ON THE USE OF NORMALIZED LPC ERROR TOWARDS BETTER LARGE VOCABULARY SPEECH RECOGNITION SYSTEMS

Rathinavelu Chengalvarayan

Speech Processing Group, Bell Labs
Lucent Technologies, Naperville, IL 60566, USA
Email: rathi@lucent.com

ABSTRACT

Linear prediction (LP) analysis is widely used in speech recognition for representing the short time spectral envelope information of speech. The predictive residues are usually ignored in LP analysis based speech recognition system. In this study, the normalized residual error based on LP is introduced and the performance of the recognizer has been further improved by the addition of this new feature along with its first and second order derivative parameters. The convergence property of the training procedure based on the minimum classification error (MCE) approach is investigated, and experimental results on *city name* recognition task demonstrated a 8% string error rate reduction by using the extended feature set as compared to conventional feature set.

1. INTRODUCTION

Speech recognizers have traditionally utilized cepstral parameters derived from linear predictive (LP) analysis due to its ability to provide a reasonable source-tract separation. This analysis assumes the speech signal to follow an all-pole model and the importance of this method lies in its relative speed of computation. A by-product of the LP analysis is the generation of an error or residual signal. If the all-pole model is perfect then the speech samples are predictable so that the residual signal is very small. The prediction residual signal essentially carries all information that has not been captured by the LP coefficients. Recent study shows that the LP residual can be exploited for enhancement of speech in the presence of additive noise [9]. In speech recognition the LP residual is usually ignored and only the LP cepstral coefficients are used as a basic feature set [1, 6].

Combining the LP ananlysis and residual analysis could potentially produce improved speech features. This can be done in several ways. For example, one could compute the cepstrum of the LP residual and then append this feature with the traditional LP cepstral features. Residual cepstrum is currently being used in speaker verification [8] and speaker identification [7] applications with great success. Another approach is to compute the LP cepstral coefficients in a conventional way and augment the normalized LPC error as an additional feature into the existing cepstral feature set. In this study both the LP cepstrum and normalized LP error have been utilized.

We restrict our presentation to only the recognizer based on hidden Markov model (HMM) approach using continuous density mixtures to characterize the states of the HMM. We call the HMM using the new feature set as HMM-II and the the model using the cepstral coefficients alone is represented by HMM-I. In this work we describe an algorithm to calculate features from residual signals and apply these features in a speaker independent continuous speech recognition experiment. We show that features obtained from the residual signals using this method contain important information for speech discrimination and can be exploited for telephone based speech recognition tasks.

2. COMPUTATION OF NORMALIZED LPC ERROR

The LP analysis converts each frame of $p+1$ autocorrelations into p LP coefficients as given in the following Durbin's recursive algorithm:

$$
\begin{aligned}
E^{(0)} &= r(0) \\
k_i &= \frac{r(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} r(|i-j|)}{E^{(i-1)}} \qquad 1 \le i \le p \\
\alpha_i^{(i)} &= k_i \\
\alpha_j^{(i)} &= \alpha_j^{(i-1)} - k_i \alpha_{i-j}^{(i-1)} \qquad 1 \le j \le i-1
\end{aligned}
$$

$$E^{(i)} = (1 - k_i^2)E^{(i-1)}$$

where the summation in the second equation is omitted for $i = 1$ and $r(i)$ is the autocorrelation coefficient of lag i. The five set of linear equations are solved iteratively for $i = 1, 2, \cdots, p$ and the final solution is given as

$$\begin{aligned} \alpha_j^{(p)} &= LPC\ coefficients \quad 1 \le j \le p \\ k_j &= PARCOR\ coefficients \end{aligned}$$

Note that the quatity $E^{(i)}$ is the prediction error for a predictor of order i. Thus at each stage of the computation the prediction error for a predictor of order i can be monitored. Also, if the autocorrelation coefficients $r(i)$ are replaced by a set of normalized autocorrelation coefficients, $R(k) = r(k)/r(0)$, the Durbin's recursive solution remains unchanged. However, the error $E^{(i)}$ is now interpreted as a normalized error and is given by:

$$V^{(i)} = \frac{E^{(i)}}{r(0)} = 1 - \sum_{k=1}^{i} \alpha_k R(k) \quad with$$
$$0 < V^{(i)} \le 1 \quad 0 \le i$$

The normalized error for $i = p$ can also be written in the form

$$V^{(p)} = \prod_{i=1}^{p} (1 - k_i^2)$$

where the quantities k_i are the PARCOR coefficients and are in the range $-1 \le k_i \le 1$ therby guranteeing the stability of the LPC analyzer. The parameter $V^{(p)}$ can vary from 0 to 1. Thus the normalized LPC prediction error is extracted for every frame of speech.

To illustrate the nature of the error signal Figure 1 shows the actual speech waveform and the corresponding normalized prediction error for the word "Chicago Illinois" spoken by a female speaker. It is observed that the normalized prediction error steadily decreases for voiced sections of speech. For unvoiced sections of speech the error starts increasing significantly higher than for voiced portion of speech. This is expected since the all-pole model for unvoiced speech is nowhere near as accurate as it is for voiced speech.

3. FEATURE ANALYSIS

The speech input is sampled at 8kHz and preemphasized using a first-order filter with a coefficient

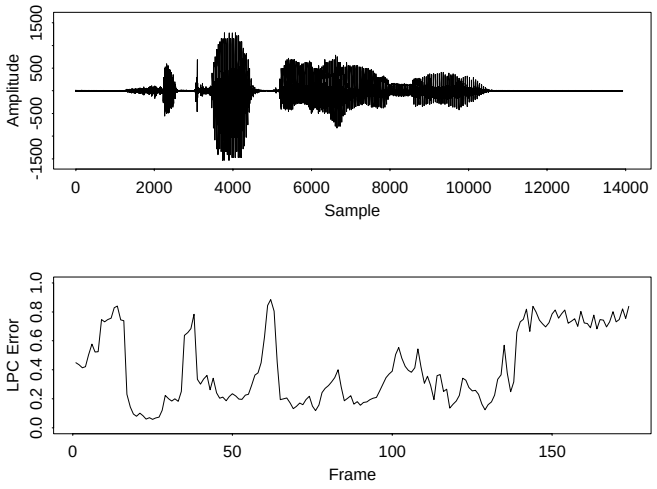

Figure 1. Speech from the word "Chicago Illinois" spoken by a female. Top plot shows the original speech and bottom plot shows the normalized LPC error.

of 0.97. The samples are blocked into overlapping frames of 30 msec in duration, where the overlap is set to 20 msec. Each frame is windowed with a Hamming window and then processed using a 10th-order LPC analyzer. The LPC coefficients are then converted to cepstral coefficients, where only the first 12 coefficients are retained. The baseline recognizer feature set consists of 36 features that includes the 12 liftered cepstral coeeficients and their first and second order derivatives [3]. The extended feature vector used in this study has 39 parameters, including the baseline 36 features, normalized lpc error parameter and their first and second order derivatives. Since the signal has been recorded under various telephone conditions and with different transducer equipment, each cepstral vector was further processed using the hierarchical signal bias removal (HSBR) method in order to reduce the effect of channel distortion [2].

4. SPEECH DATABASE

The experimental results are based on a continuous speech database containing speech utterences recorded over the telephone network in a U.S. wide data collection covering the different dialect regions. Male and female speakers were fairly equally represented. The training set consists of 9865 generic phrases and the testing set contains 3620 spontaneous utterences of *city name* followed by either a state or a country name, for example,

Seattle Washington. The large vocabulary continuous speech recognition task involves speaker independent *city name* recognition where the recognizer lexicon consists of 448 entries plus silence with one lexicon entry per word. A word insertion penalty was used which is the same for all speakers. The grammar used in the recognition is the standard word pair grammar.

5. HMM RECOGNIZER

The subword model set used in the recognition consists of 41 context independent units [4]. Each subword is modeled by a three state left-to-right continuous density HMM with only *self* and *forward* transitions. A mixture of Gaussians with diagonal covariances is employed to estimate the density function for each state. A maximum of 16 mixtures per state is allowed. The silence/background is modeled with a single state, 32 Gaussian mixture HMM. Furthermore no transition probabilities are used. The lexical representations of the sentences are obtained by preprocessing the sentence orthographic transcriptions through a text-to-speech front end. The initial model set of these 41 subword units was trained using the conventional maximum likelihood training procedure [5]. We then applied five iterations of integrated HSBR with MCE training to the initial boot model with null grammar and the number of competing string models as well as the size of HSBR codebook were set to four [2]. Make a note that the HSBR codebook is extracted form the mean vectors of HMMs coresponding to the 12 cepstral coefficients and each training utterence is signal conditioned by applying HSBR, prior to being used in MCE training and decoding.

6. RECOGNITION RESULTS

We have conducted experiments to verify the effectiveness of the proposed new feature set, using the continuous speech database, on the convergence property of the MCE training procedure and on subword recognition performance. In Figure 2 we show empirical results on the behavior of the MCE training procedure for the *city name* continuous speech recognition task. The upper graph of Figure 2 shows the string error rates as a function of the epoch (a complete pass through the entire training data set is called an epoch) of the MCE training algorithm for the testing data. The solid lines are associated with MCE-trained conventional HMM (HMM-I), and the dotted lines with HMM generated using extended feature set (HMM-II). The lower graph of Figure 2 shows the average string loss for the entire training data set as a function of the training epoch. We observed that the recognition error rate monotonically decreases with the training epoch, and the average string loss monotonically decreases, both reaching their respective asymptotic values after five epoches of the training. The average loss decreases faster for the HMM-II than for the HMM-I, indicating the effectiveness of the newly introduced feature parameter. Similar characteristics in the recognition performance are also observed. This indicates that the original objective set out for minimizing the recognition error via the MCE training is accomplished and that the MCE training may be more effective for the HMM-II than the HMM-I.

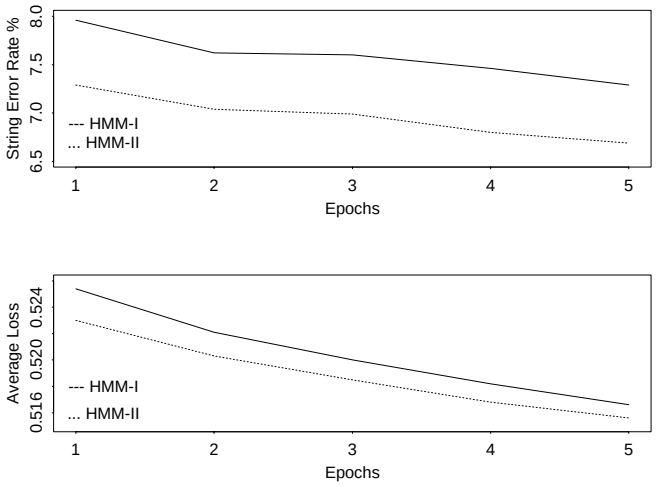

Figure 2. Convergence characteristics of the MCE training procedure. Top graph shows the word error rate for the "city name" recognition task and bottom graph shows the average string loss as a function of the training epoch.

Type of Model	ML Method	MCE Method
HMM-I	12.24%	7.29%
HMM-II	10.75%	6.69%

Table 1. Word error rate for large vocabulary "city name" recognition task using the conventional ML (left) and MCE (right) training methods.

The *city name* speech recognition results focusing on the comparative performances of the ML and MCE-trained HMM-II versus the HMM-I are

summarized in Table 1. The results shown in Table 1 can be elaborated as follows. First, under all conditions the MCE training is superior to the ML training; the MCE-based recognizer achieves an average of 35% string error rate reduction, uniformly across all types of speech models. Second, for the ML-based recognizer, the HMM-II gives about 12% string error rate reduction compared with HMM-I (error rate went down from 12.24% to 10.75%). Thirdly, for the MCE-based recognizer, the HMM-II produces 6.69% string error rate which further yields about 8% reduction in string error rate compared with the HMM-I. Finally, we noticed that the HMM-II outperforms the HMM-I by atleast 8% in error rate reduction for all cases. The results presented in Table 1 demonstrate the efficacy of extended feature set models trained by MCE for *city name* continuous speech recognition.

7. CONCLUSIONS

In this study, the normalized residual error based on LP is introduced and the performance of the recognizer has been further improved by the addition of this new feature along with its first and second order derivative parameters. We also showed how the generation of such features can be obtained as a by-product of LP analysis. A new HMM that integrates the LP cepstrum and normalized LP error is implemented and evaluated using ML and MCE training methods. The convergence property of the training procedure based on the MCE approach is investigated, which leads us to believe that the objective of minimizing the string error intended with the MCE criterion is achieved more effectively for the HMM-II than for the HMM-I.

The experimental results on *city name* recognition task yields a 8% string error rate reduction by using the HMM-II as compared to HMM-I. This shows that the normalized LP error do contain useful information and is complementary to the LP cepstral coefficients. We believe that it is also important not to ignore the last output of an LP analysis, namely the normalized prediction error factor. So far, the investigation presented in this paper and the performed experiments confirm the expected significance of the residue for automatic speech recognition.

Acknowledgements

The author would like to thank Dr. Rafid Sukkar for his ideas and support in the early stages of this work.

REFERENCES

[1] W. Chou, C. H. Lee, B. H. Juang and F. K. Soong, "A minimum error rate pattern recognition approach to speech recognition", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 8, No. 1, pp. 5-31, 1994.

[2] M. Rahim, B. H. Juang, W. Chou and E. Buhrke, "Signal conditioning techniques for robust speech recognition", *IEEE Signal Processing Letters*, Vol. 3, No. 4, pp. 107-109, 1996.

[3] C.H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini and A. E. Rosenberg, "Improved acoustic modeling for large vocabulary continuous speech recognition", *Computer Speech and Language*, Vol. 6, No.2, pp. 103-127, 1992.

[4] R. A. Sukkar and C. H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition", *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 6, pp. 420-429, 1996.

[5] B. H. Juang and L. Rabiner, "The segmental K-means algorithm for estimating parameters of hidden Markov models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 9, pp. 1639-1641, 1990.

[6] S. Young, "A review of large-vocabulary continuous-speech recognition", *Signal Processing Magazine*, Vol. 13, No. 5, pp. 45-57, 1996.

[7] J. He, L. Liu and G. Palm, "On the use of features from residual signals in speaker identification", *Proc. EUROSPEECH*, pp. 313-316, 1995.

[8] P. Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification", *Speech Communication*, Vol. 17, pp. 145-157, 1995.

[9] B. Yegnanarayana, C. Avendano, H. Hermansky and P. S. Murthy, "Processing linear prediction residual for speech enhancement", *Proc. EUROSPEECH*, pp.1399-1402, 1997.