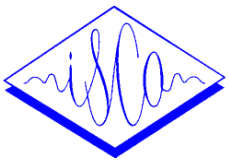

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ON THE USE OF FEATURES FROM PREDICTION RESIDUAL SIGNALS IN SPEAKER IDENTIFICATION

Jialong He, Li Liu, and Günther Palm
Department of neural information processing
University of Ulm, 89069 Ulm, Germany
email: {jialong, li, palm}@neuro.informatik.uni.ulm-de

ABSTRACT

A by-product of the LPC analysis is the generation of a prediction residual signal $e(n)$. $e(n)$ is usually ignored in the major applications of speech analysis, and only the LPC coefficients or parameters derived from the LPC coefficients are used to compose feature vectors. Since $e(n)$ carries all information that has not been captured by the LPC coefficients, an algorithm was proposed to calculate parameters from this prediction residual signal. The effectiveness of these features (named as *RCEP* coefficients) for speaker identification was evaluated. This approach yielded promising results. In an evaluation experiment in which the learning vector quantization (LVQ) networks served as classifiers, the correct identification rate for 112 male speakers was 88.8% for feature vectors composed of LPC based cepstrum (LPCC) alone, but reached 96.9% when the LPCC coefficients were combined with the *RCEP* coefficients.

I. INTRODUCTION

One of the most powerful speech analysis techniques is the method of linear predictive analysis. The importance of this method lies both in its ability to provide extremely accurate estimates of speech parameters and in its relative speed of computation. The philosophy of linear predictor is intimately related to the basic speech production model in which it has been shown that speech can be modeled as the output of a linear, time-varying system excited by either quasi-periodic pulses for voiced speech or random noise for unvoiced speech. Linear prediction parameters have been found useful in a variety of applications, such as speech coding, speech recognition, speech synthesis, and speaker recognition. A by-product of the LPC analysis is the generation of an error or residual signal $e(n)$. Theoretically if the all-pole model is perfect, the speech samples are predictable so that the residual signal $e(n)$ is very small. However, this simplified model is not suitable for nasal and fricative sounds, the detailed acoustic theory calls for both poles and zeros in the vocal tract transfer function. The prediction residual signal essentially carries all information that has not been captured by the LPC coefficients, e.g., phase, pitch information, zeros due to nasal sound, etc. In speech synthesis the residual signal is regarded as an ideal excitation of the all-pole model. In speech or speaker recognition the $e(n)$ is usually ignored, only the LPC parameters or some transformations of the LPC parameters (e.g., cepstrum or reflection coefficients) are used to compose feature vectors [1][2]. In this research we proposed an algorithm to calculate features from residual signals and applied these features in a text-independent speaker identification experiment. It was shown that features obtained from the residual signals using this method contain important information for speaker identification. Especially, the identification performance was improved significantly as features extracted from both original signals and residual signals were used jointly.

II. ALGORITHM

It has been shown that both LPC based cepstrum and FFT based cepstrum are dominant acoustic measurements of speech signals [3]. The LPC analysis approach performs spectral analysis on short segments of speech with an all-pole modeling constraint. It is faster and provides extremely accurate estimates of speech parameters if the signal can be well modeled by the all-pole model. On the other hand, FFT based cepstrum are derived based on the fact that the variations of a vocal track is much slower than the variations of an excitation signal so that they can be decoupled using homomorphic signal processing techniques [4]. The advantages of using FFT based parameters are their immunity to noise and easy to wrap frequency to non-uniform (bark or mel) scale. Here we calculated *RCEP* coefficients using FFT based method in the hope that the *RCEP* coefficients could capture extra information that are not contained in the LPC coefficients. The procedure of deriving the *RCEP* coefficients is briefly described as following.

1. Calculating LPC coefficients

Due to the time-varying nature of speech, the LPC analysis must be done on short segments of a speech signal. In our case, the Hamming window was used to do the short-time analysis and the window size was 32

msec long. A linear predictor with prediction coefficients α_k is defined as a system whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \quad (1)$$

where p is the LPC analysis order and $s(n)$ are samples of a speech signal. There are many methods to calculate the LPC coefficients. We employed the autocorrelation method so that the α_k could be calculated efficiently by the Durbin's recursive algorithm [4]. We also transformed the 16 LPC coefficients obtained from each frame of signals to 16 cepstral coefficients (LPCC) using the following relation for later use.

$$c_k = \alpha_k + \sum_{n=1}^{k-1} (n/k) c_n \alpha_{k-n} \qquad 1 \le k \le p \quad (2)$$

2. Obtaining the prediction residual signal
A prediction residual signal of LPC analysis is defined as

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \quad (3)$$

If the actual speech signal is generated by a system that is well modeled by a time-varying linear predictor of order p, then $e(n)$ is equally a good approximation to the excitation source. However, this is not the case for nasal and fricative sounds.

3. converting $e(n)$ to an one-sided autocorrelation sequence

$$q(k) = \sum_{n=1}^{N-k} e(n) e(n+k) \qquad 0 \le k \le N-1 \quad (4)$$

where N is the window size. The motivation that we transform the prediction residual signal to its autocorrelation sequence follows the work by Hernando *et. al* [5]. It has also been shown that an unwindowed autocorrelation operation on the impulse response of an all-pole system does not alter its pole structure, and the estimation of system parameters may be more reliably accomplished from the autocorrelation function since the signal-to-noise ratio (SNR) is enhanced [6]. Our pioneer experiment also confirmed that this transformation can enhance the effectiveness of features extracted from the prediction residual signal.

4. Calculating FFT based cepstrum from $q(k)$
We calculated FFT based cepstrum from $q(k)$ and defined them as *RCEP* coefficients. The procedure was similar to that of deriving a set of mel-scaled FFT based cepstrum (MFCC) [3]. The only difference is that we used $q(k)$, the one-sided autocorrelation sequence of a prediction residual signal, instead of the original speech signal $s(n)$ as the input. The major operations are: (a) appending sufficient zeros to the $q(k)$ to increase frequency resolution; (b) obtaining the magnitude spectrum of $q(k)$ with FFT; (c) forming 40

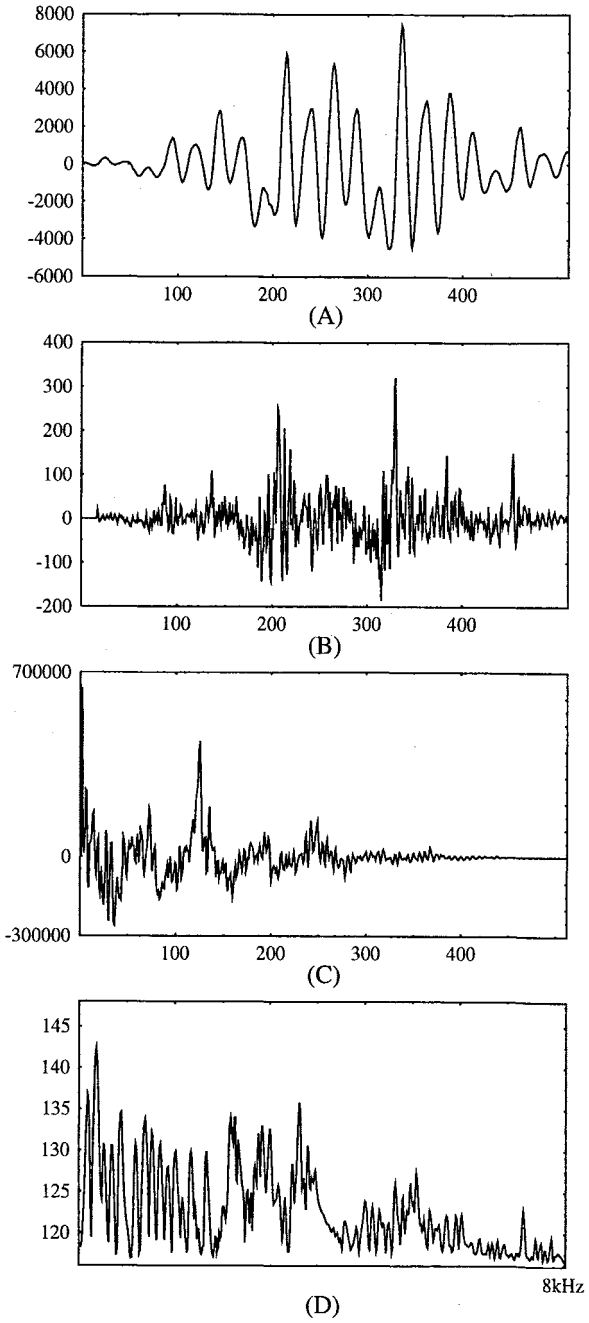

Figure 1. illustration of typical waveforms at several points of calculating *RCEP* coefficients. (A) a segment of speech weighted by a Hamming window; (B) prediction residual (error) signal; (C) one-sided autocorrelation sequence of the prediction residual signal; (D) log magnitude (in dB) of the autocorrelation sequence.

filter banks in mel-scale; (d) computing the log magnitude spectrum of $q(k)$; (e) calculating 16 *RCEP* coefficients R_1-R_{16}, normalized by R_0, with inverse DFT

$$R_k = \sum_{i=1}^{40} X_i \cos[k(i-0.5)\pi/40] \qquad 0 \le k \le 16 \quad (5)$$

where X_i is the log magnitude spectrum.

As an example, Figure 1 shows a segment of voiced speech signal and illustrates some typical waveforms obtained at several points in the analysis. For voiced speech it is expected that the prediction residual signal will be large at the beginning of each pitch period. Thus the autocorrelation of $e(n)$ is used to estimate pitch period by detecting the largest peak in the appropriate range [4]. This is also manifested in Fig. 1(c), a pitch period of about 8 msec is clearly in evidence. Since the pitch period is found to be an effective feature for speaker identification [13] we also calculated a pitch period from each frame of signals as a feature.

III. SPEECH DATABASE

The evaluation speech data were selected from the TIMIT database [7]. This database has been designed to provide speech data for the development and evaluation of automatic speech recognition systems. It contains speech data from 630 speakers, each speaking ten phonetically rich sentences. The average length of the sentences is 6.2 seconds. Since the variances between male and female voices are fairly large, it is more reasonable to evaluate a speaker identification system with data from the same gender. Hence, we chose the speech data from 112 male speakers. Eight out of ten sentences spoken by each speaker were randomly selected as training data and the other two as test ones, that is, there are 896 training and 224 test sentences. It is generally agreed that features extracted from the voiced parts of speech signals, especially from vowels, nasals, and fricatives, are more effective for speaker recognition [8]. Therefore, we implemented an automatic procedure to locate all voiced segments in speech signals and calculated feature vectors from only these segments. There are 28556 training vectors from 896 training sentences and 7442 test vectors from 224 test sentences.

IV. RESULTS

We conducted a speaker identification experiment to evaluate the effectiveness of the *RCEP* coefficients. The identification system was adapted from our on-line system [9]. The classifiers were based on the learning vector quantization (LVQ) algorithms proposed by Kohonen [10]. In our experiment we employed the LVQ3 algorithm since it provides a better performance and is self-stabilizing. The initial placements of the code vectors were decided by the LBG vector quantization algorithm [11].

The evaluation results with 16 code vectors per speaker are summarized in Table 1. Note that there are two kinds of identification rate, i.e., *frame identification rate* and *sentence identification rate*. The frame identification rate is the percentage of feature vectors that are correctly classified. It represents an identification performance on the bases of short segments of speech signals. In our case, each short segment was 32 ms long. Accordingly, the sentence identification rate is the percentage of sentences that are correctly classified. A frame-level classification for a test vector was made by nearest-neighbor comparison of the test vector with all code vectors, and the identification decision for a sentence was determined by the majority voting of all vectors from the sentence. It is seen that with the *RCEP* coefficients alone the identification performance at sentence level reached 42.4% (95/224). This result demonstrates that prediction residual signals still contain useful information for identifying speakers. Since the *RCEP* coefficients are calculated from prediction residual signals, it is expected that they may carry extra information that is not captured by the LPC coefficients and should be complementary to the LPCC coefficients. Based on this reasoning, we combined the *RCEP* coefficients with the LPCC coefficients to form a long feature vector. It is seen that the identification performance is significantly improved at both frame and sentence level with combined feature vectors.

Parameter	RCEP	LPCC	RCEP+ LPCC	RCEP+ LPCC+ Pitch
dimension	16	16	32	33
frame (%)	7.45	25.82	33.93	36.16
sentence (%)	42.4	88.8	96.9	97.3

Table 1. Identification rates at frame/sentence level for 112 male speakers using LVQ networks. The identification decision for a sentence is based on the majority voting of all vectors from the sentence. The codebook size is 16 code vectors per speaker.

Soong and Rosenberg found [12] that the LPCC and the dynamic LPCC (also known as delta cepstrum) coefficients are fairly uncorrelated and can be used jointly to improve the performance of a *text-dependent* speaker recognition system. Later the delta cepstrum, in conjunction with the LPCC, are widely used in speech

recognition systems [3]. With the analysis method proposed in this study, features obtained from the prediction residual signals were found to be complementary to the features extracted from the original speech signals. So that they could be used jointly to improve the overall performance of a *text-independent* speaker identification system. Similar to that of combining the LPCC coefficients with the delta cepstrum, where an appropriate normalization is necessary [12], the LPCC and *RCEP* coefficients also need to be scaled to the same range before their combination since they are calculated based on different metric systems. In our experiment the LPCC and the *RCEP* coefficients were concatenated to compose the feature vector , the scaling factor k was defined as

$$(c_1, c_2, c_3 \ldots c_{16}, \beta R_1, \beta R_2, \beta R_3, \ldots, \beta R_{16}) \qquad \beta = k / R_0 \tag{6}$$

Figure 2 shows the frame level classification rates as a function of k. It is seen that the classification rates are improved at first with k, reach the maximum when k = 8, and then decrease.

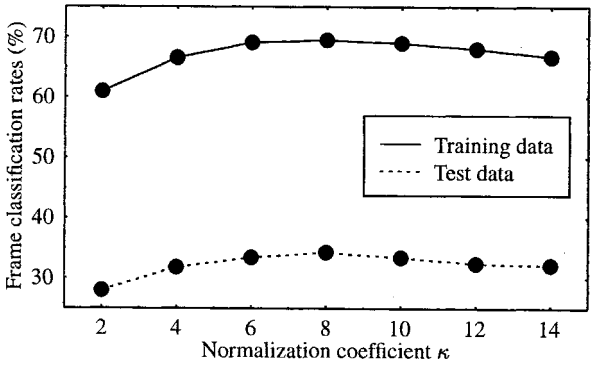

Figure 2 Classification rates at frame level for 112 speakers as a function of scaling factor k.

In summary, it was shown that the *RCEP* coefficients carry extra information not captured by the LPCC coefficients, so that they can be combined together to improve the performance of a speaker identification system. Besides, we also found from our preliminary experiments that the *RCEP* coefficients are also useful for speech recognition.

ACKNOWLEDGMENT

This work is partially sponsored by the state of Baden-Württemberg, Germany (Landesschwerpunkt Neuro-informatik).

REFERENCES

[1] S. B. David and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech, Signal Processing, ASSP-28, pp. 357-366, 1980.

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

[3] J. W. Picone, "Signal modeling techniques in speech recognition," Proc. IEEE, vol. 81, pp. 1215-1247, 1993.

[4] L. R. Rabiner and R. W. Schafer, "Digital processing of speech signals," Englewood Cliffs, NJ: Prentice-Hall, 1978.

[5] J. Hernando, C. Nadeu, C. Villagrasa and E. Monte, "Speaker identification in noisy condition using linear prediction of the one-sided autocorrelation sequence," Proc. ICSLP-94, pp. 1847-1850, Sept. 1994, Yokohama, Japan.

[6] D. Mansour and B. H. Juang, "The short-time modified coherence representation and noisy speech recognition," IEEE Trans. Acoustics, Speech Signal Proc. ASSP-37, pp. 795-804, 1989.

[7] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech communication, Vol. 9, pp. 351-356, 1990.

[8] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," J. Acoust. Soc. Am. Vol. 51, pp. 2044-2056, 1972.

[9] J. He, L. Liu and G. Palm, "A text-independent speaker identification system based on neural networks," Proc. ICSLP-94, pp. 1851-1854, Sept. 1994, Yokohama, Japan.

[10] T. Kohonen, "The self-organizing map," Proc. IEEE, vol. 78, pp. 1464-1480, 1990.

[11] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Comm., Vol. 20, pp. 84-95, 1980.

[12] 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, ASSP-36, pp. 871-879, 1988.

[13] B. S. Atal, "Automatic speaker recognition based on pitch contours," J. Acoust. Soc. Am. 52, pp. 1687-1697, 1972.