ROBUST FEATURE SELECTION FOR SCALING AMBIGUITY REDUCTION IN AUDIO-VISUAL CONVOLUTIVE BSS

Qingju Liu[*], *Syed Mohsen Naqvi*[†], *Wenwu Wang*[*], *Philip Jackson*[*], *Jonathon Chambers*[†]

[*]Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering
University of Surrey, Guildford, UK
Emails: {Q.Liu, W.Wang, P.Jackson}@surrey.ac.uk

[†]Advanced Signal Processing Group, Department of Electronic and Electrical Engineering,
Loughborough University, Loughborough, UK
Emails: {S.M.R.Naqvi, J.A.Chambers}@lboro.ac.uk

ABSTRACT

Information from video has been used recently to address the issue of scaling ambiguity in convolutive blind source separation (BSS) in the frequency domain, based on statistical modeling of the audio-visual coherence with Gaussian mixture models (GMMs) in the feature space. However, outliers in the feature space may greatly degrade the system performance in both training and separation stages. In this paper, a new feature selection scheme is proposed to discard non-stationary features, which improves the robustness of the coherence model and reduces its computational complexity. The scaling parameters obtained by coherence maximization and non-linear interpolation from the selected features are applied to the separated frequency components to mitigate the scaling ambiguity. A multimodal database composed of different combinations of vowels and consonants was used to test our algorithm. Experimental results show the performance improvement with our proposed algorithm.

1. INTRODUCTION

Multi-channel frequency domain BSS [1, 6, 15, 12, 10] has been extensively used for separating audio sources from their convolutive mixtures. Typically, independent component analysis (ICA) [3] techniques are applied to solve the instantaneous model in each frequency channel. However, to reconstruct the original sources, the separated components at each frequency bin need to be grouped and scaled correctly to the corresponding sources. These are the well-known problems in frequency domain BSS: permutation and scaling ambiguities. Many methods have been developed to solve the permutation problem [6, 15, 12, 10], while little work has been done to address the scaling ambiguity problem. A minimal distortion principle (MDP) [9] technique has been usually applied to mitigate the scaling problem.

All these approaches are applied in the audio domain. It has been shown, however, that at least human speech is bimodal, and the cross-modal interactions (also referred as the audio-visual coherence) can be exploited to improve the intelligibility of human speech embedded in cocktail party noise [16]. Such bimodality has been used in several recent studies for blind source separation [17, 18, 14, 2, 7, 8, 11]. In a previous work, we have presented a method of using the bimodality to reduce the scaling ambiguity [7]. It mainly contains two independent stages. In the off-line training stage, first, audio-visual features are extracted respectively and synchronized to form the audio-visual space; then the Gaussian mixture models (GMMs) are applied to statistically characterize the audio-visual coherence. In the separation stage, after applying ICA in each frequency channel, we calculate the scaling parameters using the audio-visual coherence, and scale the ICA-separated frequency components with those parameters to mitigate the scaling ambiguity.

However, outliers in the feature space may greatly affect the GMMs used in both the training stage and the separation stage. Using more kernels could potentially mitigate this problem but involves a higher computational complexity and the overfitting problem (i.e. more parameters often resulting in mismatch between the model and the data). Also, the final scaling parameters obtained are not smooth enough which may result in a drastic change in adjacent frequency channels of the global filter. To overcome these limitations, we present the following new contributions:

- Instead of using all the features, we choose only *robust* features with a novel frame selection scheme.
- The scaling parameters are obtained with a more flexible method using non-linear interpolation.

The remainder of the paper is organised as follows. Section 2 introduces the multi-channel frequency domain BSS. The feature extraction, selection and fusion methods are presented in Section 3. Calculation of scaling parameters is shown in Section 4. Section 5 presents experimental results, followed by the conclusions.

2. FREQUENCY-DOMAIN BSS AND ITS ASSOCIATED AMBIGUITIES

Suppose P observations $x_p(n)$ are recorded from K source signals $s_k(n)$ in a cocktail party scenario:

$$x_p(n) = \sum_{k=1}^{K} \sum_{m=0}^{+\infty} h_{pk}(m) s_k(n-m) + \xi_p(n), \qquad (1)$$

where h_{pk} represents the room impulse response filter from source k to sensor p, n is the discrete time index and $\xi_p(n)$ is the additive noise, ignored for simplicity in this work. The objective of source separation is to find a set of separation filters $\{w_{kp}\}$ that satisfy:

$$y_k(n) = \sum_{p=1}^{P} \sum_{m=0}^{+\infty} w_{kp}(m) x_p(n-m), \qquad (2)$$

where $y_k(n)$ is the recovered source signal.

© EURASIP, 2011 - ISSN 2076-1465

In frequency domain BSS, ICA algorithms for instantaneous mixtures are independently applied to the spectral components $\boldsymbol{X}(f,t)$ in each frequency bin:

$$\boldsymbol{Y}(f,t) = \mathbf{W}(f)\boldsymbol{X}(f,t), \tag{3}$$

where $\boldsymbol{X}(f,t)$ is the observation vector in frequency channel f at time-frame index t after performing the short time Fourier transform (STFT), and $\mathbf{W}(f)$ is the separation matrix. $\boldsymbol{Y}(f,t)$ is considered as a copy of $\boldsymbol{S}(f,t)$, only up to a permutation matrix $\mathbf{P}(f)$ and a diagonal matrix of gains $\mathbf{D}(f)$:

$$\boldsymbol{Y}(f,t) = \mathbf{P}(f)\mathbf{D}(f)\boldsymbol{S}(f,t). \tag{4}$$

These are the so-called permutation ($\mathbf{P}(f)$) and scaling ($\mathbf{D}(f)$) indeterminacies.

In this paper, we only consider the scaling ambiguity, and suppose the permutation problem is addressed and no global permutation occurred, i.e., all components of $Y_k(f,t)$ come from $s_k(n)$. The scaling ambiguity means $Y_k(f,t)$ is scaled with different gains in different bins f, $Y_k(f,t) = d_{kk}(f)S_k(f,t)$ where $d_{kk}(f)$ is the k-th diagonal component of $\mathbf{D}(f)$. Therefore, if we reconstruct $y_k(n)$ in the time domain, it is an FIR-filtered version of the source signal $s_k(n)$. To solve this problem, MDP [9] can be applied:

$$\mathbf{W}(f) \Leftarrow diag(inv(\mathbf{W}(f)))\mathbf{W}(f), \tag{5}$$

assuming the recovered source signal $y_k(n)$ is the received component at the k-th sensor from $s_k(n)$, which is still a filtered version of $s_k(n)$. Suppose we are just interested in the first target signal $y_1(t)$, we need to estimate a series of scaling parameters $\{\alpha(f) = 1/d_{11}(f)\}_f$ to update $Y_1(f,t) \Leftarrow \alpha(f)Y_1(f,t)$. The audio-visual coherence, described in the following sections, can be used for this problem.

3. FEATURE EXTRACTION, SELECTION AND FUSION

The relationship between the audio and visual streams can be modeled in the feature space with features extracted and fused as follows.

3.1 Audio-Visual Feature Space

- Audio feature
  The Mel-scale filterbank analysis is applied to obtain the audio feature. Denote $\mathscr{F}_l, l = 1,2,...,L$ as the group of frequency bins spanned by the l-th filter. We map the spectral power into these filters respectively to obtain $a_{\mathbf{T}l}(t)$:
  $$a_{\mathbf{T}l}(t) = \sum_{f\in\mathscr{F}_l} b_l(f)|S_{\mathbf{T}}(f,t)|^2, \tag{6}$$
  where $_{\mathbf{T}}$ denotes training and $b_l(f)$ is the magnitude of the l-th filter while $S_{\mathbf{T}}(f,t)$ is the spectral component of the training audio.
- Visual feature
  We use the same front geometric visual features as in [17, 14, 8]: the lip width (LW) and height (LH) from the internal labial contour. It is low-dimensional and therefore is robust for statistical characterization. 2-dimensional visual feature $\mathbf{v}_{\mathbf{T}}(t) = [\mathrm{LW}(t), \mathrm{LH}(t)]^T$ is extracted from the training video, and ^{T} is the transpose.
- Audio-visual feature
  We get L sets of 3-dimensional audio-visual vectors $\mathbf{u}_{\mathbf{T}l}(t) = [\mathbf{v}_{\mathbf{T}}(t)^T, a_{\mathbf{T}l}(t)]^T$ by concatenating each audio feature $a_{\mathbf{T}l}(t)$ with the visual features $\mathbf{v}_{\mathbf{T}}(t)$, corresponding to each filter.

3.2 Proposed Robust Feature Selection

Most works for multimodel fusion [1, 2, 7, 8, 14, 17, 18] employ all the extracted features. As a result, any outliers may greatly affect the fusion results. In addition, the computational complexity is high. To improve the robustness and accuracy of the estimation of the audio-visual coherence, we use a new frame selection scheme based on the dynamic characteristics of the visual feature.

At each time frame centered by the visual feature $\mathbf{v}(t) = [\mathrm{LW}(t), \mathrm{LH}(t)]^T$, we extract a short time period with $2Q+1$ frames, then calculate

$$\gamma_{\mathrm{LW}}(t) = \sigma(\mathrm{LW}(t)) + \alpha_{\mathrm{LW}}\|\mathrm{LW}(t+Q) - \mathrm{LW}(t-Q)\|, \tag{7}$$

where $\|\cdot\|$ is the Euclidean distance, $\sigma(\cdot)$ is the standard deviation over $2Q+1$ frames and α_{LW} is a weighting coefficient, chosen between 0 and 1, which weights the influence of the overall changing trend of the short time interval. Then we define a Boolean variable to determine the stationarity of this frame

$$\mathscr{S}_{\mathrm{LW}}(t) \stackrel{\mathrm{def}}{=} \begin{cases} 1, & \gamma_{\mathrm{LW}}(t) < \delta_{\mathrm{LW}}\overline{\mathrm{LW}(t)} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where δ_{LW} is a comparison coefficient, typically chosen as 0.5, and $\overline{\mathrm{LW}(t)}$ is the mean over the $2Q+1$ frames. Then we smooth between adjacent frames

$$\mathscr{S}_{\mathrm{LW}}(t) = \mathscr{S}_{\mathrm{LW}}(t-1) + \mathscr{S}_{\mathrm{LW}}(t) + \mathscr{S}_{\mathrm{LW}}(t+1), \tag{9}$$

where $+$ is a logical OR operator. In the same way, we can determine $\mathscr{S}_{\mathrm{LH}}(t)$, and the final decision is

$$\mathscr{S}(t) = \mathscr{S}_{\mathrm{LW}}(t) \cdot \mathscr{S}_{\mathrm{LH}}(t), \tag{10}$$

where $\cdot$ denotes logical conjunction.

If $\mathscr{S}(t) = 1$, the frame will be chosen, otherwise it will be discarded. The audio-visual features associated with the selected frames are used in both the training and separation stages.

This frame selection scheme effectively removes the transient period from one syllable to another. The drastic change of visual parameters in the transient period results in outliers. For example, in several frames of the transition process from /a/ to /b/, the mouth shape may look similar to that of the utterance of /o/, and those frames may be *misclassified* as the the kernel related to the utterance of /o/. The proposed frame selection scheme has mitigated this problem.

3.3 Feature Fusion

The audio-visual coherence of each filter can be statistically characterized as a GMM model with I kernels and we use $\mathscr{N}$ to denote the Gaussian distribution:

$$p_{AV}(\mathbf{u}_{\mathbf{T}l}(t)) = \sum_{i=1}^{I} \gamma_{li}\mathscr{N}(\mathbf{u}_{\mathbf{T}l}(t) \mid \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}), \tag{11}$$

where γ_{li} is the weighting parameter, $\boldsymbol{\mu}_{li} = [\mu_{li1}, \mu_{li2}, \mu_{li3}]^T$ is the mean vector and $\boldsymbol{\Sigma}_{li} = diag([\sigma_{li1}, \sigma_{li2}, \sigma_{li3}])$ is the diagonal covariance matrix of the i-th kernel associated with the

l-th filter. Each kernel is a multi-variate Gaussian distribution. We denote $\lambda_{li} = \{\gamma_{li}, \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}\}$ as the parameter set, with $\{\lambda_{li}\}$ estimated by the expectation maximization algorithm in the off-line training process.

4. SCALING AMBIGUITY CANCELLATION WITH NON-LINEAR INTERPOLATION

In our early work in [7], we have proposed the scaling ambiguity cancellation algorithm from the coherence maximization point of view. Suppose there are L filters in the feature extraction process, we obtain L scaling parameters $\alpha(\mathscr{F}_l)$ where $\mathscr{F}_l$ is the l-th group of frequency bins spanned by the l-th Mel-scale filter:

$$\alpha(\mathscr{F}_l) = \sqrt{\sum_t a_l^{\dagger}(t) / \sum_t a_l(t)}, \tag{12}$$

where $a_l(t)$ is the audio feature extracted from $Y_1(f,t)$ by equation (6) corresponding to the l-th filter, and

$$a_l^{\dagger}(t) = \sum_{i=1}^{I} c_{li}(t)\mu_{li3}, \tag{13}$$

where μ_{li3} is the third element in the mean vector $\boldsymbol{\mu}_{li}$ in equation (11) and

$$c_{li}(t) = \frac{\gamma_{li}\mathcal{N}(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV})}{p_V(\mathbf{v}(t) \mid l)}. \tag{14}$$

$\boldsymbol{\mu}_{liV} = [\mu_{li1}, \mu_{li2}]^T$, $\boldsymbol{\Sigma}_{liV} = diag([\sigma_{li1}, \sigma_{li2}])$ and $p_V(\mathbf{v}(t) \mid l)$ is the visual marginal distribution with the l-th filter:

$$p_V(\mathbf{v}(t) \mid l) = \int_{\mathbf{a}(t)} p_{AV}(\mathbf{u}_l(t)) d\mathbf{a}(t). \tag{15}$$

Since the visual distribution and audio distribution in each kernel are independent (the covariance matrix is diagonal), we have

$$p_V(\mathbf{v}(t) \mid l) = \sum_{i=1}^{I} \gamma_{li}\mathcal{N}(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV}). \tag{16}$$

We get L scaling parameters related to L filters. However, we need to scale $Y_1(f,t)$ in each frequency bin f. In [7], we simply smooth between the L scaling parameters with linear interpolation to obtain M scaling parameters $\alpha(f)$, where M is the number of frequency channels. However, the resulting parameters are not *smooth* enough. Therefore, here we apply a piecewise cubic Hermite interpolating polynomial (PCHIP) interpolation, which ensures the monotonicity and contains no extraneous "bumps" or "wiggles" [4].

5. EXPERIMENTAL RESULTS

5.1 Data

The corpus[1] used in our research contains sequences of "V1-C-V2", where "V1" and "V2" are vowels from /a/, /i/ ,/o/, /u/, and "C" stands for the consonant from /p/, /t/, /k/, /b/, /d/, /g/ or no plosive (in the case of no plosive, the sequences are "V1-V2"). There are 112 combinations recorded twice, one for training and another for testing. The audio sequences are sampled at 16 kHz while the video sampling rate is 50 Hz and the associated visual features are extracted by a chrome based system. We concatenate the 112 isolated sequences to obtain two audio signals lasting approximately 50 seconds.

The length of 20 ms (i.e. 320 samples) Hamming window is applied in STFT. 12-dimensional ($L = 12$) audio features are extracted from the Mel-scale filterbank analysis. Therefore, we get 12 sets of audio-visual features. In the frame selection process, we preserve 33.7% of features by assigning $\delta_{\text{LW}} = 0.2, \delta_{\text{LH}} = 0.3$, and $\alpha_{\text{LW}} = \alpha_{\text{LH}} = 1$. Then in the training process, GMMs of 5 ($I = 5$) kernels are applied to model the audio-visual coherence.

5.2 Test of Scaling Ambiguity Cancellation

First we need to demonstrate the effectiveness of the proposed algorithm. We get the spectrogram of the training audio, and then manually scale it with a generated function to simulate the scaling ambiguity parameters $d_{11}(f)$ shown in the solid line in Figure 1:

$$d_{11}(f) = (4 * \cos(f/15 + 1) + 6) * \exp(-f * 3/160). \tag{17}$$

If the scaling ambiguity is perfectly solved, the estimated scaling parameters $\alpha(f)$ should satisfy $G_{11}(f) = \alpha(f)d_{11}(f) = 1$, where G denotes the global filter. Figure 1 shows the global filters with ($G_{11AV}(f)$) and without ($G_{11}(f)$) scaling ambiguity cancellation.

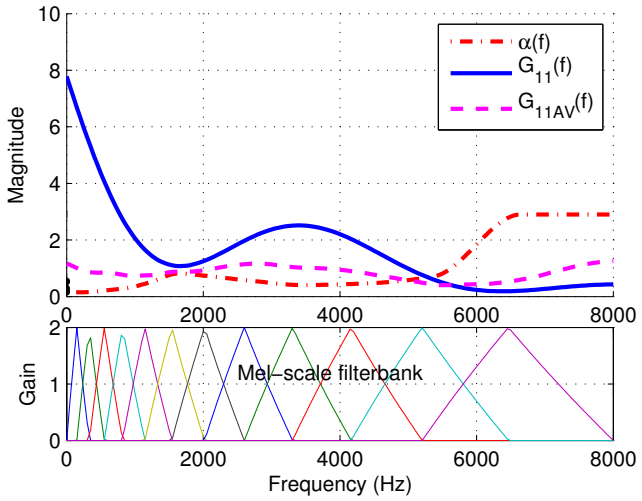

Figure 1: The lower part shows a typical Mel scale filterbank with 12 bands. The original spectrogram was amplified with different scales in different frequency channels (solid line). With the proposed algorithm, we successfully mitigate this problem by applying scaling parameters (dot-dashed) and the recovered spectrogram is very similar to the original one only up to some minor multiplication (dashed line).

Next, we compare the system performance obtained with and without the frame selection scheme. This scheme aims to discard *non-stationary* features, which improves the robustness and reduces the computational complexity in both the training and separation processes. Figure 2 shows the reserved features (cross dots) after feature selection. Figure 3 demonstrates that using the frame selection outperforms training with all the features.

[1] Thanks to B. Rivet in GIPSA-Lab for providing us with this multimodal database.

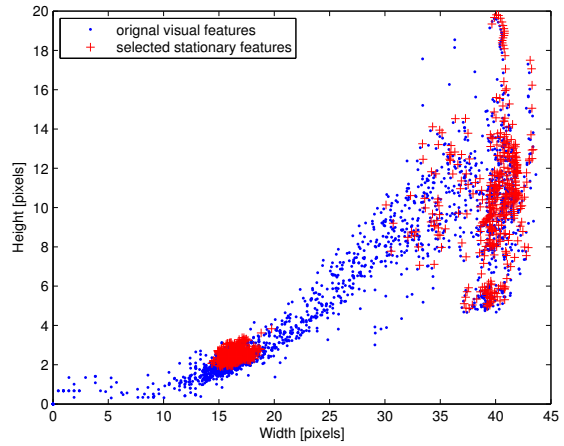

Figure 2: Visual frame selection scheme.

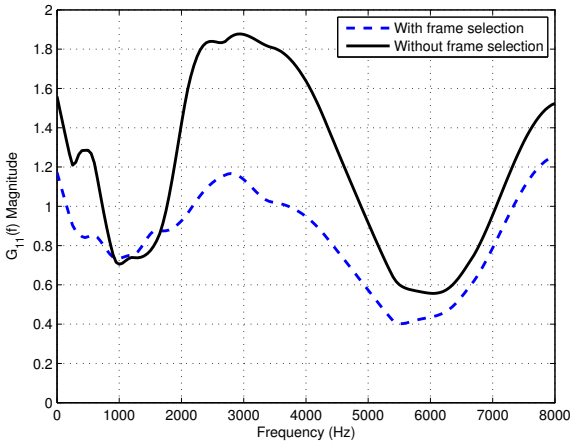

Figure 3: Global filter $G_{11}(f)$ comparison with and without frame selection (GMM parameters used in the separation stage is estimated with frame selection in the training stage). With frame selection, $G_{11}(f)$ oscillates between 0.4 to 1.2, and especially in the low frequency bins where the majority of energy resides, the oscillation is smaller than that without frame selection. Overall, $G_{11}(f)$ is smoother after the frame selection.

5.3 Applied in BSS

The algorithm is tested on convolutive mixtures synthesized on a computer. The mixing filters $\{h_{pk}\}$ are generated by the system utilizing the head related transfer functions (HRTFs) of a dummy head [5], and the length of each mixing filter is 64, which are obtained by specifying the azimuth angles of sources signals in relation to a human head. We select two periods each lasting 8 seconds from the testing audio as the source signals. The mixtures are obtained by convolving the source signals with the HRTFs, and Gaussian white noise at different signal to noise ratios (SNRs) is added to the mixtures.

In the frequency domain, the ICA technique used in [13] is applied. Since our algorithm is based on the assumption that components of $Y_1(f,t)$ come from $s_1(n)$, the permutation problem should be addressed, otherwise, it would greatly degrade the performance. Hence we have applied the correlation algorithm to group the spectral components in advance.

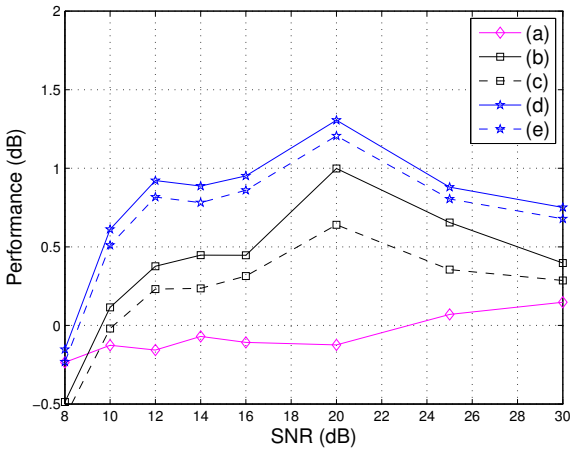

Figure 4: Performance comparison at different SNRs. (a) Without scaling adjustment (i.e. directly calculated from the frequency domain BSS). (b) Scaling with non-linear interpolation but no feature selection. (c) Scaling with linear interpolation but no feature selection. (d) Scaling with non-linear interpolation and frame selection. (e) Scaling with linear interpolation and frame selection.

To test the performance, we use the criterion as follows with normalized $S_1(f,t)$ and $Y_1(f,t)$:

$$\mathscr{P} = 10\log \frac{\|S_1(f,t)\|_F}{\| \left(|S_1(f,t)| - |Y_1(f,t)|\right) \|_F}, \qquad (18)$$

where $|\cdot|$ denotes modulus and $\|\cdot\|_F$ is the Frobenius norm, and a large value of which means good performance. The results are shown in Figure 4, which is an average of 20 independent runs with different mixing filters. From this figure, it can be observed that with the frame selection scheme, we obtain an average of 0.75 dB improvement. And with the non-linear interpolation algorithm, we gain a further 0.09 dB improvement. The performance improvement is modest since the scaling ambiguity has smaller effect (than the permutation ambiguity) on the degradation of the separated signals. We also found that even though the proposed algorithm still outperforms the conventional algorithm, the performance actually decreases when the SNR is greater than 20 dB. Our algorithm is effective with the assumption that the permutation problem has been addressed. However, if the permutation occurs, the proposition might even degrade the result. Our algorithm can be used as a supplementary step after the permutation problem is solved. Moreover, if the same visual feature is exploited to solve the permutation problem, the additional computation is neglect-able for the scaling ambiguity cancellation, which mainly lies in the off-line training process.

6. CONCLUSIONS

We have presented a new approach to address the scaling ambiguity problem encountered in the convolutive BSS system,

utilizing audio-visual coherence, which is statistically characterized in the feature space with GMMs. A new frame selection scheme has been proposed to improve the accuracy of the estimation of the audio-visual coherence. Our algorithm has been tested on a multimodal database composed of different combinations of vowels and consonants, and shows performance improvement.

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (Grant number EP/H012842/1) and the MOD University Defence Research Centre on Signal Processing (UDRC).

REFERENCES

[1] S. Amari, S. C. Douglas, A. Cichocki, and H. H. Yang. Multichannel blind deconvolution and equalization using the natural gradient. In *Proc. IEEE Int. Workshop on Wireless Commun.*, pages 101–104, 1997.

[2] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371, Aug 2010.

[3] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, Apr 1994.

[4] F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM J. on Numerical Analysis*, 17(2):238–246, 1980.

[5] B. Gardner and K. Martin. Head related transfer functions of a dummy head. Website, 1994. `http://sound.media.mit.edu/ica-bench/`.

[6] M. Z. Ikram and D. R. Morgan. A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation. In *Proc. ICASSP*, pages 881–884, 2002.

[7] Q. Liu, W. Wang, and P. Jackson. Bimodal coherence based scale ambiguity cancellation for target speech extraction and enhancement. In *Proc. Interspeech*, pages 438–441, 2010.

[8] Q. Liu, W. Wang, and P. Jackson. Use of bimodal coherence to resolve spectral indeterminacy in convolutive BSS. In *Proc. LVA/ICA*, pages 131–139, 2010.

[9] K. Matsuoka. Minimal distortion principle for blind source separation. In *Proc. SICE*, volume 4, pages 2138–2143, 2002.

[10] R. Mazur and A. Mertins. An approach for solving the permutation problem of convolutive blind source separation based on statistical signal models. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):117–126, Jan. 2009.

[11] S. M. Naqvi, M. Yu, and J. A. Chambers. A multimodal approach for blind source separation of moving sources. *IEEE Journal Selected Topics in Signal Processing*, 4(5):895–910, Oct. 2010.

[12] F. Nesta, M. Omologo, and P. Svaizer. A novel robust solution to the permutation problem based on a joint multiple tdoa estimation. In *Proc. IWAENC*, 2008.

[13] D.-T. Pham, C. Servière, and H. Boumaraf. Blind separation of speech mixtures based on nonstationarity. In *Proc. ISSPA*, volume 2, pages 73–76, Jul 2003.

[14] B. Rivet, L. Girin, and C. Jutten. Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures. *IEEE Transactions on Audio, Speech and Language Processing*, 15(1):96–108, Jan 2007.

[15] H. Sawada, S. Araki, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions on Speech and Language Processing*, 15(5):1592–1604, Sept. 2007.

[16] J.-L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93:B69–B78, 2004.

[17] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten. Separation of audio-visual speech sources: a new approach exploiting the audio-visual coherence of speech stimuli. *EURASIP Journal on Applied Signal Processing*, 2002(11):1165–1173, Jan 2002.

[18] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. Chambers. Video assisted speech source separation. In *Proc. ICASSP*, pages 425–428, 2005.