Volume 2, Issue 8, August 2012 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Throat Microphone Signals for Isolated Word Recognition Using LPC

Anuradha S. Nigade
Department of Electronics Engineering,
Bharati Vidyapeeth Deemed University, Pune (India)

J. S. Chitode
Department of Electronics Engineering,
Bharati Vidyapeeth Deemed University, Pune (India)

Abstract - Automatic speech Recognition (ASR) is technology that allows a computer to identify the words that a person speaks into microphone or telephone. One of the most difficult problems for an automatic speech recognition system resides in dealing with noises. The performance of standard ASR systems using Normal Microphone (NM) degrades even if the ambience is slightly noisy. In this system Throat Microphone (TM) signals are used for isolated word recognition. In contrast to the NM speech the TM speech is unaffected by such an ambience. We use the linear predictive coding (LPC) spectral analysis model for speech recognition. This ASR system is designed to recognize isolated devanagari words.

Keywords – Normal microphone, Throat microphone, Automatic speech recognition, Linear predictive coding

I. INTRODUCTION

Speech recognition acts as an interface between the user and the system. Its applications vary to the extent that it is a successful replacement for input devices like keyboard, mouse etc. ASR is technology that allows a computer to identify the words that a person speaks into microphone or telephone. It is a process by which computer maps an acoustic speech signal to text. Basically it is a process of converting an acoustic signal, captured by microphone or telephone to a set of words. One of the most difficult problems for an automatic speech recognition system resides in dealing with noises. When there are multiple people speaking, it is difficult to determine whether the captured audio signal is from the speaker or from other people. In addition, the recognition error is much larger when the speech is overlapped with other people's speech [1].

Applications such as entry into high-security enclosures and access control may involve noisy environments, for instance, cockpit of an aircraft. For such applications, reliable person identification is required. Speaker recognition is the task of person identification using speech as the biometric feature. A person's voice, like other biometrics (finger prints, retinal patterns or genetic structure), cannot be forgotten or misplaced unlike the use of artifacts for identification by artificial means such as keys or memorized passwords. Hence, speaker recognition is more reliable than other artifacts for person identification.

The performance of standard ASR systems using NM degrades even if the ambience is slightly noisy. In contrast to the NM speech the TM speech is unaffected by such an ambience. So in our ASR system TM signals are used for isolated word recognition. This ASR system may also be useful in situations where NM cannot be used. Speech recorded from a throat microphone is robust to the surrounding noise.

Speech recognition using throat microphone is an attractive alternative in high-noise environments, such as airplane, motorcycle, military field, factory or street crowed environment [3].

II. SYSTEM ANALYSIS

The performance of standard ASR systems using NM degrades even if the ambience is slightly noisy. In contrast to the NM speech the TM speech is unaffected by such an ambience. So in this ASR system TM signals are used for isolated word recognition. This ASR system may also be useful in situations where NM cannot be used. System is developed to recognize the isolated word of Marathi language.

In a noisy environment, the intelligibility of a close-speaking microphone speech is affected, as the microphone picks up not only the voice but also the background noise and reflections from various objects. But the intelligibility of the throat microphone signal is nearly the same as that of the signal obtained in a noise-free environment. Hence the throat microphone is a preferred choice for use in speech applications even in adverse conditions[5].

As mentioned earlier, in adverse conditions, a person's voice is less affected when recorded using a throat microphone than when using a close-speaking microphone.

LPC has widely used for speech recognition due to following reasons.

1. LPC provides a good model of the speech signal. This is especially true for the quasi steady state voiced regions of speech in which the all-pole model of LPC provides a good approximation to the vocal tract spectral envelope. During unvoiced and transient regions of the speech, the LPC model is less effective than for voiced regions, but it still provides an acceptably useful model for speech-recognition purpose.

2. The way in which LPC is applied to the analysis of speech signals leads to a reasonable source-vocal tract separation. As a result, a parsimonious representation of the vocal tract characteristics (which we know are directly related to the speech sound being produced) becomes possible.
3. LPC is an analytically tractable model. The method of LPC is mathematically precise and is simple and straightforward to implement in either software or hardware. The computation involved in LPC processing is considerably less.
4. The LPC model works well in recognition applications. Experience has shown that the performance of speech recognizers, based on LPC front ends, is better.

Block Diagram:

The block diagram of proposed ASR system

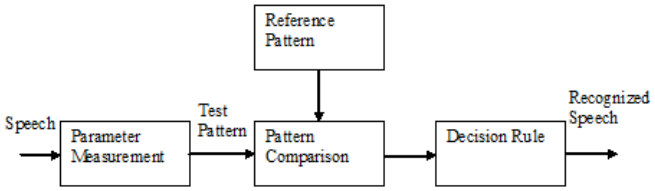


Fig1. Block Diagram of ASR

The three basic steps in the recognition model are:
1. Parameter measurement in which test pattern is created.
2. Pattern comparison
3. Decision making

A throat microphone, placed in contact with the skin surrounding the larynx near the vocal folds, picks up its vibrations and also the signals transmitted through the muscles of the speech production mechanism. The resulting signal is very similar to normal speech. Due to its proximity to the speech production system, speech recorded from a throat microphone is clean, and is not affected by environmental noise.
The Throat Microphone strapped to the throat skin near the larynx, records speech which is intelligible. The throat microphone converts the vibrations that it picks up into equivalent speech signals. Typically, the throat speech is a low amplitude signal when compared to the close-speaking microphone signal. But, it is interesting to note that the throat speech is of high quality. The throat microphone speech is relatively unaffected by background noise and reverberation effects. This signal is almost as intelligible as the close-speaking microphone speech [5]. TM provides high signal to noise ratio over a large part (0-3500 Hz) of audio frequency range. The speech signal recorded by TM is given as an input to parameter measurement block.

The function of parameter measurement block is to represent the relevant acoustic events in the speech signal in terms of compact, efficient set of speech parameters. Here speech is directly represented by the time sequence vectors. Thus test pattern T is defined as the concatenation of spectral frames over the duration of the speech such that

$$T = \{ t1, t2, t3, \ldots, ti\}$$

Where each ti is the spectral vector of the input speech at time. "i" and "T" is the total number of frames of speech. In a similar manner we define a set of reference pattern $\{ R^1, R^2, \ldots, R^v \}$ where each reference pattern R^j is also a sequence of spectral frames such that

$$R^j = \{ r1^j, r2^j, \ldots, rj^j \}$$

The LPC model is used as signal processing front end. Speech knowledge is brought into the system via training procedure.

The goal of the pattern comparison stage is to determine the dissimilarity or distance of T to each of the R^j, $1 \leq j \leq v$, in order to identify the reference pattern that has the minimum dissimilarity. Decision algorithm associates the spoken input with this pattern.

The LPC Model

The basic idea behind the LPC model is that a given speech sample at time n, s(n), can be approximated as a linear combination of the past p speech samples, such that

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \ldots\ldots + a_p s(n-p), \qquad (1)$$

Where the coefficients $a_1, a_2, \ldots\ldots\ldots\ldots, a_p$ are assumed constant over the speech analysis frame.

We convert above equation to an equality by including an excitation term, G u(n), giving:

P

$$s(n) = \sum_{i=1} a_i\, s(n-i) + G\, u(n), \qquad (2)$$

where u(n) is a normalized excitation and G is the gain of the excitation. By expressing Eq.(2) in the z- domain we get the relation

$$S(z) = \sum_{i=1}^{p} a_i\, z^{-i}\, S(z) + G\, U(z), \qquad (3)$$

leading to the transfer function

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad (4)$$

The interpretation of Eq. (4) is given in following figure, which shows the normalized excitation source, u(n), being scaled by the gain, G, and acting as input to the all-pole system, H(z) = 1/A(z), to produce the speech signal, s(n).

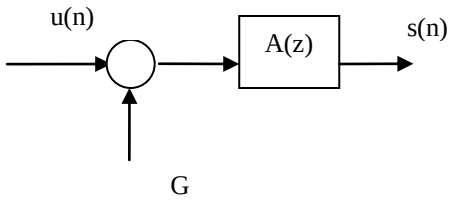


Fig 2. Linear prediction model of speech

LPC Analysis Equations

Based on the model of above figure, the exact relation between s(n) and u(n) is

$$s(n) = \sum_{k=1}^{p} a_k\, s(n-k) + G\, u(n), \qquad (5)$$

We consider te linear combination of past speech samples as the estimate $\tilde{s}(n)$, defined as

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k\, s(n-k). \qquad (6)$$

We now form the prediction error, e(n), defined as

$$E(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k\, s(n-k). \qquad (7)$$

With error transfer function

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{p} a_k z^{-k} \qquad (8)$$

Clearly, when s(n) is actually generated by a linear system of the type shown in above figure then the prediction error, e(n), will equal G u(n), the scaled excitation.

The LPC analysis approach performs spectral analysis on blocks of speech (Speech frames) with an all-pole modeling constraint. The output of the LPC spectral analysis block is a vector of coefficients (LPC parameters) that specify the spectrum of an all-pole model that best matches the signal spectrum over the period of time in which the frame of speech samples was accumulated.

The predictor coefficients are determine by the following equation

$$\Phi n(i,0) = \sum_{k=1}^{p} a\square_k \Phi n(i,k) \qquad (9)$$

Which describe a set of p equations in p unknowns.
To solve Eq.(9) for the optimum predictor coefficients (the $a\square_k$s) we have to compute Φn(i,k) for 1<=i<=p and 0<=k<=p, and solve the resulting set of p simultaneous equations. In practice, the method of solving the equations is a strong function of the range of m used in defining both the section of speech for analysis and the region over which the mean-squared error is computed.
There are two methods for this

1. Autocorrelation
2. The Covariance method

In this ASR system autocorrelation method is used.

The Autocorrelation Method-
Assume that the speech segment, sn(m), is identically zero outside the interval 0<=m<=N-1. This is equivalent to assuming that the speech signal, s(m + n), is multiplied by a finite length window, w(m), which is identically zero outside the range 0<-m<=N-1.
Thus the speech sample for minimization can be expressed as

$$Sn(m)=\begin{cases} s(m+n).w(m), & 0<=m<=N-1 \\ 0, & \text{otherwise} \end{cases} \qquad (10)$$

The purpose of the window of Eq.(10) is to taper the signal near m=0 and near m=N-1 so as to minimize the errors at section boundaries.

LPC front-end processor:

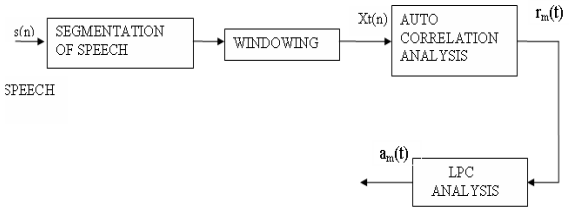


Fig.3 LPC front-end processor

The basic steps in the processing include the following:

1. **Segmentation of speech** – The input speech is divided into small segments. This segments are given as input to next block i.e. windowing.

2. **Windowing –** The next step in the processing is to window each individual segment so as to minimize the signal discontinuities at the beginning and end of each segment. The concept here is to use the window to taper the signal to zero at the beginning and end of each segment. If we define the window as w(n), 0<=n<=N-1, then the result of windowing is the signal

$$x\square_l(n) = x_l(n)w(n), \qquad 0<=n<=N-1$$

A "typical" window used for the autocorrelation method of LPC is the Hamming window, which has the form

$$w(n) = 0.54 – 0.46 \cos (2 \Pi n/N-1), \qquad 0<=n<=N-1$$

3. **Autocorrelation Analysis –** Each frame of windowed signal is next autocorrelated to give

$$r_l(m) = \sum_{n=0}^{N-1-m} x_l(n)\, x_l(n+m), \quad m=0,1,\ldots,p$$

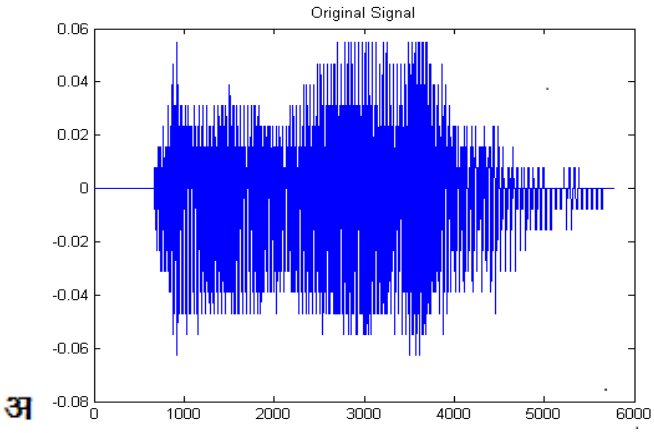
Where the highest autocorrelation value, p , is the order of the LPC analysis. A side benefit of the autocorrelation analysis is that the zeroth autocorrelation, R_l (0), is the energy of the lth frame. The frame energy is an important parameter for speech-detection systems.

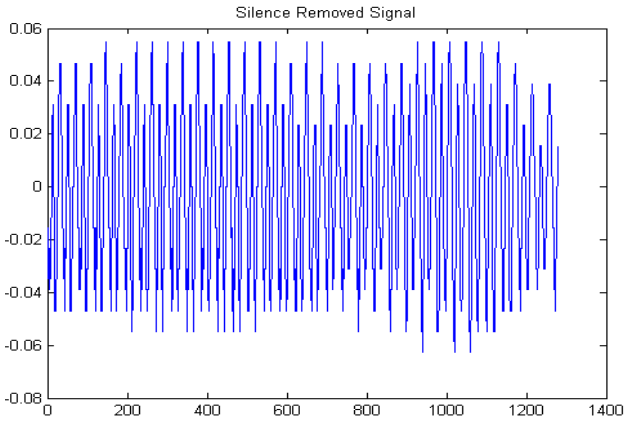
4. **LPC Analysis -** The next step is the LPC analysis, which converts each frame of p+1 autocorrelations into an "LPC parameter set" coefficients, or any desired transformation of the above sets. The formal method for converting from autocorrelation coefficients to an LPC parameter set (for the LPC autocorrelation method) is known as Durbin's method.

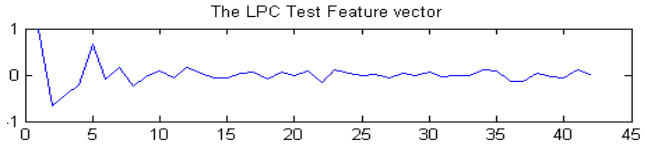
RESULT

Here are two examples for two letters –

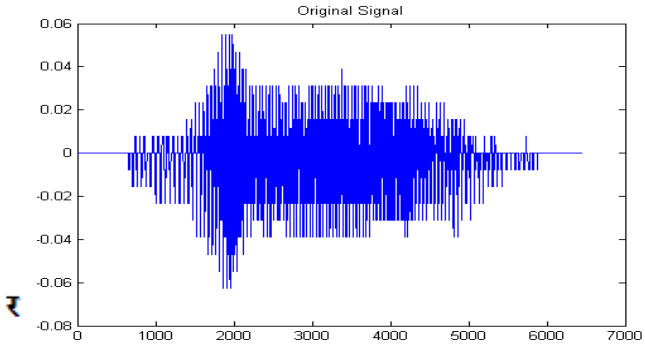
Example No. 1-

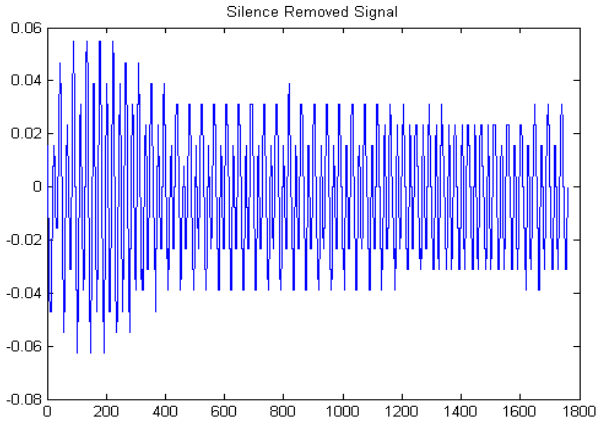


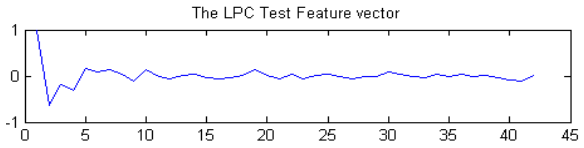




Example No. 2







III. CONCLUSIONS

This paper has described the isolated speech recognition of throat microphone signal using LPC for devanagari language. With experimentation it is shown that system has 85% recognition accuracy. Due to use of throat microphone for recording the performance of system is better even in a normal noisy ambience. Further it is proposed to recognise system for syllable recognition

REFERENCES

[1] Zhang,Z.Liu,M.Sinclair,A.Acero,L.Deng,J.Droppo,X.Huang, and Y. Zeng, Multisensory microphones for robust speech detection, enhancement and recognition, *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, Montreal, Quebec, Canada, May 2004,pp. 781-784*

[2] A. Shahina, and B. Yenanarayana , Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach, *Dept. Comp. Sci. Engg, Indian Institute of Technology, Madras, May 2007*

[3] Engin Erzin, Improving Throat Microphone Speech Recognition by Joint Analysis of Throat and Acoustic Microphone Recordings ,*supported by TUBITAK under project EEEAG 104E176, Turkey, July 17,2009*.

[4] Ren Wenxia, Zhang Huili, Lv Wenzhe, Realization of Isolated-words Speech Recognition System, *Pacific-Asia Conference on Circuits, Communications and System, 2009*

[5] Shahina, B. Yenanarayana and M.R. Kesheory, Throat Microphone signal for speaker recognition

[6] Javier Hernando and Climent Nadeu, Speech Recognition In Noisy Car Enviornment Based On OsaLPC Representation And Robust Similarity Measuring Techniques, *Signal Theory and comm.. Dept., Spain, 1994*

[7] Area V. Nefian, Luhong Liang, Xiaobo Pi, Liu Xiaoxiang, Crusoe Mao and Kevin Murphy, A Coupled HMM For Audio-Visual Speech Recognition

[8] Zhanyu Ma and Arne Leijon, A Probabilistic Principal Component Analysis Based Hidden Markov Model For Audio-Visual Speech Recognition, *Sweden, 2008*

[9] Yasunari Obuchi, Multiple-Microphone Robust Speech Recognition Using Decoder-Based Channel Selection, *Advanced Research Laboratory, Japan, 2004*

[10] John G. Ackenhusen, L.R.Rabiner, Microprocessor Implementation of an LPC-based isolated word recognizer.

[11] Mayukh Bhaowal and Kunal Chawala, Isolated word Recognition for English language using LPC,VQ and HMM.

[12] Eric Fragniere, Andre Van Schaik and Eric Vittoz, Linear Predictive coding of speech using an analog Cochlear model

