# A LINEAR PREDICTIVE FRONT-END PROCESSOR FOR SPEECH RECOGNITION IN NOISY ENVIRONMENTS

*Yariv Ephraim, Jay G. Wilpon and Lawrence R. Rabiner*

AT&T Bell Laboratories
Speech Research Department
Murray Hill, NJ 07974.

## Abstract

We investigate the performance of a recent algorithm for linear predictive (LP) modeling of speech signals, which have been degraded by uncorrelated additive noise, as a front-end processor in a speech recognition system. The system is speaker dependent, and recognizes isolated words, based on dynamic time warping principles.

The LP model for the clean speech is estimated through appropriate composite modeling of the noisy speech. This is done by minimizing the Itakura-Saito distortion measure between the sample spectrum of the noisy speech and the power spectral density of the composite model. This approach results in a "filtering-modeling" scheme in which the filter for the noisy speech, and the LP model for the clean speech, are alternatively optimized.

The proposed system was tested using the 26 word English alphabet, the ten English digits, and the three command words, "stop," "error," and "repeat," which were contaminated by additive white noise at 5-20 dB signal to noise ratios (SNR's). By replacing the standard LP analysis with the proposed algorithm, during training on the clean speech and testing on the noisy speech, we achieve an improvement in recognition accuracy equivalent to an increase in input SNR of approximately 10 dB.

## 1. Introduction

The problem of speech recognition in noisy environments has recently attracted the attention of many speech researchers. The reason is that existing speech recognition systems, which perform reasonably well in a clean or laboratory environment, fail under conditions in which high level noise is present at the recognizer input. The noise can be picked up at the source location and/or in the channel which connects the speaker and the recognizer. The first case may be more difficult to handle, since in the presence of high ambient noise the utterances to be recognized are pronounced differently than in a clean environment in which the recognizer is usually trained. This condition leads to a further mismatch between the input utterance and the corresponding stored reference pattern. Typical noise sources of interest include fans in an office environment, traffic in mobile radio communication, engine noise in aircraft communication, and channel noise over the long distance switched telephone network.

In general there are three main approaches to speech recognition in noisy environments. The first is to make existing speech recognition systems, which have proved to perform successfully in a laboratory environment, immune to noise. The second is to design speech recognition systems which are inherently robust to input interference. The third, is to train the recognizer in an environment similar to that of testing. The first approach is accomplished by taking into account the noise presense in estimating the feature vector. The second approach is far less understood and will probably require using different feature vector and/or distortion measure then those commonly used. The third approach has the obvious disadvantage that it requires training in an environment similar to that of the test.

The work presented in this paper belongs to the first category of speech recognition approaches. We use an LPC-based, dynamic time warping, speaker dependent, isolated word, speech recognition system [1], and examine the performance of a recent algorithm [2] for autoregressive (AR) modeling of the original speech given noisy speech. This algorithm was derived using minimal assumptions about the source and the noise statistics and therefore is attractive in speech recognition applications. Specifically, we assume that the noise is additive and uncorrelated with the source, and that the noisy source is quasi-stationary. We, however, do not require exact knowledge of the probability distribution (PD) of either the source or the noise, nor that the original speech has the structure of the AR model. The modeling is entirely based on the sample spectrum of the noisy source, similar to the way that LP modeling of clean sources only uses the sample spectrum of the original source. Note that since the noise is assumed to be additive and uncorrelated with the source, our work applies better to channel noise rather than to source noise, since the pronunciation effect mentioned above is ignored here.

The proposed algorithm estimates the AR model of the original source through an appropriate composite modeling of the noisy source. The composite model consists of an AR model for the source and an additive parametric model (e.g., moving average (MA)) for the noise. The modeling is achieved by minimizing the Itakura-Saito distortion measure between the sample spectrum of the noisy source and the power spectral density of the composite model, *over all parameters* of the source and the noise models. This results in a "filtering-modeling" scheme in which the filter for the noisy speech, and the AR model for the clean speech, are alternatively optimized [2]. The Itakura-Saito distortion measure was chosen since it is an information theoretic distortion measure for sources which are not strictly stationary and autoregressive moving average (ARMA) composite models which are of interest in this work [3].

The above algorithm was used to recognize spoken versions of the 26 word English alphabet, the ten English digits, and the three command words, "stop," error," and "repeat," which were contaminated by additive white noise. The alphabet vocabulary was chosen since it contains highly confusable sets of words which make the recognition task non-trivial and therefore sensitive to noise. White noise was chosen, since it is believed to be the most harmful noise for recognition, as it equally attacks all the frequency components of the speech. The variance of the noise was adjusted



1324 CH2396-0/87/0000-1324 $1.00 

to a fixed value across each word. This results in words which are degraded by uncorrelated stationary noise with some overall SNR. Noisy speech with overall SNR of 5, 10, 15, and 20 dB was examined.

Three cases, which differ in the way the AR models of the speech were estimated during training and testing, to produce templates and test patterns, respectively, are considered. In all cases the templates were generated from clean speech and testing was done on noisy speech. In the first case, standard LP analysis (e.g., the Levinson algorithm [4]) was used in generating both the templates and the test patterns. In the second case, standard LP analysis was used to generate the templates, and the proposed algorithm was used in generating the test patterns. In the third case, the proposed algorithm was exclusively used in generating both the templates and the test patterns. The first strategy corresponds to ignoring the noise presense in the input signal and applying the recognizer as if the input speech is clean. This strategy results in the worst recognition accuracy. Comparing with this case, the second and the third strategies provide improvement in recognition accuracy equivalent to an increase of about 5 dB and 10 dB, respectively, in input SNR.

## 2. AR modeling algorithm

Let $Y=X+V$, where $X$, $V$, and $Y$ denote, respectively, the $K$-dimensional random vectors of the source, the noise, and the noisy source. Let $Y_\theta$ be the Fourier transform of $Y$. Let $\sigma^2/|A_\theta|^2$ denote the power spectral density of an AR model for the source. $A_\theta$ is the Fourier transform of the sequence $\{1,a_1,a_2,\cdots,a_p\}$ which, together with $\sigma$, constitute the parameters of the AR model. Let $f_\theta$ be the power spectral density of the noise model. $f_\theta$ is assumed to be dependent on a finite number of parameters $(f_0,f_1,\cdots,f_q)$. For the case of white noise, which is of interest here, $f_\theta=\lambda$, a constant. Assume that the sample spectrum of the noisy source, $|Y_\theta|^2$, is strictly positive on $0\le\theta\le 2\pi$. The modeling problem, as formulated in [2], is that of finding $\sigma^2/|A_\theta|^2$ and $f_\theta$ (or equivalently their respective parameters) which minimize the Itakura-Saito distortion measure between the sample spectrum of the noisy source $|Y_\theta|^2$ and the power spectral density of the composite model $\sigma^2/|A_\theta|^2+f_\theta$. This distortion measure is given by

$$d(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2+f_\theta) = \int_0^{2\pi}\left[\frac{|Y_\theta|^2}{\sigma^2/|A_\theta|^2+f_\theta}-\ln\frac{|Y_\theta|^2}{\sigma^2/|A_\theta|^2+f_\theta}-1\right]\frac{d\theta}{2\pi}, \quad (1)$$

and it is proven in [2] that it achieves its minimum by some composite model.

The modeling is performed as follows. We first define a filter

$$|H_\theta|^2 \triangleq \frac{\sigma_h^2/|A_\theta^h|^2}{\sigma_h^2/|A_\theta^h|^2+f_\theta}, \quad (2)$$

where $\sigma_h^2/|A_\theta^h|^2$ is defined similarly to $\sigma^2/|A_\theta|^2$ but the two power spectral densities are not necessarily the same, and rewrite (1) as

$$d(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2+f_\theta) = d(|Y_\theta|^2|H_\theta|^2,\ \sigma^2/|A_\theta|^2)\,|_{\sigma_h^2/|A_\theta^h|^2=\sigma^2/|A_\theta|^2}. \quad (3)$$

Now we focus on $d(|Y_\theta|^2|H_\theta|^2,\ \sigma^2/|A_\theta|^2)$ and alternatively minimize it once over all filters of the class (2) assuming that the AR model is given, and then over all AR models assuming the filter is given. The minimizing filter obtained in this way satisfies the constraint in (3) [2]. Each of these two phases reduces, or at least does not increase, the value of the distortion measure and thus a descent algorithm results.

When the AR model $\sigma^2/|A_\theta|^2$ is given, the filter which minimizes $d(|Y_\theta|^2|H_\theta|^2,\ \sigma^2/|A_\theta|^2)$, called the minimum distortion filter, is given by [2]

$$\nu_1(\sigma^2/|A_\theta|^2) = \frac{\sigma^2/|A_\theta|^2}{\sigma^2/|A_\theta|^2+f_\theta^*} \quad (4)$$

where $f_\theta^*$ satisfies

$$d(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2+f_\theta^*) = \inf_{f_\theta} d(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2+f_\theta). \quad (5)$$

Such a filter is guaranteed to exist for MA, AR and ARMA noise models; however, it might not be unique. In this case we apply an arbitrary selection rule to choose one filter from all possible minimum distortion filters and consider a "point-to-set" mapping, $\nu(\sigma^2/|A_\theta|^2)$: $\sigma^2/|A_\theta|^2 \to \{\nu_1(\sigma^2/|A_\theta|^2)\}$, from the given AR model to the set of minimum distortion filters. When the filter $|H_\theta|^2$ is given, the AR model which minimizes $d(|Y_\theta|^2|H_\theta|^2,\ \sigma^2/|A_\theta|^2)$ is the unique stable linear predictive model for the linearly filtered noisy source $|Y_\theta|^2|H_\theta|^2$. Let this model be denoted by $\mu(|H_\theta|^2)$. Combining the above two steps we have the algorithm $T_Y$: $(\sigma^2/|A_\theta|^2) \to \mu(\nu(\sigma^2/|A_\theta|^2))$ which generates a sequence of AR models when starting from some initially given AR model.

Let

$$\rho(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2) \triangleq d(|Y_\theta|^2\nu_1(\sigma^2/|A_\theta|^2),\ \sigma^2/|A_\theta|^2) = d(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2+f_\theta^*) \quad (6)$$

be the distortion which is associated with each generated AR model and the corresponding optimal filter for that model. Then, the formal statement of the algorithm is as follows.

*The fixed point algorithm*

(0) Initialization: Given a sample spectrum $\{|Y_\theta|^2>0,\ 0\le\theta\le 2\pi\}$, an initial AR model $(\sigma^2/|A_\theta|^2)_0$, and a threshold $\varepsilon>0$, calculate $\rho(|Y_\theta|^2,\ (\sigma^2/|A_\theta|^2)_0)$ and set $m=1$.

(1) Given $(\sigma^2/|A_\theta|^2)_{m-1}$, calculate

$$(\sigma^2/|A_\theta|^2)_m = T_{Y_1}((\sigma^2/|A_\theta|^2)_{m-1}) = T_{Y_1}{}^m((\sigma^2/|A_\theta|^2)_0)$$

where, $T_{Y_1}(.)\in T_Y(.)$.

(2) Compute $\rho(|Y_\theta|^2,\ (\sigma^2/|A_\theta|^2)_m)$.

(3) If $\rho(|Y_\theta|^2,\ (\sigma^2/|A_\theta|^2)_{m-1}) - \rho(|Y_\theta|^2,\ (\sigma^2/|A_\theta|^2)_m) \le \varepsilon$ stop. Otherwise set $m\to m+1$ and go to (1).

The convergence of the model sequence $(\sigma^2/|A_\theta|^2)_m$, generated by $T_Y$, to the set of fixed points of this mapping is proved in [2] provided that the sequences of vector parameter corresponding to the source AR model and the filter model are in a compact set of the Euclidean space $R^{p+1}\times R^{q+1}$.

The fixed point algorithm was implemented in the frequency domain using the FFT algorithm. It was tailored for white noise whose power spectral density is modeled by a constant $\lambda$. For a given AR model $\sigma^2/|A_\theta|^2$, the minimum distortion filter was obtained by minimizing $d(|Y_\theta|^2,\ \sigma^2/|A_\theta|^2+\lambda)$ over all $\lambda\ge 0$ using the Fibonacci approach. For a given filter $|H_\theta|^2$, the LP coefficients were estimated from the sample correlation obtained from $|Y_\theta|^2|H_\theta|^2$, by the "autocorrelation approach" [4]. The Itakura-Saito distortion measure was calculated in the frequency domain using simple numerical integration.

An FFT of 1024 points was used in all of our experiments in order to prevent aliasing in calculating the sample correlation of the linearly filtered noisy signal $|Y_\theta|^2|H_\theta|^2$ when frames of 300 samples and AR models of 8-th order

are used (see Section 3), and also for achieving reasonable accuracy in calculating the Itakura-Saito distortion measure.

The initial model used here is naturally chosen to be the AR model $(\sigma^2/|A_\theta|^2)_N$ which is directly obtained from the noisy source. However, in order to prevent the algorithm from converging to a fixed point in which $|H_\theta|^2 \equiv 1$, we excluded the noise model $\lambda=0$ from the set over which the filter for the initial model (but not for subsequent models generated from the given initial model) is optimized, i.e., the initial filter was optimized over $\lambda \geq c_0 > 0$. Furthermore, the value of $c_0$ was experimentally chosen so that a "good" initial filter results. Specifically, $c_0$ was determined in according with the dynamic range $\delta$ of the initial model as follows:

$$c_0/\sigma_N^2 = \begin{cases} 2 & \text{for} \quad 0 \leq \delta < 10 \\ 1 & \text{for} \quad 10 \leq \delta < 60 \\ 0.1 & \text{for} \quad 60 \leq \delta \end{cases} \tag{7}$$

The dynamic range $\delta$ is defined as the ratio between the average power in the $M$ highest energy bands and the $M$ lowest energy bands of $(\sigma^2/|A_\theta|^2)_N$. We used 32 equally spaced frequency bands on $0 \leq \theta \leq \pi$ and chose $M=8$ when an FFT of 1024 points was used. The rule of (7) was motivated by the interpretation of the dynamic range as an estimate of the segmental SNR and by the fact that for high SNR we wish to start the iterative procedure with a filter whose spectrum is close to unity, while for low SNR the initial filter should be such that it strongly affects the noisy input. The specific values and ranges which appear in (7) were experimentally determined.

## 3. Speech recognition system

The speech recognition system used in our experiments is completely described in [1], [5] and its main features will be summarized here for completeness. Figure 1 shows a block diagram of the recognizer in its training and testing modes. The clean analog speech, recorded over a local dialed-up telephone line, is first bandpass filtered to 100-3200 Hz and then sampled at 6.67 kHz. The digitized clean speech was manually endpointed to determine the boundaries of each word. The endpoints obtained in this way were used in all of our experiments including those in which noise was added to the signal. In this manner we eliminate the effect of errors in endpoint detection on the recognition accuracy and focus only on the recognition process itself. For each word, 8-th order AR modeling is applied to 300 sample frames of the digital signal. Adjacent frames overlap by 200 samples. The AR modeling is either done by the "autocorrelation method" [4] or by the fixed point algorithm, as discussed in Section 1. Neither windowing nor pre-emphasis was applied in either case. The fixed point algorithm was applied using a stopping threshold of $\varepsilon=0.01$. The algorithm usually converged in a few iterations. For example, for 0 dB segmental SNR, convergence is achieved in less than three iterations on average [2]. The LP coefficients obtained from each analysis frame are transformed into cepstral coefficients using the well known recursion given in [4, p. 230] and then windowed by

$$w(k) = \begin{cases} 1+0.5L\sin(\pi k/L) & 1 \leq k \leq L \\ 0 & \text{otherwise,} \end{cases} \tag{8}$$

where $L=12$, as proposed in [5]. The windowed cepstral coefficients constitute the feature vector in our recognizer and the distance measure used here is the usual Euclidean distance in the cepstral domain [5].

The recognizer is trained to give two templates for each word from a five token training set using a variant of the K-means or the generalized Lloyd clustering algorithm [6]. The distance per word is thus considered as the minimal distance to the two templates. Given the templates for each word, the recognition is done using standard dynamic time warping techniques.

## 4. Recognition results

The speech recognition system described in Section 3 was used to recognize spoken versions of the 26 word English alphabet, the ten English digits, and the three command words "stop," "error," and "repeat," which were artificially contaminated by uncorrelated, additive, zero mean, Gaussian white noise. The statistics of the noise were arbitrarily chosen and its specific form does not effect the performance of the fixed point algorithm as we have argued and demonstrated in [2]. The recognition results were obtained from four talkers (two females and two males), each speaking the 39 word vocabulary 15 times in random order. Five replications were used for training and the remaining ten repetitions were used for testing. Tables 1-3 and Figures 2-4 show the recognition accuracy obtained for each of the three cases discussed in Section 1. The results are separately given for the alphabet, the digits, and for the entire 39 word vocabulary. Each recognition score in these tables and figures represents the average accuracy obtained for the four talkers. The following notation is used.

SNR-input SNR of the tested speech.

CSS-templates were generated from clean speech, and test patterns were generated from noisy input speech, at the indicated SNR's ($\infty$ corresponds to clean speech), using the same standard LP analysis.

CSF-templates were generated from clean speech using standard LP analysis. Test patterns were generated from noisy input speech, at the indicated SNR's, using the fixed point algorithm.

CFF-templates were generated from clean speech, and test patterns were generated from noisy input speech, at the indicated SNR's, using the fixed point algorithm.

| SNR | 5 dB | 10 dB | 15 dB | 20 dB | $\infty$ dB |
|---|---|---|---|---|---|
| CSS | 30.0 | 56.7 | 83.5 | 96.0 | 99.7 |
| CSF | 60.2 | 84.7 | 96.0 | 98.7 | 99.7 |
| CFF | 84.7 | 93.5 | 97.5 | 99.7 | 99.7 |

Table 1: Digit recognition accuracy scores.

| SNR | 5 dB | 10 dB | 15 dB | 20 dB | $\infty$ dB |
|---|---|---|---|---|---|
| CSS | 26.0 | 41.9 | 59.8 | 70.7 | 92.7 |
| CSF | 44.0 | 60.5 | 73.1 | 81.5 | 92.2 |
| CFF | 56.3 | 69.8 | 79.3 | 85.8 | 92.8 |

Table 2: Alphabet recognition accuracy scores.

| SNR | 5 dB | 10 dB | 15 dB | 20 dB | $\infty$ dB |
|---|---|---|---|---|---|
| CSS | 19.2 | 37.3 | 59.6 | 73.8 | 93.8 |
| CSF | 40.8 | 61.6 | 75.5 | 84.4 | 94.1 |
| CFF | 58.7 | 72.4 | 81.5 | 88.5 | 94.2 |

Table 3: 39-word (Alphabet, digits, and the three command words) recognition accuracy scores.

Tables 1-3 show that if the proposed fixed point algorithm replaces the standard LP analysis in the recognition

system, then an improvement in recognition accuracy equivalent to an *increase of about 10 dB* in input SNR is achieved. If, however, the training is done using the standard LP analysis, and testing is done by the fixed point algorithm, then an equivalent improvement in input SNR of about 5 dB is achieved. The better results obtained in the first case are due to the similar distortion the fixed point algorithm introduces in creating the templates and the test patterns, which makes recognition easier.

*References*


[1] L. R. Rabiner and S. E. Levinson "Isolated and connected word recognition-Theory and selected applications, " *IEEE Trans. Comm.*, vol. COM-29, No. 5, pp. 621-659, May 1981.

[2] Y. Ephraim, "An information theoretic approach for autoregressive modeling of noisy sources," submitted for publication.

[3] Y. Ephraim, H. Lev-Ari, and R. M. Gray, "Asymptotic minimum discrimination information measure for asymptotically weakly stationary processes," submitted for publication.

[4] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, NY, 1976.

[5] B. H. Juang, L. R. Rabiner and J. G. Wilpon, "On the use of bandpass liftering in speech recognition, " *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 85-88, April 1986.

[6] J. G. Wilpon and L. R. Rabiner, "A modified K-means clustering algorithm for use in isolated word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, No. 3, pp. 587-594, June 1985.


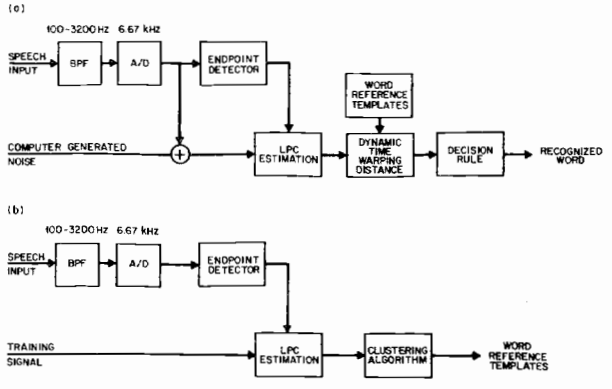


Fig. 1: Block diagram of the recognizer. (a)-Testing mode. (b)-Training mode.

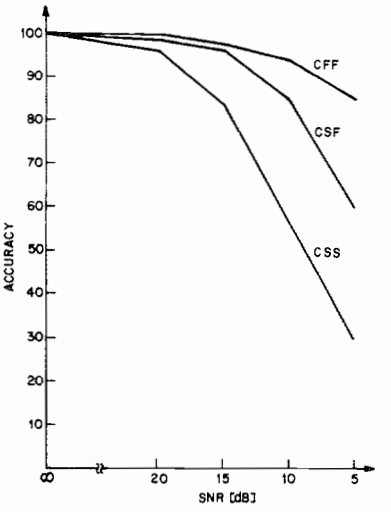


Fig. 2: Digit recognition accuracy scores.

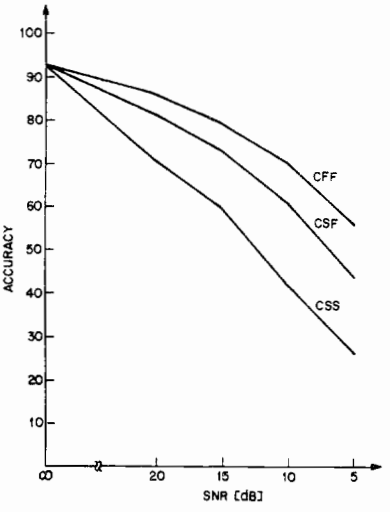


Fig. 3: Alphabet recognition accuracy scores.

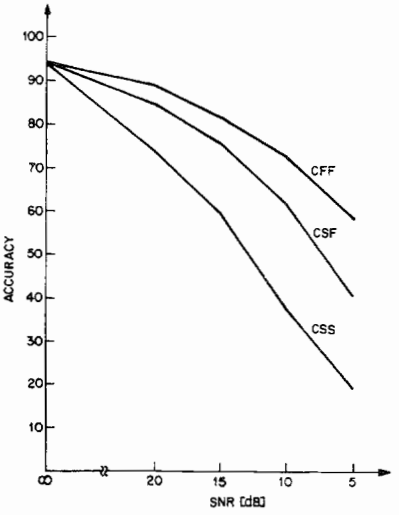


Fig. 4: 39-word (alphabet, digits, and the three command word) recognition accuracy scores.