Determining Number of Speakers in Multi-Speaker Condition with Additive Noise

Namratha N.
M.Tech Student, Signal Processing,
Siddaganga Institute of Technology,
Tumakuru, Karnataka (India).

R.Kumaraswamy, PhD
Professor & HOD, Dept. of ECE,
Siddaganga Institute of Technology,
Tumakuru, Karnataka (India).

ABSTRACT
The performance of speaker recognition system considerably degrades if the sample used for speaker recognition task has voices from different speakers in the close vicinity. Solutions to these problems are needed, especially for signals collected in a practical environment, such as in a room with background noise and reverberation. This paper presents a method of determining number of speakers in multi speaker condition using excitation source information. Speech in a multi speaker environment are collected using two spatially separated microphones which results in time delay of arrival of speech signals with respect to a given speaker. This time delay is estimated from the cross correlation function of Hilbert envelopes of LP Residual signals. Thus by estimating the difference in time delay for different speakers the number of speakers can be determined. The performance of the proposed method is evaluated by adding different types of noise to the clean speech signal which illustrates the robustness of the proposed method.

Keywords
Excitation source information, Instants of glottal closure (GCIs), Linear prediction(LP) residual, Hilbert envelop (HE) , Time delay estimation, different types of noises.

1. INTRODUCTION
Multi-speaker condition is one of the most challenging conditions because the desired signal is the speech from the desired speaker and the interference signal is also speech from the speakers in the vicinity besides noise and reverberation. The interference signal is still challenging because of its non-stationary property and similarity to the desired target speech. Determining number of speakers in multi-speaker environment attains paramount importance which is often the initial step in many signal processing applications such as array processing, automation transcription of video via audio tracks and audio indexing, source localization and audio tracking of moving speakers, audio recording in a conference or meetings which may contain more than one speaker.

Speech is produced as an outcome of time varying vocal tract system driven by time varying excitation caused by the vibrating vocal cords at the glottis [1]. For a voiced speech the excitation source consists of impulses like excitation around the instants of glottal closure (GCI) called as epochs within each pitch period [7]. The relative spacing of these instants of significant excitation in voiced sound remains unchanged at different microphone locations but differ only by a fixed time delay with respect to relative distance of the microphone from the speaker [2].

Around the instants of glottal closure the speech signal exhibits a high SNR (signal to noise ratio) relative to other regions because of damped sinusoidal components present in the speech signal due to the resonance of vocal tract system. The linear prediction (LP) residual of the multi speaker signal is obtained in order to highlight the high SNR regions by using the method of auto-correlation. The advantage of using LP residual is that it removes the second order correlation among the samples and produces large amplitude fluctuations around the instants of glottal closure, but the cross correlation of LP residual signal doesn't yield prominent peaks as the large amplitude fluctuations will be of random polarity around the GCI's.

Hence, the high SNR regions around the instants of significant excitation are high lightened by computing Hilbert envelop of LP residual signal [3]. The time delay corresponding to different speakers can be estimated using cross correlation function of the multi speaker signals [2]. The advantage of obtaining cross correlation for the multi speaker signal is to avoid unambiguous peaks which may be due to noise and reverberation at each time delay. Thus the position of dominant peaks in cross correlation function gives the time delay due to all the speakers in multi speaker signals. The concept of time delay is exploited in order to determine the number of speakers.

The organization of the paper is as follows: Section 2 provides description of database. Section 3 gives a brief overview of the algorithm followed by a brief description of each step involved. Experimental results of number of speaker determination in different noisy condition are presented in Section IV followed by conclusion and References.

2. DATABASE
Different types of noises such as (stationary)white noise ,and (non- stationary) train noise, machine gun noise, factory noise and type writing noise which are taken from NOISEX-92 database are added to clean multi speaker speech signals to evaluate its effects. Noises are added to multi-speaker signal at 20db, 15db, 10db and 5db SNR's after being scaled.

3. ALGORITHM DESCRIPTION
The data of multi-speakers are collected using two spatially separated microphones as shown in Fig.1 which results in time delay of arrival of speech signal from a given speaker. In this section an overview of how this concept of time delay is exploited in order to determine the number of speakers is presented. S1 and Sn are the no of speakers.

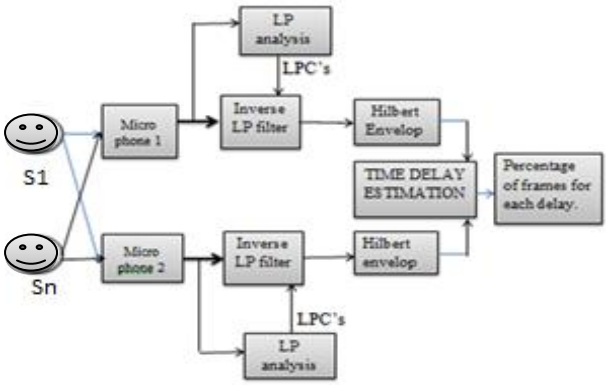

Fig 1: Block diagram of the proposed method

3.1 Preprocessing of Multi-speaker signals

Speech consists of two types of sounds, voiced an unvoiced. The voiced sound decreases by 6db octave which results in falling resonance spectrum. It isvery crucial to even out the spectrum before analyzing it. If this is not done only the peaks at lower frequencies will be found .This straightening is easily done with low pass filtering by choosing filter coefficients slightly lower than unity to avoid instability as given in the following expression,

$$H_{lp} = \frac{1}{1-0.95\,Z^{-1}} \quad (1)$$

Voiced speech is the output of a quasi-stationary vocal tract system excited by quasi periodic puffs of air produced due to vibration of vocal folds. Even though the vibration of vocal folds produces sequences of glottal pulses, the significant excitation occurs only at the instants of glottal closure within each pitch period, called epochs [4]. Around the instants of significant excitation speech signal exhibits high SNR (signal to noise ratio). To determine the time delay the coherence of high SNR regions of the speech signal at two microphones is exploited.

The regions of high SNR in the speech signal are high lightened using linear prediction (LP) residual using the method of autocorrelation. During LP analysis of the speech signal s[n], each sample of s[n] is estimated from the weighted sum of past p samples [6]. If s[n] is the present sample, then its predicted sample ŝ[n] is given by,

$$\hat{s}[n] = -\sum_{k=1}^{p} \alpha_k\, s[n-k] \quad (2)$$

Where p is the order of prediction and α_k are LPC's [Linear prediction coefficients]. The LP (linear prediction) residual or the prediction error e[n] is obtained by,

$$e[n] = s[n] - \hat{s}[n] \quad (3)$$

$$e[n] = s[n] + \sum_{k=1}^{p} \alpha_k\, s[n-k] \quad (4)$$

The linear prediction is performed by autocorrelation method where the autocorrelation function is defined [5] as,

$$R[k] = \sum_{n=k}^{N} s[n]\, s[n-k] \quad (5)$$

Where 'N' is the length of the analysis window. The autocorrelation coefficients R[k] obtained is converted to LP coefficients α_k using Levinson Durbin algorithm.

The LP residual obtained correspond to the estimate of the excitation source of speech signal. Further in order to highlight the high SNR regions around the instants of glottal closure the Hilbert envelop h[n] of the LP residual signal e[n] is computed which is given by,

$$h[n] = \sqrt{e^2(n) + e_h^2(n)} \quad (6)$$

Where $e_h[n]$ is the Hilbert transform of LP residue e[n] which is obtained by interchanging real and imaginary parts of Discrete Fourier Transform (DFT) of e[n] and then taking IDFT (Inverse Discrete Fourier Transform).

$$e_h[n] = IDFT\{E_h[k]\} \quad (7)$$

Where,

$$E_h[k] = \begin{cases} -j\,E[k], & k = 0,1,\ldots\ldots\ldots\ldots\ldots\ldots\ldots\left(\frac{N}{2}\right)-1 \\ +j\,E[K], & k = \left(\frac{N}{2}\right),\left(\frac{N}{2}+1\right),\ldots\ldots\ldots(N-1) \end{cases}$$

Where N is the number of DFT points used and E[K] is the DFT of e[n].

3.2 Determining Number Of Speakers Using Time Delay Estimation

Other than the large amplitudes around the instants of significant excitation, the Hilbert envelope computed also contains a considerable amount of small positive values. So, in order to clearly indicate the epochs, the regions around the instants of significant excitations are further emphasized by dividing square of each sample of HE of LP residual by the moving central average over a short window around the sample. The emphasized / preprocessed Hilbert Envelope $g_i[n]$ is computed as [2] follows,

$$g_i[n] = \frac{h_i^2[n]}{\frac{1}{2M+1}\sum_{m=n-M}^{n+M} h_i\,[m]}\,, \quad (8)$$

$$where \;\; i \in \{1,2,\ldots.p\}$$

Since we consider multi-speaker signals collected using two spatially separated microphones, p=2.Therefore we compute $g_1[n]$ and $g_2[n]$ indivually for the two microphone signals. M is the number of samples corresponding to 4ms duration.

Assuming that the speakers are stationary with respect to microphones and are not positioned along the perpendicular bisector along the line joining of two microphones there exists a fixed time delay of arrival of speech signal.

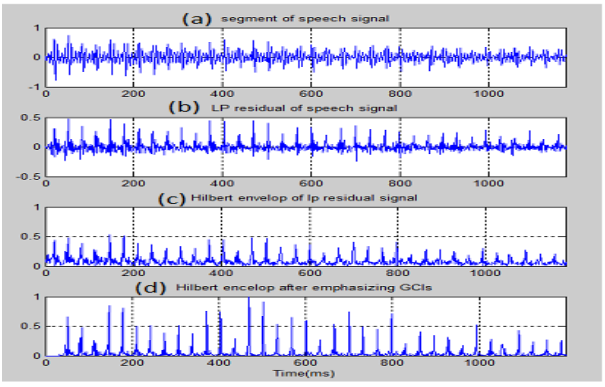

Fig 2: (a) 150ms of noisy speech signal (b) LP residual of the speech signal (c) and (d) Hilbert envelope and emphasized HE respectively.

The time delay corresponding to different speakers can be estimated using normalized cross-correlation function of the emphasized Hilbert envelope of LP residual signals $g_1[n]$ and $g_2[n]$ which is computed as,

$$\rho_{12}[l] = \frac{R_{12}[l]}{\sqrt{R_{11}[0]\,R_{22}[0]}}, l = 0, \pm1, \pm2, \ldots \pm L-1$$

$$\rho_{12}[l] = \frac{\sum_{n=i}^{L-|p|-1} g_1[n]*g_1[n-l]}{\sqrt{\sum_{n=0}^{L-1} g_1^2[n]*\sum_{n=0}^{L-1} g_2^2[n]}} \quad (9)$$

Where $i = l, p = 0$ for $l \geq 0$ and $i = 0, p = l$ for $l < 0$, L is the length of the segment of the Hilbert envelop. The cross correlation is computed over an interval of 2L+1 lags which corresponds to the interval greater than largest expected delay. The largest expected delay is estimated from the approximate positions of the speakers and the microphones in the recording environment[10]. The number of prominent peaks in cross-correlation function corresponds to the number of speakers. However, this is not always true because all the speakers may not contribute to the voiced sounds in the segment that is taken to compute cross correlation and there might be some fallacious peaks in the cross correlation function due to non-speech or noisy frames, which may not correspond to delay due to a speaker [2]. To overcome the above problems, this delay is computed for successive frames of 50ms duration shifted by 5ms duration. The number of frames corresponding to each delay is accumulated over the entire multi speaker data which helps in determining number of speakers, as well as their respective delays.

4. RESULTS

Experiments were conducted on two and three multi speaker data collected using two spatially separated microphones. Different types of noise were added to multi-speaker data at different SNRs [9] to prove the robustness of the algorithm in determining number of speakers accurately. The peaks in the histogram plot indicate the number of speakers.

4.1. Results for white noise added to speech signal at different SNR's.

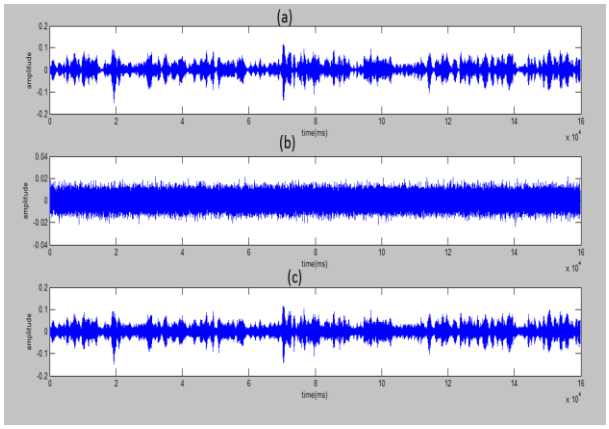

Fig 3. (a) A segment of clean multi-speaker speech signal (b) Generated white noise and (c) Signal plus white noise.

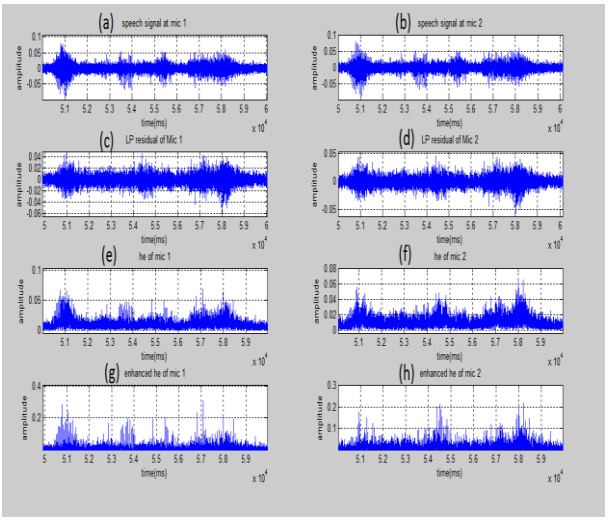

Fig 4. (a) and (b) segment of clean multi-speaker signal at mic 1 and mic 2 (c) and (d) LP residual signals (e) and (f) Hilbert envelope of LP residual signals (g) and (h) Emphasized Hilbert envelope signal.

4.1.1 At 5db SNR

The results for two and three speakers database with white noise at 5db SNR is shown in Fig 5.

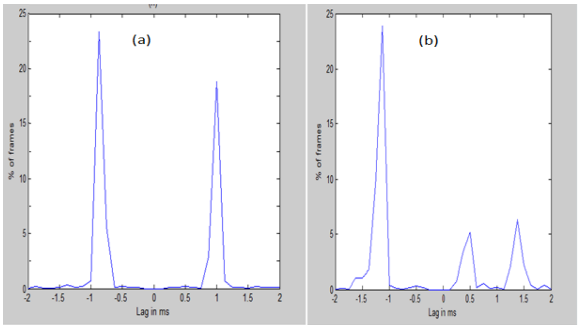

Fig 5. Percentage of frames for each delay in millisecond (ms) for (a) Two (b) Three speaker data at 5db SNR.

4.1.2 At 10db SNR

The results for two and three speakers database with white noise at 10db SNR is shown in Fig 6.

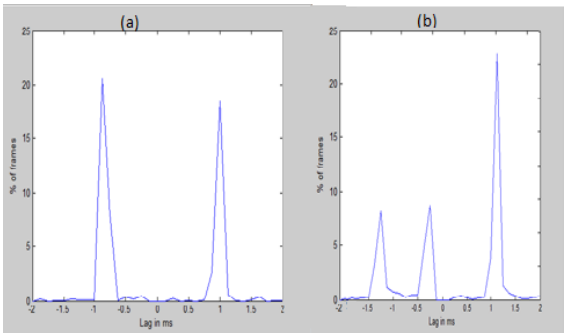

Fig 6. Percentage of frames for each delay in millisecond (ms) for (a) Two (b) Three speaker data at 10db SNR.

4.1.3 At 15db SNR

The results for two and three speakers database with white noise at 15db SNR is shown in Fig 7. As the SNR increases the peaks are more prominent.

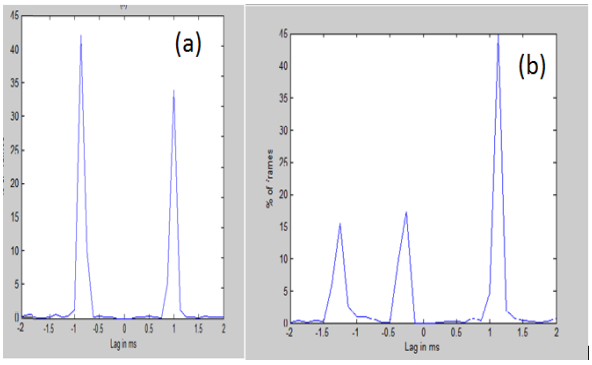

Fig 7. Percentage of frames for each delay in millisecond (ms) for (a) Two (b) Three speaker data at 15db SNR.

4.2. Results for machine gun noise added to speech signal at different SNR's.

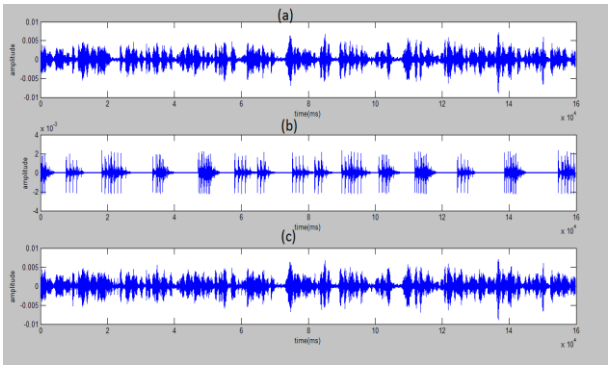

Fig 8. (a) A segment of clean multi-speaker speech signal (b) Generated white noise and (c) Signal plus gun noise.

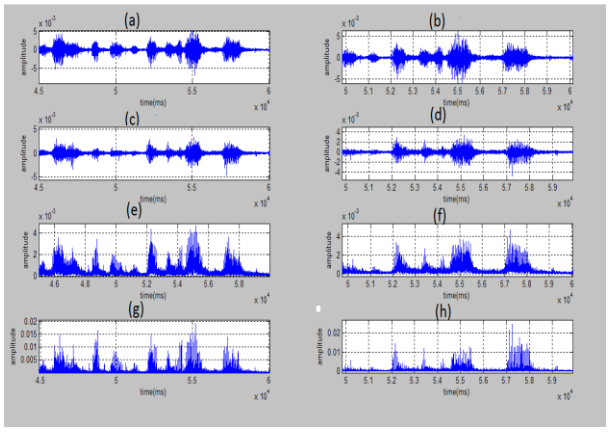

Fig 9. (a) and (b) segment of clean multi-speaker signal at mic 1 and mic 2 (c) and (d) LP residual signals (e) and (f) Hilbert envelope of LP residual signals (g) and (h) Emphasized Hilbert envelop signal.

4.2.1 At 5db SNR

The results for two and three speakers database with machine gun noise at 5db SNR is shown in Fig 10.

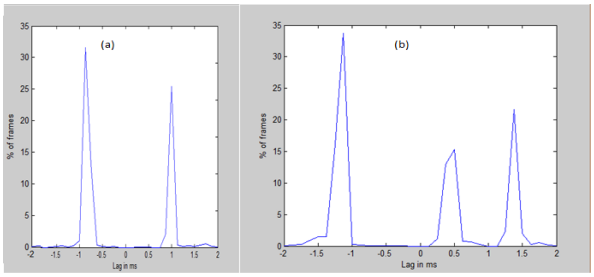

Fig 10. Percentage of frames for each delay in millisecond (ms) for (a) Two (b) Three speaker data at 5db SNR.

4.2.2 At 10db SNR

The results for two and three speakers database with machinegun noise at 10db SNR is shown in Fig 11.

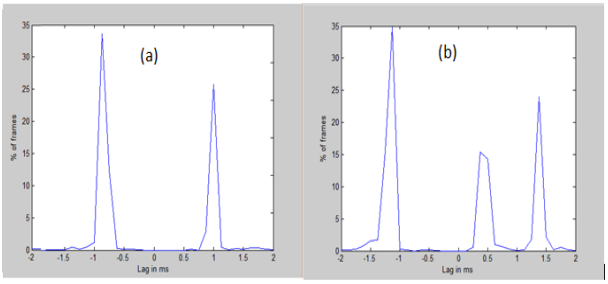

Fig 11. Percentage of frames for each delay in millisecond (ms) for (a) Two (b) Three speaker data at 10db SNR.

4.2.3. At 15db SNR

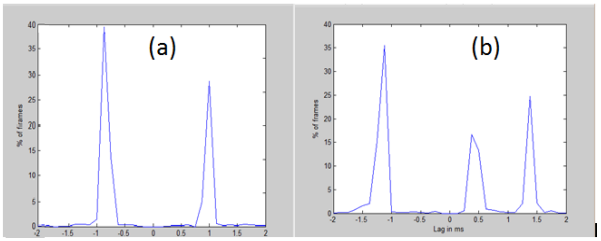

Fig 12. Percentage of frames for each delay in millisecond (ms) for (a) Two (b) Three speaker data at 15db SNR

The results for two and three speakers database with machine gun noise at 15db SNR is shown in Fig 12.

5. CONCLUSION

The concept of time delay estimation is exploited in order to determine the number of speakers. The experimental results show that the location of peaks is not altered and is very similar to the results with clean speech signal. Thus the proposed algorithm is robust not only for clean speech data but also in different noisy condition in determining number of speakers. Present work is done by considering two and three multi-speaker database, which can be further extended to four, five and six or more multi-speaker database. Further, Different types of non-stationary and stationary noises can also be added in order to demonstrate the robustness of the proposed algorithm

6. REFERENCES

[1] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition.

[2] Kumara Swamy.R., Sri Rama Murty. K., & Yegnanarayana.B, "Determining number of speakers from multispeaker speech

[3] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin,"Processing of reverberent speech for time-delay estimation," IEEE Trans. Speech Audio Process., vol. 13, no. 6, pp. 1110–1118, Nov.2005.

[4] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval,"IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.

[5] J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[6] K. Sri Rama Murty, Vivek Boominathan, and Karthika Vijayan, "Allpass modeling of lp residual for speaker recogni-tion," in International Conference on Signal Processing and Communications, SPCOM, July 2012, pp. 1–5.

[7] Ananthapadmanabha, T. V., & Yegnanarayana, B. (1979). "Epoch extraction from linear prediction residual for identification of closed glottis interval". IEEE Transactions on Acoustics, Speech, and Signal Processing, 27, 309–319.

[8] Krishnamoorthy, P., & Prasanna, S. R. M. (2007).Processing noisy speech by noise components subtraction and speech components enhancement. In Proc. int. conf. systemics and cybernetics informatics, Hyberabad, India.

[9] Berouti, M., Schwartz, R., & Makhoul, J. (1979) " Enhancement.

[10] Of speech corrupted by acoustic noise". In Proc. IEEE int. conf.acoust., speech, signal process (pp. 208–211).Smits, R., & Yegnanarayana, B. (1995) ," Determination of instants of significant excitation in speech using group delay function". *IEEE Transactions on Speech and Audio Processing*, *3*, 325–333.