Bessel Features for Estimating Number of Speakers from Multispeaker Speech Signals

Ashok Kumar P.V, Balakrishna L
Department of Electrical Engineering
Blekinge Institute of Technology
Karlskorna, Sweden
vapo08@student.bth.se, bkla08@student.bth.se

Chetana Prakesh, Surayakanth V. Gangashetty
Speech and Vision Laboratory
International Institute of Information Technology,
Hyderabad, India
chetana@research.iiit.ac.in, svg@iiit.ac.in

***Abstract*— In this paper, we explore Bessel features to determine the number of speakers from multispeaker speech signals collected simultaneously from a pair of spatially separated microphones. The arrival of the speech signals from speaker to microphones gives the time delays of given speaker. The time delays can be estimated by performing the cross-correlation to the band limited multispeaker signals collected at the two microphones. Signals are band limited using a finite number of Bessel coefficients. The computer simulation results demonstrate the proposed method is efficient compared to existing methods.**

Keywords: Bessel Expansion, Multispeaker Signals, Time Delay Estimation, and Underdetermined Case

I. INTRODUCTION

In this paper we address a problem of estimating the number of speakers from multispeaker signals collected at a pair of spatially separated microphones. It is also necessary to separate speech of the individual speakers from the multispeaker environment. We need the solution for these problems, especially for signals collected in a practical environment, such as in a room with background noise and reverberation. Several theoretical methods were proposed in the literature for detection of the number of sources whose mixed signals are collected from multiple sensors. The existing method explored the time delays of arrival of speech signals between the two microphones for a given speaker. These time delays are estimated by applying cross-correlation to the Hilbert Envelopes of linear prediction residuals of the multispeaker signals collected from two microphones [1]. The position of dominant peaks in the cross-correlation function of the multispeaker signals give the time delays due to all the speakers at the pair of microphones. Thus, the number of peaks indicates the number of speakers. The method fails if the direct components are masked by high levels of ambient noise and reverberation [1]. All methods, consisting of N observations collected at *p* sensors from *q* sources, is give by :

$$x[n] = As[n] + v[n], \qquad n = 1,2, \dots \dots \dots . N \tag{1}$$

Where, $x[n] = [x_1[n]\ x_2[n] \dots \dots . . x_i[n] \dots \dots \dots x_p[n]]^T$

$$s[n] = [s_1[n]\ \ s_2[n] \dots \dots . . s_j[n] \dots \dots \dots s_q[n]]^T$$

$$v[n] = [v_1[n]\ v_2[n] \dots \dots . . v_i[n] \dots \dots \dots v_p[n]]^T$$

Here, $x_i[n]$ is the mixed signal at the i^{th} sensor, $s_j[n]$ is the signal generated from the j^{th} sources and $v_i[n]$ is the additive noise of the i^{th}sensors. N is the number of observations and A=$[a_{ij}[n]]_{p\times q}$, is mixing matrix. T indicates the transpose operation. The j^{th}column vector of the mixing matrix $A([a_{1j}[n]\ a_{2j}[n] \dots a_{pj}[n]]^T)$ gives the array response associated with the j^{th} source signal. The i^{th} row vector of the mixing matrix $A([a_{i1}[n]\ a_{i2}[n] \dots a_{iq}[n]]^T)$ gives the mixing weights for the source signals collected at i^{th} sensor.

For determining the number of sources, three cases are considered: over determined case ($p > q$), well determined case ($p = q$), and underdetermined case ($p < q$). Most of the cases for estimating the number of speakers use normally generated mixed signals based on model (1). Practical signals are collected from a number of speakers speaking simultaneously have much more variability due to noise and reverberation, besides delay and decay of the direct sound due to distance of the microphones from the speaker.

In this method, we assume that the speakers are stationary with respect to the microphones; there exists a fixed time delay of arrival of speech signal for a given speaker. The time delays corresponding to different speaker is estimated by performing the cross-correlation function on the band limited multispeaker signals. Positions of dominant peaks in the cross-correlation function of the multispeaker signals give the time delays due to all the speakers at the pair of microphones. However, in general the cross-correlation function of the multispeaker signals does not show unambiguous prominent peaks at the time delays. This is mainly because of the damped sinusoidal components in the speech signal due to resonances of the vocal tract, and also because of the effects of reverberation and noise. These effects can be reduced by band limiting the signals to low frequency components using appropriate range of Bessel coefficients. Bessel functions provide the desired properties of the speech signals as both of them have regular zero crossing and decaying amplitudes and also the Bessel features are efficient in representing speech-like waveform [2].

The paper organized as follows: In Section 2, we describe the Bessel expansion of signals. The construction of band limited signals for the given speech signal is also described. Section 3 describes an approach for detection of number of speakers using cross-correlation on band limited multispeaker speech signals. In Section 4, we study the performance of proposed approach for the estimation of number of speakers.

II. ESTIMATION OF BANDLIMITED MULTISPEAKER SIGNALS USING BESSEL EXPANSION

The Bessel expansion is suitable for non-stationary signals representation [3]. The wave equation inside cylindrical structures (tubes) includes the first kind of Bessel function. The vocal tract can be modeled as organ pipe like cylindrical tubes with a sound source at one end (the larynx or voice box) and open at other ends (the lips or nose). It is a good approximation to choose the first kind Bessel functions in terms of naturalness for representing the sounds produced in the vocal tract it could be approximated as acoustic tubes for short-time intervals analysis [4].

A. Coefficients Computation Using Bessel Expansion

The zero-order Bessel series expansion of the signal $x(t)$ considered over some arbitrary interval $(0, a)$ is expressed as in [3] [4] [5] [6] [7].

$$x(t) = \sum_{m=1}^{Q} C_m J_0\left(\frac{\lambda_m}{a} t\right) \tag{2}$$

Where $\{\lambda_m, m = 1,2,3 \ldots . Q\}$ are the ascending-order positive roots of $J_0(\lambda) = 0$, and $J_0\left(\frac{\lambda_m}{a} t\right)$ are the zero-order Bessel functions. A speech signal is approximated as a linear combination of the orthogonal Bessel functions. The sequence of Bessel functions forms an orthogonal set on the interval $0 \le t \le a$ with respect to the weight t.

$$\int_0^a t J_0\left(\frac{\lambda_m}{a} t\right) J_0\left(\frac{\lambda_n}{a} t\right) dt = 0, \quad \text{for } m \neq n \tag{3}$$

The Bessel coefficients C_m are computed by using the following equation:

$$C_m = \frac{2 \int_0^a t x(t) J_0\left((\lambda_m / a) t\right) dt}{a^2 [J_1(\lambda_m)]^2} \tag{4}$$

With $1 \le m \le Q$, where Q is the order of the Bessel expansion and, $J_1(\lambda_m)$ is the first order Bessel functions. The Bessel expansion order Q must be know a priori. The intervals between successive zero crossings of the Bessel function $J_0(\lambda)$ increases slowly with time and approaches to π in the limit. If order Q is unknown, then in order to cover full signal bandwidth, the half of the sampling frequency,Q must be equal to length of the signal. The Bessel series expansion coefficients C_m are unique for a given signal; similarly as the Fourier series coefficients are unique for a given signal. However, unlike the sinusoidal basis functions in the Fourier series, the Bessel functions decay over time. This feature of the Bessel functions makes the Bessel series expansion suitable for the analysis of non-stationary signals such as speech [3] [4] [5] [6] [7]. Speech signals are sampled at 44.1 kHz. We calculate the Bessel coefficients by framing the multispeaker signals using (4). We know that, the highest frequency component of the signal is half the sampling frequency (fs/2), we band limiting the signal to half the highest frequency component of the signal (fs/4). On the extracted coefficients of zero order Bessel function; we performed inverse Bessel transform method to synthesize the band limited signals. Fig.1 shows the multispeaker speech signal collected from Mic-1 and the corresponding band limited signal. The band limited multispeaker signals are used to estimate the time delays.

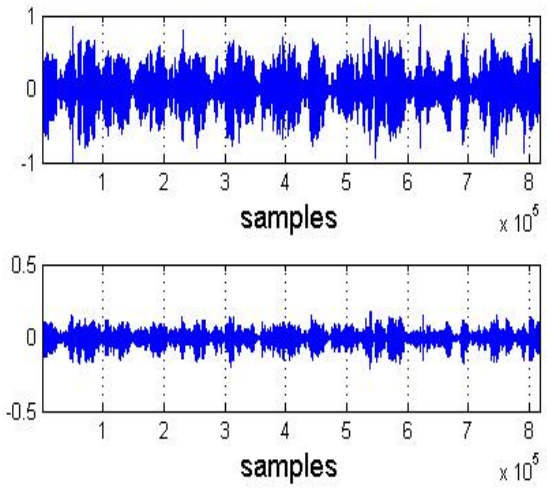

Figure.1. (a) Multispeaker speech signal (b) Band limited multispeaker speech signal using Bessel expansion (fs/4).

III. CROSSCORRELATION OF THE BAND LIMITED MULTISPEAKER SPEECH SIGNALS

The band limited signal of multispeaker signal collected from i^{th} microphone is given by:

$$g_p[m] = \sum_{m=1}^{Q} C_m J_0\left(\frac{\lambda_m}{a} t\right) \quad , m \in \{1,2, \ldots \ldots, Q\} \tag{5}$$

Where Q is the number of samples corresponding to signal duration, and p is the number of microphones. In this paper, we consider multispeaker signals collected using a pair of microphones, and hence *p=2*. The cross-correlation function of the band limited signals derived from the multispeaker signals is used to determine the number of speakers. The cross-correlation function $r_{12}[n]$ between the band limited signals $g_1[m]$ and $g_2[m]$ is computed as:

$$r_{12}[n] = \frac{\sum_{m=z}^{N-k-1} g_1[m]\, g_2[m-n]}{\sqrt{\sum_{m=z}^{N-k-1} g_1^2[m] \sum_{m=z}^{N-k-1} g_2^2[m]}}\,, \quad n = 0, \pm 1, \pm 2, \ldots \ldots \pm L \tag{6}$$

Where $z = n$, $k = 0$ for $n \ge 0$, and $z = 0$, $k = n$ for $n < 0$, and N is the length of the segments of the band limited signal. The cross-correlation function is computed over an interval of $2L+1$ lags, where $2L+1$ corresponds to an interval greater than the largest expected delay. The largest expected delay can be estimated from the approximate position of the speakers and microphones in the room. The location of the peaks with respect to the origin (zero lags) of the cross-correlation function corresponds to the time.

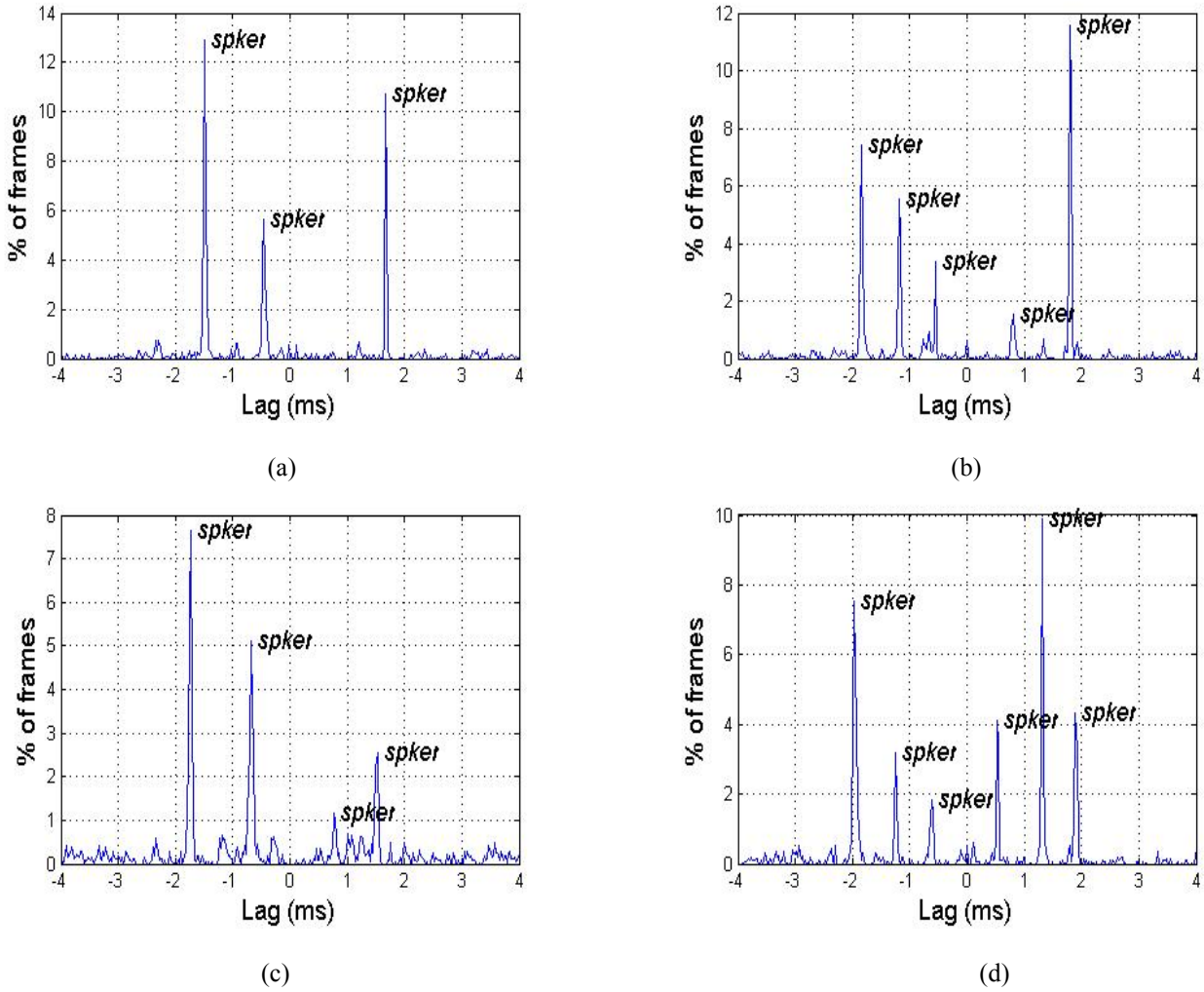

Figure.2. Percentage of frames for each delay in milliseconds for (a) three speakers, (b) four speakers, (c) five speakers, (d) six speakers.

Delay between the microphone signals for all the speakers. The number of prominent peaks should correspond to the number of speakers. However, in practice, this is not always true because of the following reasons: *1)* all the speaker may not contribute to voiced sounds in the segment used for computing the cross-correlation function and *2)* there could be spurious peaks in the cross-correlation function, which may not correspond to the delay due to the speaker.

Hence, we rely only on the delay due to the most prominent peak in the cross-correlation function [1] [8]. This delay is computed from the cross-correlation function of successive frames of *50* ms duration shifted by *5* ms. since the different regions of speech signals may provide evidence for the delays corresponding to different speakers, the number of frames corresponding to each delay is accumulated over the entire data. This helps in the determination of the number of the speakers, as well as their respective delays. Thus, by collecting number of frames corresponding to each delay over the entire data, there will be large evidence for the delays corresponding to the individual speakers. Fig.2 shows the evidence in favor of each delay, for a recording consisting of the speech from three, four, five and six speakers. The figure shows prominent peaks corresponding to the three, four, five and six speakers.

IV. STUDIES ON ESTIMATION OF NUMBER OF SPEAKERS IN MULTISPEAKER SPEECH SIGNAL

In this experiment, we assumed that all the speakers were stationary with respect to microphones. Total experiment is performed in a laboratory environment for different speakers and signals are obtained in these conditions consists of background noise and reverberation. The average reverberation time is 0.5 sec. The average distance between speaker and microphone is 1.5 meter and microphones are separated by 1 meter. All the speakers and microphones were positioned approximately in the same plane. It ensures that each speaker produce different time delays at the corresponding microphones. All the speakers were spoken simultaneously during the entire recording and we assumed that the noise and reverberation components in the room are significantly less than the direct component of speech from each speaker. The speech signals are sampled at 44.1 kHz. We measured distances between the speakers from both the microphones. The actual time delay of arrival τ of speech signals at Mic-1 and Mic-2 located at distances d_1 and d_2 respectively, from a speaker is given by:

$$\tau = \frac{d_1 - d_2}{c} \tag{7}$$

Where c is speed of sound in air. A negative time delay (lead) indicates that the speaker is nearer to Mic-1 relative to Mic-2 [1] [8]. We computed Bessel coefficients for mixed speech signals collected from the Mic-1 and Mic-2. The Bessel coefficients and zero order Bessel functions are used to synthesize the band limited signals using (2).

TABLE I. Comparisons of proposed method estimated time delays τ_p, existing method time delays τ_e and theoretically computed time delays τ.

Number of speakers	Speaker id	d_1 (ms)	d_2 (ms)	τ (ms)	τ_e (ms)	τ_p (ms)
3	Spkr-1	0.45	0.98	-1.5	-1.47	-1.49
	Spkr-2	0.93	1.11	-0.51	-0.47	-0.47
	Spkr-3	1.48	0.91	1.63	1.69	1.66
4	Spkr-1	0.55	1.14	-1.7	-1.72	-1.73
	Spkr-2	1.01	1.23	-0.63	-0.65	-0.67
	Spkr-3	1.43	1.17	0.74	0.81	0.76
	Spkr-4	1.21	0.68	1.5	1.5	1.52
5	Spkr-1	0.6	1.24	-1.83	-1.8	-1.84
	Spkr-2	0.88	1.29	-1.2	-1.13	-1.19
	Spkr-3	1.30	1.49	-0.54	-0.56	-0.55
	Spkr-4	1.42	1.14	0.80	0.81	0.80
	Spkr-5	1.16	0.54	1.77	1.81	1.80
6	Spkr-1	0.4	1.08	-1.9	-2	-1.98
	Spkr-2	0.82	1.29	-1.3	-1.25	-1.26
	Spkr-3	1.19	1.4	-0.6	-0.59	-0.62
	Spkr-4	1.39	1.18	0.6	0.56	0.54
	Spkr-5	1.42	0.96	1.3	1.31	1.3
	Spkr-6	1.4	0.75	1.9	1.94	1.9

The signal is band limited to half of the highest frequency component of the signal (fs/4). The cross-correlation function of the band limited signals of the multispeaker signals is used to estimate the time delays τ_p. The percentage of the frames for each delay (in ms) for three, four, five, and six speakers are shown in Fig.2. The number of dominant peaks in the Fig.2 corresponds to the number of speakers and location of the peaks corresponds to the time delays due to the different speakers. Table 1 lists the actual time delay τ obtained from the measured distances d_1 and $d_2$using (7). Here, d_1 and d_2 are the distance between the microphones and corresponding speaker. The estimated time delay τ_p is obtained by our proposed approach, τ_e is the existing method time delay [1]. Table 1 show that the actual time delays τ and estimated time delays τ_p from proposed approach are very closer compare to existing method time delays τ_e. This indicates the effectiveness of the proposed method in determining the number of speakers and their corresponding time delays form multispeaker speech signals.

V. SUMMARY AND CONCLUSIONS

The proposed method exploits the time delay of arrival of speech signals between the two microphones for a given speaker and also recognizes the number of speakers. When multispeaker speech signals are band limited to low frequencies components using Bessel expansion, we obtained source information of the signal. It works in underdetermined case, where the number of sensors far less than the number of sources. In this study microphones are stationary, so time delays are constant. The proposed method demonstrated for the case where the time delays are distinct for each speaker and it is more efficient than the existing method as per results

REFERENCES

[1] R. Kumar Swamy, K. Sri Rama Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Processing Lett*. vol. 14. No. 7, pp 481-484, July 2007.

[2] C. S. Chen, K. Gopalan, P. Mitra, " Speech signal analysis and synthesis via Fourier Bessel representation", *proc. IEEE International Conference on Acoustics, speech and signal Processing*, vol. 2, pp 497-500, Tampa, Florida, March 26-29, 1985.

[3] J.Schroeder, "Signal processing via Fourier–Bessel series expansion," *Digital Signal Processing*, vol. 3, pp. 112-124, 1993.

[4] F. S. Gurgen, C. S. Chen, "Speech enhancement by Fourier-Bessel coefficients of speech and noise," *IEEE Proceedings*, vol. 137, no. 5, pp 290-294, October 1990.

[5] R .B. Pachori, P. Sircar, "EEG signals analysis using FB expansion and second-order TVAR," *Signal Processing*, vol. 88, pp. 415-420, 2008.

[6] R .B. Pachori, P. Sircar, "Analysis of multicomponent AM-FM signals using FB-DESA method," *Digital Signal Processing,* vol. 20, pp 42-62. 2010.

[7] R .B. Pachori, P. Sircar, "Speech analysis using Fourier-Bessel expansion and discrete energy separation algorithm," *Proc IEEE Digital Signal Processing workshop,* pp. 423-428, 2006.

[8] B. Yegnanarayana, R. Kumar Swamy, and K. Sri Rama Murty, "Determining mixing parameters from multispeaker data using speech- specific information," *IEEE Trans. Audio, Speech and Language Processing*, vol. 17, no. 6, pp 1196-1207, August 2009.