ENERGY-BASED MULTI-SPEAKER VOICE ACTIVITY DETECTION WITH AN AD HOC MICROPHONE ARRAY

Alexander Bertrand*, Marc Moonen

Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
E-mail: alexander.bertrand@esat.kuleuven.be; marc.moonen@esat.kuleuven.be

ABSTRACT
In this paper, we propose an energy-based technique to track the power of multiple simultaneous speakers using an ad hoc microphone array with unknown simultaneous microphone positions. By considering the short-term power of the microphone signals, the problem can be converted into a non-negative blind source separation (NBSS) problem. By exploiting the prior knowledge that the source signals are non-negative and well-grounded, very efficient algorithms can be used to solve this NBSS problem, based only on second order statistics. We provide simulation results that demonstrate the effectiveness of the presented algorithm.

Index Terms— Signal detection, Random arrays, Voice activity detection

1. INTRODUCTION
Many speech processing algorithms make use of a voice activity detector (VAD), i.e., an algorithm that decides whether a speech source is active or not. However, most VAD’s assume that there is a single speech source, and are therefore unreliable in scenario’s with multiple speakers. Furthermore, it is sometimes desirable that the VAD is able to distinguish between different speakers, e.g., in noise reduction algorithms where the noise signal is a speaker that interferes with the target speaker.

Since different speakers have different positions, the design of a multi-speaker VAD can rely on spatial information collected by multiple microphones. In [1], a far-field multi-speaker VAD is proposed for a microphone array with known microphone positions. The algorithm uses independent component analysis (ICA), K-means clustering, and beam-pattern analysis, which makes it very complex. In this paper, we use an energy-based approach that does not exploit any prior knowledge on the geometry of the array. It is suited for applications that make use of an ad hoc microphone array with widely spaced microphones (e.g., [2,3]). This is for instance the case in video conferencing applications where each participant brings a device with built-in microphones, such as a laptop or PDA. Since most of these devices have WiFi technology, they can be linked to form an ad hoc network [2,4]. The presented algorithm also does not assume any accurate synchronization between the microphone sampling clocks, which is very convenient, e.g., in the mentioned scenario with different devices. The VAD algorithm provides an estimate of the instantaneous power of each speech signal at each microphone.

By using short-term power measurements at the different microphones, the multi-speaker VAD problem can be converted into a blind source separation problem with non-negative sources, which can be solved efficiently with second order statistics only. We provide simulation results to demonstrate the effectiveness of the presented algorithm.

2. PROBLEM STATEMENT AND DATA MODEL
Consider a scenario with $N$ speakers and an ad hoc microphone array with $J$ microphones. It is assumed that the microphones are spatially distributed such that the captured power from any speech source varies over the different microphones. We assume that the number of speakers $N$ is known. If not, a prior step is needed to estimate $N$ from the microphone signals, e.g., with PCA.

The $N$ speakers produce the speech signals $\tilde{s}_n[t]$, $n = 1 \ldots N$, where $t$ denotes the sample time index. Let $L$ denote the block length over which the instantaneous power of a signal is measured. We define the signal $s_n[k]$ as

$$s_n[k] = \frac{1}{L} \sum_{l=0}^{L-1} \tilde{s}_n[kL+l]^2$$

(1)

i.e. $s_n[k]$ contains the instantaneous power of the signal $\tilde{s}_n$ at sample time $kL$ ($k$ is a frame index). The $s_n[k]$ signals are stacked in an $N$-dimensional vector $s[k]$. In the sequel, we will use the symbol $s$ without the index $[k]$ to refer to the underlying random process that generates the samples $s[k]$. Similarly to (1), we define the instantaneous power in the $j$-th microphone signal as

$$y_j[k] = \frac{1}{L} \sum_{l=0}^{L-1} \tilde{y}_j[kL+l]^2$$

(2)

where $\tilde{y}_j[t]$ denotes the $j$-th microphone signal. The $y_j[k]$ signals are stacked in a $J$-dimensional vector $y[k]$.

If we assume that the signals $\tilde{s}_n$, $n = 1 \ldots N$, are mutually independent, and if we neglect reverberation effects over the block edges, we can model $y[k]$ according to

$$y[k] \approx As[k], \ \forall k \in \mathbb{N}$$

(3)
where $A$ is a $J \times N$ mixing matrix, for which the element $[A]_{jn}$ denotes the power attenuation between speaker $n$ and microphone $j$. It is assumed that the mixing matrix $A$ has full column rank. Notice that $L$ yields a trade-off between time resolution and model mismatch. The larger the value of $L$, the better the approximation (3) holds, but the worse the time resolution becomes. Furthermore, if there is significant reverberation, this will also affect the approximation (3) (especially when $L$ is small). However, we will demonstrate in section 4 that our VAD algorithm is still able to provide satisfying results under limited reverberation.

Our goal is to find both $A$ and $s[k]$, which would allow us to compute the instantaneous power of each speaker at each microphone, and then to run a VAD for each speaker separately. Notice that this is a blind source separation (BSS) problem in which the source signals are non-negative. In [5], this is referred to as a non-negative independent component analysis (NICA) problem. Expression (3) can also be described in the frequency domain to allow for a multi-speaker VAD in separate frequency bins. However, as with all frequency domain BSS problems, a post-processing stage must then be added to resolve the permutation ambiguity between the different frequency bins. We will not take this into consideration in this paper.

Notice that we did not incorporate any noise in the data model.

However, a localized noise source with non-stationary noise power, can readily be included in $s$ as an additional source signal. On the other hand, diffuse noise with stationary power results in a constant noise floor, which can be easily estimated and subtracted from $y[k]$. If required, noise estimation techniques, such as [6–8], can be used to track the power of a non-stationary diffuse noise. In the sequel, we assume that either noise power is subtracted from the signal or that localized noise sources are included in $s$, so that (3) is satisfied. In section 4, simulation results will demonstrate that the proposed VAD algorithm can still provide satisfying results when some residual noise power remains in $y[k]$. The residual noise then results in a non-zero noise floor on the unmixed signals.

3. SOLVING THE NON-NEGATIVE BSS PROBLEM

3.1. Well-grounded sources

The prior knowledge on the non-negativity of the source signals in $s$ can be exploited to design algorithms that are simpler compared to traditional ICA algorithms. In this paper, we exploit an additional assumption, i.e. the sources are assumed to be well-grounded [9]. This means that all sources have a non-zero pdf in any positive neighborhood of zero, i.e. $\forall \delta > 0: Pr(s_n < 0) > 0$, for all source signals $s_n$, $n = 1 \ldots N$. Because speech signals typically have an on-off behavior, the signals $s_n$, $n = 1 \ldots N$, can be assumed to be well-grounded.

In [5], the non-negative principal component analysis (NPCA) algorithm is introduced, which solves NICA problems with well-grounded source signals. NPCA is a gradient-based learning algorithm, and its performance heavily depends on the chosen learning rate, as we will demonstrate in section 4.

To avoid a step size search, we will use a multiplicative NICA (M-NICA) algorithm instead, which also exploits the well-grounded properties of the source signals [10]. M-NICA is a fixed-point type algorithm that has the facilitating property that it does not depend on a user-defined learning rate. In the next section, we will briefly describe M-NICA. Even though the simulation results of our speaker dependent VAD are performed in a real-time context, we will describe the algorithm in batch-mode, for the sake of an easy exposition. For a detailed description of an adaptive sliding window implementation of M-NICA, we refer to [10].

3.2. The M-NICA algorithm

Assuming that the source signals $s$ are non-negative and well-grounded, it can be shown that it is sufficient to find an $N \times J$ unmixing matrix $K$ such that the entries in the unmixed signal $\hat{s} = Ky$ are mutually uncorrelated and non-negative [9,10]. Therefore, M-NICA is entirely based on second order statistics.

Assume we collect a $J \times M$ data matrix $Y$ that contains $M$ samples $y[k], k = 0 \ldots M − 1$, in its columns. The goal is to find an $N \times M$ matrix $S = KY$ such that the rows of $S$ are uncorrelated and only contain non-negative numbers. The following fixed-point type algorithm is used to generate such a matrix [10]:

1. Initialization:

(a) $\forall n = 1 \ldots N, \forall m = 1 \ldots M: [S]_{nm} ← [Y]_{nm}$

(b) Replace $Y$ by its best rank $N$ approximation by means of

$$\{U, Σ, V\} ← \text{SVD}(Y)$$

where $Σ$ is the $N \times N$ diagonal matrix containing the $N$ largest singular values$^1$ of $Y$ on its diagonal, and where the corresponding left and right singular vectors are stored in the columns of $U$ and $V$ respectively.

2. Decorrelation step:

$$\forall n = 1 \ldots N, \forall m = 1 \ldots M :$$

$$[S^+]_{nm} ← [S]_{nm} \frac{SS^TΛ_1^{-1}S + SS^TΛ_1^{-1}S + A_2S}{SS^TΛ_1^{-1}S + SS^TΛ_1^{-1}S + A_2S}_{nm}$$

(6)

with

$$S = \frac{1}{M}S 1_M 1_M^T$$

$$C_z = (S − S)(S − S)^T$$

$$A_1 = D\{C_z\}$$

$$A_2 = D\left\{\left(Λ_1^{-1}C_z\right)^2\right\}$$

(7) (8) (9) (10)

where $1_M$ denotes an $M$-dimensional column vector in which each entry is 1, and where $D\{X\}$ denotes the operator that sets all off-diagonal elements of $X$ to zero.

3. Signal subspace projection step:

$$\forall n = 1 \ldots N, \forall m = 1 \ldots M :$$

$$[S]_{nm} ← \max\left(\left[\begin{array}{c} S^T \nabla \nabla^T \end{array}\right]_{nm}, 0 \right) .$$

(11)

4. Return to step 2.

In the decorrelation step (6), the elements of the matrix $S$ are updated to decrease the mutual correlation between the rows of $S$. Since $S$ is initialized with non-negative elements, the decorrelation step (6) will preserve the non-negativity due to its multiplicative nature. However, the rows of the resulting matrix $S$ are no longer in

$^1$Notice that, if noise were present, this step will remove some noise from the observations. In the noise-free case, $Y$ has exactly $N$ non-zero singular values.
the signal subspace defined by the rows of \( \mathbf{Y} \). Therefore, the matrix \( \mathbf{S} \) is projected to the row space of \( \mathbf{Y} \) in (11). For a more detailed derivation of the updating formulas, we refer to [10].

When a fixed point of (6)-(11) is found, the elements in each row of \( \mathbf{S} \) correspond to samples of the unmixed signal \( \hat{s}[k] \). The mixing matrix \( \hat{\mathbf{A}} \) that corresponds to \( \hat{s} \), can then be computed as

\[
\hat{\mathbf{A}} = \mathbf{Y} \mathbf{S}^T \left( \mathbf{S} \mathbf{S}^T \right)^{-1} .
\]

Notice that there always remains a permutation and scaling ambiguity between the columns of \( \hat{\mathbf{A}} \) and the signals in \( \hat{s} \). However, in the multi-speaker VAD application, we are interested in the speech energy of each target speaker in each microphone signal. Let \( v_{jn}[k] \) denote the speech energy of speaker \( n \) in microphone \( j \) at time instant \( k \). Each value \( v_{jn}[k], j = 1 \ldots J, n = 1 \ldots N, k = 1 \ldots M \) can then be estimated as

\[
\hat{v}_{jn}[k] = \left( \hat{\mathbf{A}} \right)_{jn} \hat{s}_n[k] .
\]

4. SIMULATIONS

In this section, we provide simulation results for the multi-speaker VAD algorithm based on M-NICA. To compare, we also provide simulation results for the case where (3) is solved with NPCA, with different learning rates \( \eta \) (for a description of this algorithm, we refer to [5]). We simulate a cubical room (5m \( \times \) 5m \( \times \) 5m) with \( N = 3 \) randomly placed speakers (\( \odot \)), all of them talking simultaneously, and \( J = 6 \) randomly placed microphones (\( \square \)), as shown in Fig. 1. The microphone signals are generated by means of the image method [11]. Unless stated otherwise, we compute the instantaneous energy of each target speaker in each microphone signal. Let

\[
\text{SER} = \frac{1}{JN} \sum_{j,n} 10 \log_{10} \left( \sum_k (\hat{v}_{jn}[k] - [\mathbf{A}]_{jn} s_n[k])^2 \right) ,
\]

where \( \hat{v}_{jn}[k] \) is defined by (13). Since we consider a sliding window implementation, the SER is computed over the \( K \) samples in the sliding window, and thus updated for each window shift.

We use the mean of the signal-to-error ratios (SER) to assess the performance of the multi-speaker VAD algorithm, i.e.

\[
\text{SER} = \frac{1}{JN} \sum_{j,n} 10 \log_{10} \left( \frac{1}{K} \sum_k (\hat{v}_{jn}[k] - [\mathbf{A}]_{jn} s_n[k])^2 \right) ,
\]

4. CONCLUSIONS

In this paper, we have presented a technique to track the power of multiple simultaneous speakers with an ad hoc microphone array with unknown microphone positions. Since the technique is energy-based, an accurate synchronization between the different microphone signals is not required. By using short-term power measurements at the different microphones, the multi-speaker VAD problem can be converted into a non-negative blind source separation (NBSS) problem, which can be solved efficiently based on second order statistics only. The effectiveness of the multi-speaker VAD algorithm decreases. We model residual noise by adding a stationary white noise source to each microphone signal. Each microphone signal has an equal amount of residual noise, and no noise power is subtracted from \( y[k] \). Fig. 3(b) shows the SER as a function of the signal-to-noise ratio (SNR) at the microphone with highest SNR. It is observed that the VAD algorithm still produces an output with satisfactory SER, as long as the SNR due to residual noise is sufficiently low. It should be noted that the decrease in SER is mainly due to a constant noise floor in the unmixed signals. The speech segments that have a higher power than this noise floor can still be detected, and are observed to be properly separated.
VAD has been demonstrated with adaptive sliding window simulations. The M-NICA algorithm presented here is observed to provide better overall results compared to NPCA [5], and has the additional advantage that it does not depend on a user-defined learning rate.

6. REFERENCES


