A NOVEL APPROACH TO AUTOMATED SOURCE SEPARATION IN MULTISPEAKER ENVIRONMENTS

Robert M. Nickel and Ananth N. Iyer

Department of Electrical Engineering
The Pennsylvania State University
University Park, PA 16802
rmn10@psu.edu, ani103@psu.edu

ABSTRACT

We are proposing a new approach to the solution of the cocktail party problem (CPP). The goal of the CPP is to isolate the speech signals of individuals who are concurrently talking while being recorded with a properly positioned microphone array. The new approach provides a powerful yet simple alternative to commonly used methods for the separation of speakers. It is based on the observation that the estimation of the signal transfer matrix between speakers and microphones is significantly simplified if one can assure that during certain periods of the conversation only one speaker is active while all other speakers are silent. Methods to determine such exclusive activity periods are described and a procedure to estimate the signal transfer matrix is presented. A comparison of the proposed method with other popular source separation methods is drawn. The results show an improved performance of the proposed method over earlier approaches.

1. INTRODUCTION

Up to date, the most successful approaches to solving the cocktail party problem (CPP) employ blind source separation (BSS) techniques based on an assumption of statistical independence of the sources [1, 2]. The goal is to find an unmixing system that maximizes a "measure of independence" from the reconstructed source signals. Generally, one distinguishes between methods that consider *instantaneous mixing* and methods that address *convolutive mixing* [1, 2]. In this paper we are laying the foundation for an alternative solution to the instantaneous mixing case. Extensions to convolutive mixing are explored in a separate study [3].

The crux of the aforementioned approach is to find a good (and mathematically tractable) "measure for independence." It was shown that a suitable objective in finding a solution is provided by the minimization of a *contrast function* which is a function of the PDF of the observed signals [2]. In the context of speech and audio signal processing a suitable contrast function is usually derived from a likelihood measure and/or the INFOMAX concept [4]. The optimization procedure that minimizes a contrast function (and thus provides a solution to the BSS problem) is generally called an independent component analysis (ICA) [2]. Even though the BSS techniques have achieved remarkable results in the separation of mixed audio signals, they are still suboptimal in regard of separation of speech, partially because they usually do not permit an exploitation of the highly structured nature of speech.

The alternative method for speech separation that is presented in this paper is not based on an assumption of independence but is merely based on a simple observation of the source signals. The estimation of the transfer matrix or its inverse becomes trivial if we are able to assert that during certain periods of time only one source is active and all other sources are silent. If during the total observation time each source went at least once through such an *exclusive activity period* (EAP) then the full transfer matrix can be estimated. Conversational speech is generally characterized by prolonged pauses from individual speakers and hence the probability of exclusive activity periods for individual speakers is very high, especially if we are restricting ourselves to small groups.

The caveat of the newly proposed method is that one needs to reliably detect exclusive activity periods from the observed signals. The detection of such periods becomes possible if we are focusing on the unique structure of speech signals and in particular the unique structure of voiced segments of the speech signals.

The paper is organized as follows: section 2 introduces setup and notation for the considered BSS problem. Section 3 summarizes the employed methods for EAP detection and the estimation of the resulting de-mixing matrix. An experimental evaluation of the proposed method is presented in sections 4 and 5.

2. BACKGROUND AND NOTATION

The considered scenario is described by the following mathematical model: we have K speakers in a room, each of which produces a speech signal $s_k[n]$ (for $k = 1, 2, \ldots, K$). The K speech signals are captured by M microphones ($M \geq K$), each of which must be placed such that the acoustic wave-

forms measured must be significantly different from microphone to microphone[1]. The signal recorded at microphone m will be referred to as the observed signal and is denoted as $x_m[n]$ (for $m = 1 \ldots M$). To streamline the notation it is beneficial to introduce the following row-vectors:

$$\mathbf{s}_k = [\, s_k[1]\,,\, s_k[2]\,,\, \ldots\,,\, s_k[P]\,]^{\mathrm{T}} \quad (1)$$

$$\mathbf{x}_m = [\, x_m[1]\,,\, x_m[2]\,,\, \ldots\,,\, x_m[P]\,]^{\mathrm{T}}, \quad (2)$$

where P is the observation segment length in samples. For simplicity we will use the notation $\mathbf{s}_k$ and $\mathbf{x}_m$ in three ways: (i) to indicate vectors that encompass the entire recording length (see section 4), (ii) to indicate one of a successive set of 40 msec long segments of the recording (see section 3.1), and (iii) to indicate an exclusive activity period which spans multiple successive segments of length 40msec (see section 3.2). Matrices are formed from the vectors as $\mathbf{S} = [\, \mathbf{s}_1\,,\, \mathbf{s}_2\,,\, \ldots\,,\, \mathbf{s}_K\,]^{\mathrm{T}}$ and $\mathbf{X} = [\, \mathbf{x}_1\,,\, \mathbf{x}_2\,,\, \ldots\,,\, \mathbf{x}_M\,]^{\mathrm{T}}$. If we assume instantaneous mixing then the connection between the source signal matrix $\mathbf{S}$ and the observed signal matrix $\mathbf{X}$ is determined by the constant mixing matrix $\mathbf{A}$:

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

In general, the solution to the CPP is defined so as to determine an inverse/de-mixing matrix $\mathbf{W}$ such that matrix $\hat{\mathbf{S}}$ from

$$\hat{\mathbf{S}} = \mathbf{W}\,\mathbf{X} \quad (4)$$

provides an estimate for the rows of matrix $\mathbf{S}$. The matrix $\mathbf{W}$ is considered a de-mixing matrix when the matrix product $\mathbf{W}\,\mathbf{A}$ is a permutation matrix.

3. METHODS

The proposed source separation algorithm consists of two main parts: i) the detection of EAP segments and ii) the determination of the de-mixing matrix $\mathbf{W}$. A block diagram of the proposed algorithm is shown in figure 1. The details of the method are described in sections 3.1 and 3.2.

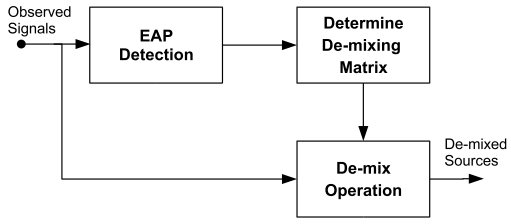

Fig. 1. A block diagram of the proposed source separation algorithm. The de-mixing matrix $\mathbf{W}$ is estimated from EAP sections of the observed signals $\mathbf{X}$. The source estimates $\hat{\mathbf{S}}$ are obtained after equation (4).

[1]Appropriate placement of the microphones is crucial to ensure that the resulting estimation problem is not being ill-conditioned.

3.1. EAP Detection

The detection of EAPs is facilitated by the following definition of a normalized signal-to-interference ratio (SIR) γ:

$$\gamma = \tfrac{1}{K-1}\, \max_k \left(K\, \frac{\mathbf{s}_k^{\mathrm{T}}\,\mathbf{s}_k}{\sum_{i=1}^{K} \mathbf{s}_i^{\mathrm{T}}\mathbf{s}_i} - 1 \right). \quad (5)$$

If we let the $\mathbf{s}_k$'s denote successive 40 msec long segments of the source signals then we obtain an SIR measure γ for each segment. Note that the SIR is normalized such that $0 \le \gamma \le 1$ (with $\gamma = 1$ indicating a perfect EAP event). Unfortunately, we cannot access the true underlying SIR since we cannot measure the source signals $\mathbf{s}_k$ directly. Instead, we are estimating γ from the observations $\mathbf{x}_m$. The proposed estimator is based on the following three features:

(i) **A Periodicity Measure.** The pitch determination algorithm (PDA) developed by Medan *et. al.* [5] aims to determine the pitch of a voiced speech segment. The method involves computing normalized inner products between adjacent, variable length speech segments. The inner product is maximized when the segment length equals the pitch period. The normalized inner product at the pitch provides a measure of periodicity f_p for the given signal segment $\mathbf{x}_m$.

(ii) **The Harmonic-to-Signal Ratio.** The energy of a voiced speech segment is concentrated around the harmonic frequencies of its pitch. The harmonic-to-signal ratio (HSR) f_h is defined as the ratio of spectral energy around the pitch harmonics versus the overall energy of signal $\mathbf{x}_m$. Its computation is achieved by comparing the energy of the output of a pitch synchronous comb-filter with the total signal energy [6].

(iii) **The Spectral Autocorrelation.** Spectral autocorrelation measures are used in many PDAs (in addition to temporal correlation measures) to combat pitch-doubling/pitch-halving errors. We are employing the normalized spectral autocorrelation proposed in [6]. The maximum autocorrelation value f_s in the pitch frequency range (50Hz to 500Hz) is determined and used as a feature for the SIR estimation.

The SIR is estimated via $\hat{\gamma} = \Phi(f_p, f_h, f_s)$ in which $\Phi(\ldots)$ is a three dimensional, second order polynomial. The optimal polynomial coefficients are chosen to minimize the least-squares error between the estimated SIR (ESIR) $\hat{\gamma}$ and the true underlying SIR γ for the training set described in section 4. The resulting normalized correlation coefficients between f_p, f_h, f_s, $\hat{\gamma}$ and γ are shown in table 1. The EAP detection is performed with a threshold test. A segment is flagged as an EAP if $\hat{\gamma}$ is greater or equal to a threshold $\bar{\gamma}$ (see section 4). A time segment is considered an exclusive activity period if the corresponding segments $\mathbf{x}_m$ are *all* flagged as EAPs across all channels $m = 1 \ldots M$.

Table 1. Correlation Coefficients for SIR Estimation

Feature	Correlation with γ
Periodicity Measure f_p	0.5838
Harmonic-to-Signal Ratio f_h	0.4360
Spectral Autocorrelation f_s	0.4329
Polynomial Estimate $\hat{\gamma}$	0.6338

3.2. Estimation of the De-mixing Matrix

During a true exclusive activity period it is readily verified that the observed signals $\mathbf{x}_i$ are scalar multiples of the one active source signal $\mathbf{s}_q$. We can obtain an estimate for $\mathbf{s}_q$ from each channel i by multiplying the channel output $\mathbf{x}_i$ with a channel specific scalar w_i:

$$\hat{\mathbf{s}}_i^q = w_i \, \mathbf{x}_i. \tag{6}$$

Ideally, we want $\hat{\mathbf{s}}_i^q = \hat{\mathbf{s}}_j^q$ for all i and j, which leaves us with an infinite number of choices for the w_i's if all of the $\mathbf{x}_i$'s are truly linearly dependent. Due to imperfect EAP detection and background noise, however, we are generally not able to achieve $\hat{\mathbf{s}}_i^q = \hat{\mathbf{s}}_j^q$ for all i and j for any set of channel weights w_i. Instead, we may seek to choose the w_i's such that the deviation between each $\hat{\mathbf{s}}_i^q$ and their average $\bar{\mathbf{s}}^q = \frac{1}{M}\sum_{j=1}^{M} \hat{\mathbf{s}}_j^q$ is minimized, i.e.

$$\underset{w_i}{\text{minimize}} \; \mathcal{C}^q = \sum_{i=1}^{M} \| \, \hat{\mathbf{s}}_i^q - \bar{\mathbf{s}}^q \, \|^2 \text{ subject to } \| \, \bar{\mathbf{s}}^q \, \|^2 = \zeta^2. \tag{7}$$

The constraint is necessary to avoid the trivial (yet meaningless) solution $w_i = 0$ for all i. Expanding the terms of the cost function leads to

$$\mathcal{C}^q = \sum_{i=1}^{M} w_i^2 \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i - \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{M} w_i w_k \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_k, \tag{8}$$

which can be compactly written in matrix notation as

$$\mathcal{C}^q = \mathbf{w}^{\mathrm{T}} [\, \mathbf{R}_d - \tfrac{1}{M} \mathbf{R} \,] \, \mathbf{w} \tag{9}$$

with $\mathbf{R} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, $\mathbf{R}_d = \mathrm{diag}(\mathbf{R})$ and $\mathbf{w} = [\, w_1 \, , \, w_2 \, , \, \ldots \, , \, w_M \,]^{\mathrm{T}}$. With Lagrange multiplier λ we can cast equation (7) into the Lagrange function

$$L(\mathbf{w}, \lambda) = \mathbf{w}^{\mathrm{T}} [\, M \, \mathbf{R}_d - (1 + \lambda) \, \mathbf{R} \,] \, \mathbf{w} - \lambda \, M^2 \, \zeta^2. \tag{10}$$

Differentiating $L(\mathbf{w}, \lambda)$ with respect to $\mathbf{w}$ and equating to zero results in the condition

$$\mathbf{R}_d \, \mathbf{w} = \tfrac{\lambda + 1}{M} \, \mathbf{R} \, \mathbf{w}. \tag{11}$$

Condition (11) is satisfied by the generalized eigenvectors of matrices $\mathbf{R}$ and $\mathbf{R}_d$ (with $\frac{\lambda+1}{M}$ being the generalized eigenvalues). It is readily shown that out of the M generalized eigenvectors we must choose the one $\mathbf{w}^q$ with the smallest eigenvalue to minimize the cost function $\mathcal{C}^q$.

The transpose of the solution $\mathbf{w}^q$ establishes the q's row of the de-mixing matrix $\mathbf{W}$. We must observe at least one EAP from every source to obtain a full estimate of $\mathbf{W}$. The decision of whether a set of disjoint EAPs belongs to the same source or to different sources is done with a simple hierarchical clustering algorithm[2]. If multiple estimates of $\mathbf{w}^q$ (from disjoint EAPs) are cast into the same cluster then their (renormalized) centroid is used as the corresponding row in $\mathbf{W}$.

4. EXPERIMENTS

We evaluated the performance of the proposed method with mixing/de-mixing trials over speech data from the TIMIT[3] database. The TIMIT dataset contains recordings of 10 phonetically rich sentences from 630 speakers of 8 major dialects of American English. The corpus is stored in 16bit/16kHz waveform files for each utterance. One subset of files is strictly reserved for training and another subset is reserved for testing.

During each mixing/de-mixing trial we randomly chose M utterances $\mathbf{s}_i$ from the corpus. The M utterances were mixed with a random mixing matrix $\mathbf{A}$ according to equation (3) to produce M observations $\mathbf{x}_i$. The elements of $\mathbf{A}$ were chosen as independent, uniformly distributed random numbers over the interval $[0, 1]$. The observations $\mathbf{x}_i$ were then subjected to the proposed de-mixing procedure to produce M source estimates $\hat{\mathbf{s}}_i$.

The quality of the de-mixing process was measured with the following signal-to-noise ratio (SNR):

$$\mathrm{SNR} = \min_{\mathrm{P} \in \mathbb{P}} \tfrac{10}{M} \sum_{[i,j] \in \mathrm{P}} \log \left[\frac{\mathbf{s}_i^{\mathrm{T}} \mathbf{s}_i}{(\mathbf{s}_i - \eta \, \hat{\mathbf{s}}_j)^{\mathrm{T}} (\mathbf{s}_i - \eta \, \hat{\mathbf{s}}_j)} \right]. \tag{12}$$

The scaling factor η is chosen as $\eta = \hat{\mathbf{s}}_i^{\mathrm{T}} \mathbf{s}_j / (\hat{\mathbf{s}}_i^{\mathrm{T}} \hat{\mathbf{s}}_i)$ to account for the unknown scaling of the reconstructed signals. Parameter P refers to a set of M index pairs $[i, j]$ with $i = 1 \ldots M$ and $j = 1 \ldots M$ and such that each number between 1 and M is only used once for i and once for j. $\mathbb{P}$ is the entirety of all possible index sets P. The minimization over P accounts for the unknown signal permutations introduced by $\mathbf{WA}$ as discussed in section 2.

In a first experiment we ran several sets of 256 random mixing/de-mixing trials with $M = 2$ over the *training subset* of the corpus. For each set of 256 trials we chose a different EAP decision threshold $\bar{\gamma}$ as described in section 3.1. The resulting average SNR (averaged over all trials) as a function of $\bar{\gamma}$ is displayed in figure 2. The optimal threshold, i.e. the one that produced the highest average SNR, was found to be $\bar{\gamma} = 0.85$.

[2]It is assumed that the underlying number of sources is known and that all sources have at least one EAP.

[3]The TIMIT database is available through the *Linguistic Data Consortium* (LDC) at the University of Pennsylvania (www.ldc.upenn.edu).

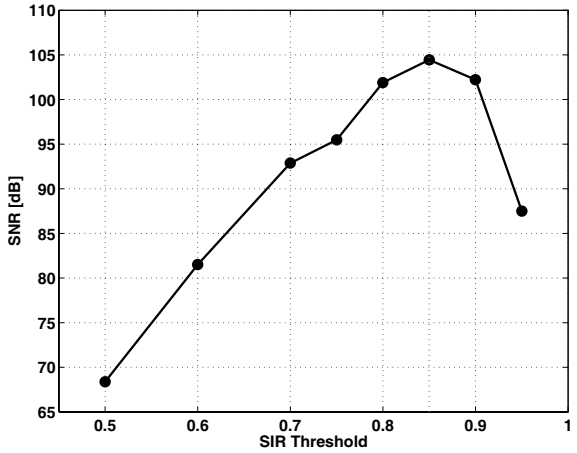

Fig. 2. Determination of the EAP decision threshold $\bar{\gamma}$. The average SNR is maximized in the two channel case ($M = 2$) for $\bar{\gamma} = 0.85$.

In a second experiment we ran several sets of 256 random mixing/de-mixing trials over the *testing subset* of the corpus. This time, we kept $\bar{\gamma}$ fixed at 0.85 and varied the number of sources/channels M. The resulting average SNR (averaged over all 256 trials per M) as a function of M is shown in figure 3 (EAPD).

5. RESULTS

The performance of the proposed exclusive activity period detection (EAPD) algorithm and two other popular BSS methods, FastICA[4] [7] and AMUSE[4] [8], are presented in figure 3 for comparison. For each method the average SNR and the standard deviation over all 256 trials per M are shown (via I-bars).

The proposed EAPD method clearly outperforms the other two methods for smaller source numbers. As the number of simultaneously talking sources M increases, the number (and quality) of available EAP sections naturally decreases. As a result, the performance of the proposed method declines. All methods performed around the same average SNR for 6 and more simultaneous sources (with a slight edge of FastICA and EAPD over AMUSE). Increased errors in EAP detection, however, lead to a much larger standard deviation of the EAPD method in comparison to FastICA and AMUSE at higher source numbers.

6. CONCLUSIONS

We presented a new approach to the solution of the cocktail party problem with instantaneous mixing. Instead of insisting on independence between sources and samples, we exploited the fact that speech is generally characterized by frequent pauses (EAPs). These pauses can be used for a *one-channel-at-a-time* estimation of the unknown mixing matrix. Experiments have shown that the proposed method can outperform common BSS methods, especially for smaller source numbers.

[4]We employed software provided by the original developers of the FastICA and AMUSE algorithms for the comparison. All third part software was run with default parameters.

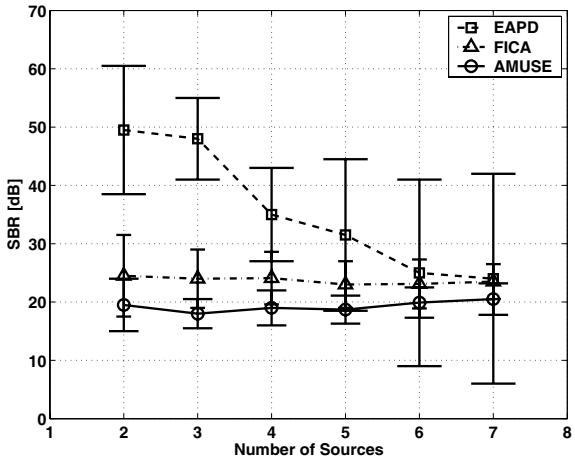

Fig. 3. An SNR comparison between the proposed method (EAPD) and two other popular BSS methods: FastICA (FICA) [7] and AMUSE [8].

It should be noted that the presented simulation results were obtained from (unrealistically) harsh conditions for the EAP detection. All speakers were talking at the exact same time. The resulting EAP sections were, hence, short and generally of poor quality. More realistically, speakers tend to listen and respond in a dialog which makes the detection of long, high quality EAP section much more probable.

7. REFERENCES

[1] N. Mitianoudis and M. E. Davies, "Audio source separation: solutions and problems," *International Journal of Adaptive Control and Signal Processing*, vol. 18, no. 3, pp. 299–314, Apr. 2004.

[2] T. W. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer Academic Publishers, 1998.

[3] R. M. Nickel, "Blind multichannel system identification with applications in speech signal processing," *Proc. Int. Conf. on Comp. Intell. for Modelling Control and Automation (CIMCA), 2005, Vienna, Austria*, Nov. 2005.

[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*, John Wiley & Sons., 2001.

[5] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39-1, pp. 40–48, January 1991.

[6] A. M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, John Wiley & Sons, 2 edition, November 2004.

[7] Hyvärinen A., "A family of fixed-point algorithms for independent component analysis," in *IEEE Int. Conf. on Acoustics, Speech Signal Processing (ICASSP'97)*, 1997.

[8] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley, New York, 2003.