech 30 dB SNR	Noisy speech 20 dB SNR	Noisy speech 10 dB SNR
LP Cepstrum	89.7 93.7 92.7	81 86 83.3	44.3 58 57
ACW	89.7 93 93	85 89.7 89	61.7 60.7 62.7
ACW2	88 91.7 94	84 88 86.7	51.3 60.7 63
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	90.7 92 93.7	87.3 85.3 87	62 65 63.3
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	84.7 92.3 94.3	82.3 87.7 91.3	50 56.7 61

TABLE V
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR SPEECH INFLUENCED BY DIFFERENT CHANNELS (TIMIT DATA BASE). THE THREE SUCCESS RATES CORRESPOND TO CODEBOOK SIZES OF 16, 32, AND 64

Feature (with mean removal)	Test Condition		
	IRS channel	CMV channel	CPV channel
LP Cepstrum	59 63 68	51 56 63	48 49 56
ACW	60 71 73	56 62 66	61 61 62
ACW2	58 68 63	56 64 65	54 59 58
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	63 71 72	58 66 70	63 67 68
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	63 68 62	56 59 58	51 57 58

features show the best performance. However, the difference in performance among all the features (except the ARMA cepstrum) is very slight. The ARMA cepstrum definitely shows a much lower performance.

B. Testing on Noisy Speech

In this experiment, the test speech is degraded by different types of noise. First, consider additive white Gaussian noise (AWGN). Table II shows the results for various SNR values. As the SNR decreases, the mismatch between the training and test conditions becomes more glaring and the performance for all the features decreases. When the SNR is 30 dB, the ARMA cepstrum clearly shows the worst performance. The performance of the various other features is about the same with the ACW2 having a slight edge. For the lower SNR values, the disparity between the performance of the ARMA cepstrum and of the other features becomes less. The PFL1 features is the best for an SNR of 20 dB.

The test speech is now corrupted by colored noise that is generated by passing white Gaussian noise through a recursive linear predictive filter computed from a frame of speech corresponding to a sustained vowel. Table III shows the results for various SNR values. Due to the inferior performance of the ARMA cepstrum for clean speech and white noise, we do not find it necessary to consider it for the colored noise condition. Again, as the SNR decreases, the performance for all the features decreases. For an SNR of 30 dB, the performance of all the features is similar. For the lower SNR values, the PFL2 feature is the best particularly for a codebook size of 64.

Consider the case when the test speech is corrupted by babble noise. Table IV shows the results for various SNR values. Again, the ARMA cepstrum is not considered. For SNR values of 30 dB and 20 dB, all the features show a similar performance. When the SNR is 10 dB, the ACW and PFL1 features are the best for a small codebook size of 16. When the codebook size is 32, the PFL1 is the best feature. An increase in the codebook size to 64 shows a nearly equivalent performance among the ACW, ACW2, PFL1, and PFL2 features. The PFL1 is the generally preferred feature.

For speech degraded by any type of noise (that we consider) at a relatively high SNR of 30 dB, the features show a similar performance. As the SNR decreases, differences in performance among the features begin to emerge. The new features do as well or better than the conventional LP cepstrum. However, the best feature depends on the type of noise.

C. Testing on Speech Subjected to Channel Effects

In this section, we present the results for test speech subjected to different types of channel effects. When clean speech is influenced by a channel, an additive component manifests itself on the cepstrum of the clean speech. It has been shown that removing the mean of the cepstrum attempts to deemphasize this additive cepstral component and improves performance [4]. Since all the features we consider are cepstral type features, we show the results when mean removal is done. For the LP cepstrum, a better method of mean removal known as *pole filtered mean removal* has been recently proposed [22]. Note that we do not consider pole filtered mean removal in this paper.

For the TIMIT data base, the test speech is obtained by passing each utterance through three types of channels, namely, 1) the intermediate reference mask (IRS) channel, 2) the continental mid voice (CMV) channel [13], [14], and 3) the continental poor voice (CPV) channel [13], [14]. All three are representative of telephone channels. Table V depicts the results. The cepstral features based on the pole-zero transfer functions are almost always better than the conventional LP cepstrum. The improvement over the conventional LP cepstrum depends on the type of channel. For the CPV channel, the PFL1 feature is better than the LP cepstrum by a factor of at least 12% depending on the codebook size.

TABLE VI
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR THE SAN DIEGO PORTION OF THE KING DATA BASE. THE THREE SUCCESS RATES CORRESPOND TO CODEBOOK SIZES OF 16, 32, AND 64

Feature (with mean removal)	Test Condition	
	Within the great divide	Across the great divide
LP Cepstrum	73.1 71.6 74.6	49.6 50.8 48.1
ACW	72.6 75.0 76.5	52.3 58.9 59.2
ACW2	65.4 73.1 74.5	45.8 47.7 50.4
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	75.0 77.0 79.8	57.4 62.3 62.3
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	71.1 74.0 76.9	52.4 51.6 53.1

TABLE VII
IDENTIFICATION SUCCESS RATE AS A PERCENT FOR THE NUTLEY PORTION OF THE KING DATA BASE. THE THREE SUCCESS RATES CORRESPOND TO CODEBOOK SIZES OF 16, 32, AND 64

Feature (with mean removal)	Test Condition	
	Within the great divide	Across the great divide
LP Cepstrum	25 30 29.5	22.8 24.8 25.0
ACW	34 35 38.5	28.8 31.2 32.4
ACW2	31 31.5 35.5	23.6 24.8 25.6
Postfilter PFL1 $\alpha = 1$, $\beta = 0.9$	35.5 38 38	31.6 32.0 34.0
Postfilter PFL2 $\alpha = 1$, $\beta = 0.9$	30.5 32 33.5	26.8 28.0 28.4

Tables VI and VII depict the results for the San Diego and Nutley portions of the King data base, respectively. We first discuss the results in Table VI for the San Diego portion and relate them to two issues, namely, mean removal and frame selection based on LP poles. Energy thresholding is always performed. First, consider testing "within the great divide." Due to the relatively lower mismatch between the training and testing conditions, all of the features show a similar performance. However, the ACW and PFL1 features depict a slightly better performance. When frame selection based on LP poles is done, mean removal improves performance by 14% to 18% for all the features. An experiment was done to compare the performance of the conventional LP cepstrum with and without frame selection based on LP poles. When no mean removal is done, the improvement due to frame selection is 3% to 4% depending on the codebook size. With mean removal, the improvement due to frame selection is 3% to 8%. Frame selection does enhance robustness. In [7], a baseline performance (LP cepstrum without frame selection) was compared to the ACW feature in which frame selection was done. If we do the same comparison of the baseline performance with the features based on pole-zero transfer functions, a more glaring disparity is seen particularly with mean removal. Now, consider testing "across the great divide." For codebook sizes of 16 and 32, the ACW, PFL1, and PFL2 features are better than the LP cepstrum. Moreover, the PFL1 is clearly the best and the ACW is the second best. The superiority of the ACW and PFL1 features is maintained for a codebook size of 64. When frame selection is done, mean removal improves performance by 23 to 45% for all the features. With mean removal and no frame selection, the performance of the LP cepstrum is between 9% to 14% less than with frame selection. This again shows the enhancement of robustness due to frame selection. As in [7], a comparison of the LP cepstrum without frame selection to the other features with frame selection reveals a more glaring difference. Finally, note that we try to emulate a more practical scenario by using less training data than what is used in [18].

Now, consider the results in Table VII for the Nutley portion of the King data base. The identification success rates are consistently lower than for the San Diego portion since the Nutley portion is more noisy [18]–[20]. This disparity in the results for the two portions has also been recorded in [18]–[20]. The ACW and PFL1 features depict the best performance for both "within" and "across the great divide." When frame selection based on LP poles is done, mean removal improves performance by 3% to 9% for all the features.

VI. SUMMARY AND CONCLUSIONS

In this paper, various new cepstral features based on pole-zero transfer functions are examined with respect to robustness to noise and channel effects. The benchmark is the conventional LP cepstrum based on the all-pole LP transfer function. This all-pole function is converted in different ways into pole-zero transfer functions from which the cepstral feature is obtained. Two of the pole-zero functions, namely, the ACW and ACW2 are based on a partial fraction expansion of the LP all-pole function. A subsequent normalization of the residues is the key to enhancing robustness. The ACW spectrum emphasizes the formants. Another two pole-zero functions (PFL1 and PFL2) are based on the concept of a postfilter which was initially configured for speech enhancement. The PFL1 and PFL2 cepstra are equivalent to applying a weight to the conventional LP cepstrum. Like the ACW spectrum, the PFL1 spectrum emphasizes the formants. Another method of getting a pole-zero transfer function is to consider an ARMA analysis of speech.

Experiments are conducted using both the TIMIT and King data bases. A vector quantizer classifier is used. The performance under mismatched training and testing conditions is a good measure of robustness. There is some variation in the relative robustness of the features for different conditions. However, the ACW, PFL1, and PFL2 cepstrum perform as well as or better than the LP cepstrum for all the test conditions. For specific cases, the ACW and PFL1 cepstrum is clearly better than the LP cepstrum. These cases are:

1) speech corrupted by additive white Gaussian noise (SNR of 20 dB) with a codebook size of 16;
2) speech corrupted by babble noise (SNR of 10 dB) with a codebook size of 16;
3) speech influenced by the CPV channel;
4) when testing is done "across the great divide" for the San Diego portion of King (codebook sizes of 32 and 64);
5) for the Nutley portion of the King data base.

In view of this, the ACW cepstrum and the PFL1 cepstrum are the preferred features. Note that both the ACW spectrum and the PFL1 spectrum show similar characteristics in that the formants are emphasized and there is no spectral tilt. This implies that for robust speaker identification, the formants are extremely important. Moreover, an accurate representation of the entire spectral envelope either by LP analysis or by ARMA analysis is not the best way of providing robustness. The overall spectral envelope changes when speech is corrupted by

a channel and/or noise. However, the formants by themselves are more intact.

REFERENCES

[1] G. R. Doddington, "Speaker recognition—Identifying people by their voices," *Proc. IEEE*, vol. 73, pp. 1651–1664, Nov. 1985.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.

[3] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Englewood Cliffs, NJ: Prentice-Hall, 1978.

[4] 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.

[5] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1992.

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

[7] K. T. Assaleh and R. J. Mammone, "New LP-derived features for speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 630–638, Oct. 1994.

[8] M. S. Zilovic, R. P. Ramachandran, and R. J. Mammone, "A fast algorithm for finding the adaptive component weighted cepstrum for speaker recognition," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 84–86, Jan. 1997.

[9] V. Ramamoorthy, N. S. Jayant, R. V. Cox, and M. M. Sondhi, "Enhancement of ADPCM speech coding with backward adaptive algorithms for postfiltering and noise feedback," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 364–382, Feb. 1988.

[10] K. K. Paliwal, "On the performance of the frequency-weighted cepstral coefficients in vowel recognition," *Speech Commun.*, vol. 1, pp. 151–154, May 1982.

[11] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1414–1422, Oct. 1987.

[12] B.-H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass filtering in speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 947–954, July 1987.

[13] J. Kupin, "A wireline simulator (software)," CCR-P, Apr. 1993.

[14] D. J. Rahikka and R. A. Dean, "Secure voice transmission in an evolving communications environment," in *7th Ann. West. Conf. Expos.*, Anaheim, CA, Jan. 1986, pp. 1–16.

[15] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B.-H. Juang, "A vector quantization approach to speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Tampa, FL, Mar. 1985, pp. 11.4.1–11.4.4.

[16] A. E. Rosenberg and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes," *Comput. Speech Lang.*, vol. 22, pp. 143–157, 1987.

[17] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COMM-28, pp. 84–95, Jan. 1980.

[18] Y. Kao, J. S. Baras, and P. K. Rajasekaran, "Robustness study of free-text speaker identification and verification," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Minneapolis, MN, Apr. 1993, pp. II-379–II-382.

[19] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 639–643, Oct. 1994.

[20] Y. Kao, L. Netsch, and P. K. Rajasekaran, "Speaker recognition over telephone channels," in *Modern Methods of Speech Processing*, R. P. Ramachandran and R. J. Mammone, Eds. Boston, MA: Kluwer, Sept. 1995, pp. 299–321.

[21] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, "Speaker recognition using neural networks versus conventional classifiers," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 194–205, Jan. 1994.

[22] D. Naik, "Pole-filtered cepstral mean subtraction," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Detroit, MI, Apr. 1995.

Mihailo S. Zilovic was born in Belgrade, Yugoslavia, on July 26, 1961. He received the Dipl.Eng. degree from Belgrade University, Belgrade, Yugoslavia, in 1986, the M.E.E. degree from The City College of New York, in 1989, and the Ph.D. degree from the City University of New York in 1993.

From 1993 to 1995, he served as a Research Assistant Professor at the Computer Aids for Industrial Productivity Center, Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway. Since 1995, he has been with Bellcore (Bell Communication Research), Red Bank, NJ. His main research interests are in network performance analysis, speech processing, and multidimensional system theory.

Ravi P. Ramachandran (S'87–M'90) was born in Bangalore, India, on July 12, 1963. He received the B.Eng. degree (with great distinction) from Concordia University, Montreal, P.Q., Canada, in 1984, and the M.Eng. and Ph.D. degrees from McGill University, Montreal, in 1986 and 1990, respectively.

From January to June 1988, he was a Visiting Postgraduate Researcher, University of California, Santa Barbara. From October 1990 to December 1992, he worked in the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ. From January 1993 to August 1997, he was a Research Assistant Professor at the Computer Aids for Industrial Productivity Center, Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ. Also, from July 1996 to August 1997, he was a Senior Research Scientist at T-NETIX Inc., Piscataway. Since September 1997, he has been an Associate Professor in the Department of Electrical Engineering, Rowan University, Glassboro, NJ. His main research interests are in speech processing, data communications, and digital signal processing.

Richard J. Mammone (S'75–M'81–SM'86) is a Professor of electrical and computer engineering at Rutgers University, Piscataway, NJ, and a Principal Investigator of the University's Computer Aids for Industrial Productivity Center. He is also a founder of SpeakEZ, Inc., Piscataway, NJ, and chief Technical Advisor for T-NETIX, Inc., Englewood, CO. His research areas include speech processing and neural networks. He is a frequent consultant to industry and government agencies. He has published numerous articles and edited several books and special issues of international journals.

Dr. Mammone was the Senior Editor for Chapman & Hall, London, U.K., for neural networks. He is a founding member of the Technical Committee on Neural Networks for the IEEE Signal Processing Society. He has been a Guest Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has also been Associate Editor of *Pattern Recognition*, IEEE TRANSACTIONS ON NEURAL NETWORKS, and *IEEE Communications* magazine. He is listed in Marquis' *Who's Who in the World* and *Who's Who in Science and Engineering*. His speaker recognition technology was a finalist in the 1995 Computer World Smithsonian Award for developing new technologies for business and related services. He holds more than a dozen patents.