14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

MULTI-SPEAKER VOICE ACTIVITY DETECTION USING ICA AND BEAMPATTERN ANALYSIS

S. Maraboina [1], *D. Kolossa* [2], *P.K. Bora* [1] *and R. Orglmeister* [2]

[1] Indian Institute of Technology, Guwahati, Assam, India.

[2] Technical University of Berlin, Electronics and Medical Signal Processing,
Einsteinufer 17, D-10587 Berlin, Germany
phone: + (0049) 3031423889, fax: + (0049) 31422120, email: m.suneeth@iitg.ernet.in
web: http://www.emsp.tu-berlin.de/

ABSTRACT

Voice activity detection is a necessary preprocessing step for many applications like channel identification or speech recognition. The problem can be solved even under noisy conditions by exploiting characteristics of speech and noise signals. However, when more speakers are active simultaneously, these methods are generally unreliable, since multiple speech signals may overlap completely in the time-frequency plane. Here, a new approach is suggested which is applicable in multi-speaker scenarios also, owing to its incorporation of higher order statistics. Here, independent component analysis is used to obtain estimates of the clean speech and the angles of incidence for each speaker. Subsequently, these estimates can help to correctly identify the active speaker and perform voice activity detection.

The suggested approach is robust to noise as well as to interfering speech and can detect the presence of single speakers in mixtures of speech and noise, even under highly reverberant conditions at 0dB SIR.

1. INTRODUCTION

Multi-speaker voice activity detection is an important preprocessing step for many noise reduction and channel estimation methods e.g. in [3]. It greatly improves the speed and reliability of speaker recognition systems and it also reduces the error rate and decreases the computational effort of speech recognizers. The problem of speaker identification is complex enough as it is, when the speakers are given individual microphones. But, practically, this might not be possible in real situations, where microphones are placed close to each other picking up neighboring speakers along with the ambient noise. Furthermore, unknown channel characteristics make the situation more difficult and challenging. Labeling of the active segments of speakers may be done using computationally less expensive single-channel methods. In many cases, speech and non-speech parts can be separated based on the energy levels of speech and proper thresholds which are dictated by the amount of noise present [5]. But the same principle cannot be extended to multi-speaker voice activity detection, when energy levels of all individual speech signals in the mixture can be almost equal and signals may overlap completely in the time-frequency plane. Especially in this case of multiple, simultaneously active speakers, robustness greatly improves when localization information is learned and used online.

In this paper, a new approach is suggested to detect the activity of the speakers using frequency domain independent component analysis (ICA) and subsequent beampattern analysis. The ability of the proposed method of frequency domain ICA to separate the convolutive mixtures has been shown e.g. in [1]. Although frequency domain ICA can separate the speakers in convolutive mixtures, it cannot recover them completely because of the permutation problem that arises after the separation of sensor outputs. So ICA must be applied in conjunction with another method which can solve the permutation problem. Beampattern analysis [4] is an efficient method to resolve the permutation problem. In this method, spatial features are exploited to rearrange the permuted frequency components. An efficient algorithm with a low error rate has been implemented using the same spatial features which are computed for detecting the permutations in the frequency domain. Voice activity detection (VAD) is applied on the speech mixture to detect activity of each of the present speakers and to assign labels identifying which speaker is active in which segments of the recorded signal.

The remainder of the paper is organized as follows. In section 2, the underlying principles of frequency domain ICA and beampattern analysis are explained. Section 3 presents the actual algorithm of beampattern-based VAD proposed for multi-speaker voice activity detection. Section 4 presents the experimental results obtained on real recordings. Section 5 concludes the paper with observations on the applicability and quality of the suggested approach, as well as with suggestions for further developments.

2. FREQUENCY DOMAIN ICA AND BEAMPATTERN ANALYSIS

Below sections briefly explain frequency domain ICA and beampattern analysis. They highlight the main principles that are being exploited in the described multi-speaker voice activity detection algorithm.

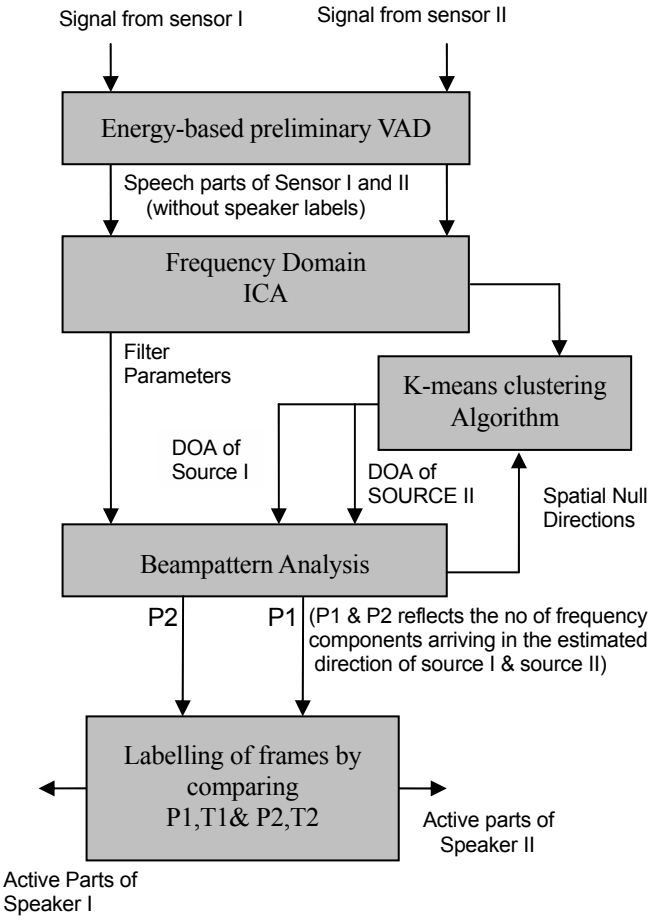

Figure 1: Proposed beampattern-based VAD. T1 and T2 are thresholds pertaining to Source and Source II, respectively.

2.1 Frequency Domain ICA

In convolutive mixtures, where sensor outputs contain source signals convolved with the room impulse response the output can be mathematically represented as convolution of source signal and room impulse response as below:

$$\mathbf{x}(t) = \mathbf{A}(t) * \mathbf{s}(t) . \quad (1)$$

This convolutive model in the time domain is difficult to treat using ICA when compared to a simple instantaneous or anechoic model. But it can be reduced to an instantaneous mixing model, if transformed to the frequency domain. Short time Fourier transform is performed frame by frame on the sensor outputs. A spectrogram is constructed aligning all the calculated STFT coefficients over time. Now, each frequency bin can be considered over the entire time interval as shown below.

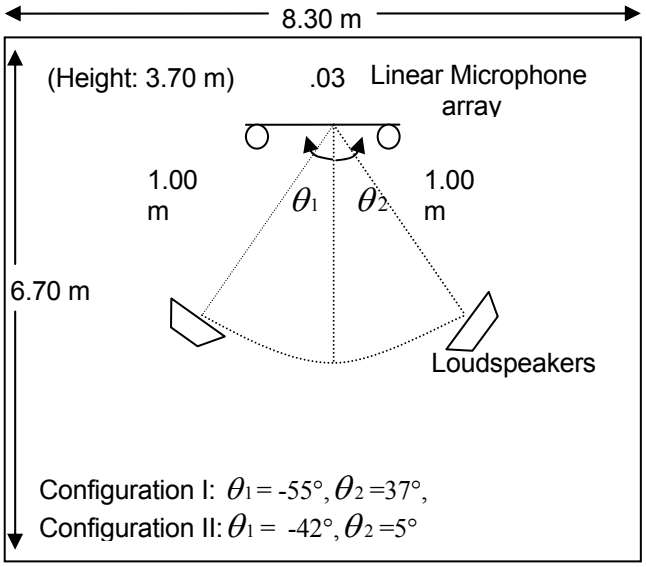

Figure 2: Layout of the reverberant room used in the experiments.

$$\mathbf{X}(\Omega,t) = [X_1(\Omega,t),\ldots, X_k(\Omega,t)]. \quad (2)$$

The corresponding mixing model pertaining to each frequency bin can be represented as

$$\mathbf{X}(\Omega,t) = \mathbf{A}(\Omega)\mathbf{S}(\Omega,t) \quad (3)$$

with **X, A** and **S** representing the recorded signals, the mixing matrix and the sources, respectively. Thereafter, a complex JADE algorithm based on joint diagonalization of the fourth order cumulant matrices is applied to determine the demixing filters for each frequency bin [2]. Unfortunately, performance of frequency domain ICA is severely degraded by the permutation problem inherent in all convolutive source separation tasks solved in the frequency domain.

2.2 Beampattern Analysis

The spatial properties of the demixing filters estimated by the ICA stage can be exploited to solve the permutation problem. Directivity patterns of demixing filters are calculated to obtain the spatial zero directions or nulls. This method has been successfully used to solve the permutation problem in e.g. [4]. Assuming that all incoming signals are planar waves obeying the far-field sound propagation model, the directivity patterns of a demixing filter $W(f)$ and their spatial zeros with respect to broadside are calculated using the function

$$F_l(f,\theta_l) = \sum_{k=1}^{K} W_{lk}(f)\exp[-j2\pi f d_k \sin\theta_l / c], \quad (4)$$

$$f = \Omega / 2\pi \tag{5}$$

where $W_{lk}(f)$ is the *l,k*'th element of $W(f)$, d_k is the distance between reference microphone and the *k*'th microphone, *c* is the velocity of sound and *l* designates the source. One can estimate the direction of arrival of sources by observing the beampatterns of ICA filters as defined in the above equation.

For beampattern-based VAD, in this paper a 2x2 mixing model is considered. For such a system, we can observe that for each filter, nulls exist in the directions of the main interferences. Since higher order statistics are used, which are zero for Gaussian noise, in the majority of cases these correspond to the speaker directions. So spatial zero directions for the above model can be used for identifying the active speakers in a given frame. These are calculated by finding the angles at which the beampattern function is minimum.
Mathematically, they are defined as

$$\theta_1(f) = \arg\min_{\theta} |F_1(f,\theta)|, \tag{6}$$

$$\theta_2(f) = \arg\min_{\theta} |F_2(f,\theta)|. \tag{7}$$

This model allows spatial nulls to vary over frequency, allowing for accurate modeling of reverberation effects, also.

2.3 K-Means Clustering Algorithm

In the suggested algorithm, estimates of directions of arrival (DOA) of sources are required to perform accurate detection of active parts of the speakers. An approximate estimation of the DOAs of sources is possible by observing the null directions of unmixing coefficients obtained in each segment of speech and clustering them using the K-means algorithm. Since there are two sources in the model assumed, one can start clustering the null angles with two centriods. Generally, if the number of sources is known already, then a simple K-means algorithm based on Euclidian distance as a criterion is sufficient. An acceptable performance can be obtained considering the fact, that all the null directions observed in each segmented speech are highly clustered around the original DOAs of the sources. The algorithm is briefly described below.

- Start with the initial assumed seeds C_1 and C_2
- Until there are no changes in the cluster centroids, collect the set of all null angles obtained in each segmented speech.
    - Use the estimated centriods C_1 and C_2 to classify the data into clusters by assigning them to the closest centriod.
    - For i =1,2;
        - Replace the C_i with centroids of the new clusters.
    - End For
- End Until

3. BEAMPATTERN-BASED VAD

Although conventional frequency domain ICA separates the speech signals, it has to be applied in conjunction with other methods like beampattern analysis to resolve the permutation problem. The spatial features of the demixing filters, which are exploited to solve the permutation problem, can also be used for detecting the active parts of the speakers. Even though spatial nulls are dependent on frequency, almost all the frequency components have approximately similar beampatterns, provided that they are arriving from a same direction or source. So, the frequency components pertaining to a particular source can be computed by observing the spatial nulls and successively comparing with an estimated DOA of the source. Hence, estimation of DOA of source signals is crucial. In this algorithm, the DOA of source signals is estimated using the K-means clustering algorithm (Sec 2.3). The overall algorithm is presented in a block diagram (see Fig. 1).

[Step 1] Voice activity detection is performed as a preprocessing step on the sensor outputs. VAD separates the speech and non-speech segments based on the energy levels and zero crossing rates as described in [5]. It significantly decreases the number of false detections of active speech segments, because in non-speech segments, ICA outputs themselves are not reliable indicators.

[Step 2] For each speech segment, demixing filter parameters are estimated over 500ms-blocks using frequency domain ICA. A complex JADE algorithm based on the joint diagonalization of fourth order cumulant matrices is used to find the inverse mixing matrix and to compute estimated source signals.

[Step 3] Spatial nulls are calculated for all frequency bins using beampattern analysis. The beampattern defined in Eq. (4) is calculated for every high energy frequency component and spatial nulls are determined by finding the minima using Eq. (6) and (7). All these nulls are compared with the estimated DOA of source I and source II. Now, a segment can be labeled as an active part of speaker I, if a high percentage of the demixing filters have nulls similar to its estimated mean DOA. The same principle is followed to identify the active parts of speaker II. But before that, prior information about the DOA of the source signals is needed. A sequential k-means clustering algorithm has been implemented for an approximate estimation of the source's DOA. The number of clusters that has to be considered depends on the number of sources present. As mentioned above, spatial nulls of the demixing filters are oriented only in the direction of sources.
If the sources are not quickly moving, it is possible to estimate the number of sources as well as their directions by the k-means clustering algorithm. Few initial segments of microphone outputs are used to determine the approximate DOA of the source signals.

VAD	FAR (%)	FRJ(%)	ERR(%) (without SER)
Speaker I	39.9	3.7	44.6
Speaker II	54.7	7.2	61.9

Table1: Results obtained when only VAD is applied on the real room recordings.

MSAD Algorithm	FAR (%)	FRJ (%)	SER(%)	ERR(%) (with SER)
Speaker I	24	9.9	1.95	35.83
Speaker II	16	9.9	0.03	25.93

Table 2: Results obtained when Beampattern-Based VAD is applied on the real room recordings.

[Step 4] In the last step of the algorithm, to detect active segments of the speakers, thresholds are defined for the 2x2 model, one for each speaker. These thresholds are the minimum number of frequency components having similar spatial nulls that must be present in a segment to declare a particular speaker to be probably active. The proper setting of these thresholds is crucial for detection of active speaker. These thresholds are set by observing the number of frequency components apparently arriving from the direction of the sources when the input is noise or silence only. For best results, the thresholds are set to twice the values that are observed.

4 EXPERIMENTS AND RESULTS

4.1 Test data

In the experiments, speech signals by four different speakers obtained from the TI-Digits database were played back and recorded. All speech signals are 4 minutes long. Here, the 2x2 mixing model was considered. Therefore, only two different speakers are played back at any time using loudspeakers. Microphone outputs are recorded while playing two males in one case and one female and one male speaker in the other. Loudspeakers are placed at a distance of 1 m from the microphone array. In the first configuration, loudspeakers are placed at -55° and 37° relative to broadside and in the second, speakers are placed relatively close to each other with angles -42° and 5°(refer to Fig 2). In both the configurations, a simple model is considered by placing the speakers in two different quadrants with respect to the broadside direction of the microphone array. All the recordings are performed in a lab room of 6.7 m x 8.3 m x 3.7 m. Distance between the microphones is crucial and determines the beampattern of the demixing filters. It also results in spectral aliasing if the distance between the microphones d does not obey the inequality

$$d \leq c / f_{\text{Samp}}, \tag{8}$$

where c is the velocity of the sound in air and f_{Samp} is the sampling frequency of the signal. But if the maximum frequency f_{Max} present in the signal is known, a distance $d'=c/f_{\text{Max}}$ can be chosen to avoid aliasing problems. which is actually greater than the maximum distance d allowed according to Eq. (8). Considering c = 345m/s and f_{Samp} = 16 kHz, the maximum frequency f_{Max} will be 8 kHz. Thus, the maximum distance d', that should be maintained between the microphones to avoid spectral aliasing is 4.31 cm. Here, d = 3 cm is chosen.

In order to calculate the reverberation time (t_{60}), a method based on time-stretched pulse (TSP) signals is implemented [6]. An approximate estimation of the acoustic transfer function (ATF) is possible by transmitting a time stretched pulse and convolving the sensor output with the filter having the inverse characteristics of same TSP. Reverberation time can be calculated by backward integration of the impulse response. Here, t_{60} of 300 ms is observed in the recorded room at the described loudspeaker setup.

The mixtures are designed with frequent speaker changes and with speaker overlaps during more than 50% of the time.

4.2 Results

The performance of the multi-speaker voice activity detection algorithm is assessed using false alarm rate (FAR), false rejection rate (FRJ), speaker error rate (SER) and total error rate (ERR). The FAR for each mixture is defined as follows:

$$\text{FAR} = \frac{\text{No. of frames falsely labeled as active}}{\text{Total no of frames}}. \tag{9}$$

The FRJ is determined by

$$\text{FRJ} = \frac{\text{No. of frames falsely rejected}}{\text{Total no of frames}} \tag{10}$$

and the SER is determined by

$$\text{SER} = \frac{\text{No. of frames falsely labelled as other speaker}}{\text{Total no of frames}}. \tag{11}$$

The total error is calculated by summing all above errors,

$$\text{ERR}=\text{FAR}+\text{FRJ}+\text{SER}. \tag{12}$$

The false alarm rate (FAR) gives an estimation of number of frames that are wrongly identified as active parts of a speaker, while the rate of false rejections (FRJ) shows, how

many actual active parts of the speaker are ignored by the algorithm and labelled as a noise or silence. The speaker error rate (SER) gives precisely the number of frames wrongly identified as an active part of a speaker, when actually another speaker is active. The parameter FRJ is more significant for speech recognizers, as rejection of frames or features of the desired speaker always deteriorates the performance of speech recognizers. It is usually required to be low for speech recognition applications. Similarly, a low FAR is crucial for algorithms which identify the channels of speakers in reverberant environments. Since channel adaptation methods fail to estimate channels correctly during the noisy or double talk intervals, a low value of FAR is desired.

Using the above performance parameters, results are presented in Table 1 and Table 2. Table 1 shows that the FAR of speaker I and speaker II are 39.9% and 54.75% respectively and FRJ is observed to be 3.7% and 7.2% when only VAD is performed on the recordings. Table 2 presents the actual results of the algorithm. FAR of speaker I and speaker II are observed to be 24% and 16% and the corresponding SERs are 1.95% and .03%. It can be observed that the average FAR decreased by 22.3% and the average speaker error rate has dropped to less than 1%. One can also notice that simple VAD has a very low FRJ when compared to the ICA-based algorithm, but at the expense of high speaker error rates which are more than 50% for both speakers, leading to error rates of 44.6% and 61.9% respectively. A moderate increase in the FRJ can be acceptable, considering the very low values of SER and overall error rate ERR.

Beampattern-based VAD was capable of detecting the active parts with considerable efficiency, even though the speaker overlap is greater than 50% and recordings are performed with microphones placed equally far from both sources, leading to an SIR of about 0 dB.

5 CONCLUSIONS

In this paper, a new approach for multi-speaker voice activity detection is presented. This beampattern-based multichannel VAD is intended for scenarios, in which not only voice activity detection but also speech separation is necessary, e.g. for subsequent speech recognition. When that is the case, beampattern-based VAD is an example of how source separation results can be further processed to obtain reliable voice activity detection, even when more than one speaker is active simultaneously. For this purpose, spatial features of demixing filters have proven essential in distinguishing the different speaker signals.

The resulting algorithm has been tested on highly reverberant recordings and found to have a lower error rate and much higher sensitivity than previously possible. Furthermore, it is shown to be almost perfectly selective with a speaker error rate less than 1%.

REFERENCES

[1] W. Baumann, D. Kolossa and R. Orglmeister, "Convolutive source separation based on a beamforming model," *Proc. ICASSP* 2003, pp.357-360, 2003.

[2] J. F. Cardoso and A. Suloumiac, "Blind beamforming for non Gaussian signals", *IEEE Proc,* Vol.140 (6), pp. 362-370, December 1993.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator", *IEEE Transactions on Acoustics, Speech and Signal processing,* Vol. 32, No. 6, pp.1109-1121, Dec. 1984.

[4] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions,"*Proc. ICASSP* 2000, Vol.5, pp. 3140-3143, 2000.

[5] L.R. Rabiner and M.R. Sambur "Algorithm for determining the endpoints of isolated utterances". *The Bell System Technical Journal,*Vol. 54, No. 2, pp. 297-315, 1975.

[6] N. Aoshima, "Computer-generated pulse signal applied for sound measurement", *J.Acoust. Soc. Am.* 69, 1484-1488 (1981).