Bimodal Coherence based Scale Ambiguity Cancellation for Target Speech Extraction and Enhancement

Qingju Liu, Wenwu Wang, Philip Jackson

Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, UK

{Q.Liu, W.Wang, P.Jackson}@surrey.ac.uk

Abstract

We present a novel method for extracting target speech from auditory mixtures using bimodal coherence, which is statistically characterised by a Gaussian mixture modal (GMM) in the off-line training process, using the robust features obtained from the audio-visual speech. We then adjust the ICA-separated spectral components using the bimodal coherence in the time-frequency domain, to mitigate the scale ambiguities in different frequency bins. We tested our algorithm on the XM2VTS database, and the results show the performance improvement with our proposed algorithm in terms of signal to interference ratio (SIR) measurements.

Index Terms: speech extraction, bimodal coherence, audio-visual, Gaussian mixture model (GMM), independent component analysis (ICA), scale ambiguity

1. Introduction

Human speech production and perception are bimodal by nature: speech is produced by articulatory organs coupled with contemporary visual movements of these organs; speech is perceived by auditory and visual senses. The correlation between audio and visual stimulus can be used to extract target speech from visual speech mixtures as shown in recent works [1] - [4]. Sodoyer et al. [1] addressed the separation problem for an instantaneous stationary mixture of decorrelated sources. Wang et al. [2] implemented a similar idea by applying the Bayesian framework to the fused feature observations for both instantaneous and convolutive mixtures. Rivet et al. [3] used the bimodal coherence to address the permutation and scale ambiguities in the spectral domain. With the time sparsity constraint, Casanovas et al. [4] detected temporal audio-visual structures represented by atoms taken from redundant dictionaries, and extracted sources from a soundtrack.

However, the audio mixing process generates convolutive mixtures, and the algorithm proposed in [1] cannot deal with this situation. The work in [2] considered convolutive mixtures, but the number of taps for the mixing filters is relatively small. When modeling the coherence, [3] chose a high-dimensional audio feature vector to represent the audio modality, therefore the model was sensitive to outliers. The method in [4] used visual features only to determine the number of sources and the active periods when each person is speaking. The scale ambiguity problem with the extracted source components is not addressed in [1, 2, 4].

In this paper, we consider the convolutive model [2, 3, 5] with the assumption of non-Gaussianity and independence constraints of the sources. In the off-line training process, the power spectrum of the audio signal is mapped into Mel-scale filter-banks as the audio features; the geometric-type visual features

are extracted from the training video associated with the target speaker. After the synchronization and fusion, we obtain the audio-visual feature vector for the characterisation of the bimodal coherence using a GMM model. The audio-visual coherence is then applied to address the scale ambiguity in the time-frequency domain. The remainder of the paper is organised as follows. An overview of ICA-based speech extraction for the cocktail party problem is presented in Section 2. Section 3 introduces our bimodal feature extraction and fusion method. Details of the scale ambiguity cancellation algorithm exploiting the audio-visual coherence are presented in Section 4. The simulation results are analysed and discussed in Section 5. Finally Section 6 concludes the paper.

2. Speech extraction

2.1. Model

The speech mixing process for a cocktail party scenario can be approximated as a convolutive model:

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

$\mathbf{x}(n) = \mathbf{H} * \mathbf{s}(n) + \boldsymbol{\xi}(n)$ is its matrix form, where $\mathbf{x}(n) = [x_1(n), ..., x_P(n)]^T$ is the observation vector at the discrete time index n; $\mathbf{s}(n) = [s_1(n), ..., s_K(n)]^T$ is the source vector and $\boldsymbol{\xi}(n)$ the additive noise vector; $\mathbf{H}$ is the mixing matrix whose entry h_{pk} represents the room impulse response filter from source k to sensor p and $*$ denotes a convolution.

Suppose we are just interested in the target speech $s_1(n)$, the objective of source extraction is to find a set of separation filters $\{w_{1p}(m)\}$ that satisfy:

$$\hat{s}_1(n) = y_1(n) = \sum_{p=1}^{P} \sum_{m=0}^{+\infty} w_{1p}(m) x_p(n-m), \qquad (2)$$

or in matrix form $\hat{s}_1(n) = y_1(n) = \mathbf{w}_1 * \mathbf{x}(n)$ where $\mathbf{w}_1 = [w_{11}, ..., w_{1P}]^T$ is the separation vector whose entry w_{1p} is the impulse response filter from observation p to the estimate of source 1.

2.2. Time-frequency domain method

Source extraction can be performed in the time-frequency domain where independent component analysis (ICA) [6] algorithms can be applied in each frequency bin f, as depicted by the upper dashed box in Figure 1 (not including the dashed lines). The convolutive model of equation (1) becomes a set of instantaneous models after performing the short-time Fourier transform (STFT) to the observations $\mathbf{X}(f, t) = \mathbf{H}(f)\mathbf{S}(f, t)$,
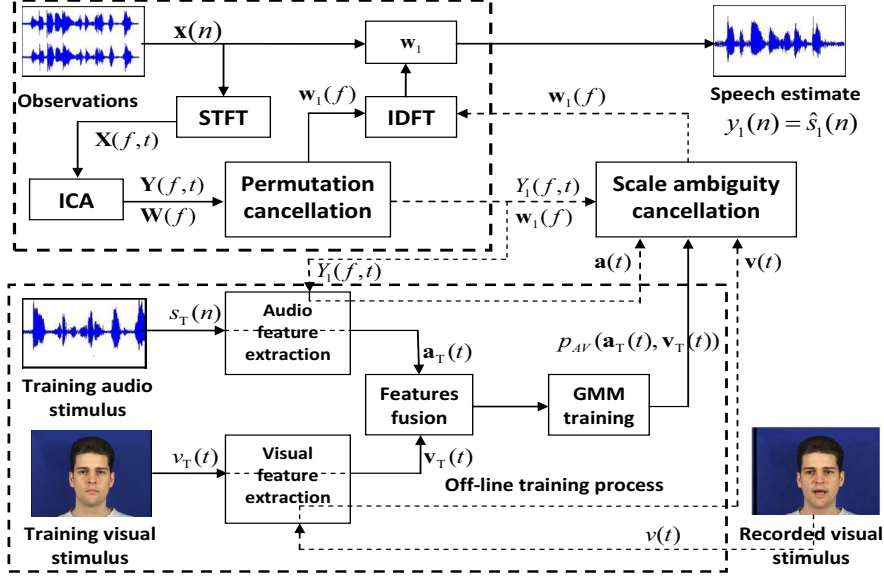
Figure 1: *Flow of bimodal target speech extraction.*

where t is the time-frame index. Then ICA is applied separately to the spectral components $\mathbf{X}(f,t)$ to obtain the independent outputs $\mathbf{Y}(f,t) = [Y_1(f,t),...,Y_K(f,t)]^T$. $\mathbf{Y}(f,t)$ is considered as a copy of $\mathbf{S}(f,t)$, only up to a permutation matrix $\mathbf{P}(f)$ and a diagonal matrix of gains $\mathbf{D}(f)$:

$$\mathbf{Y}(f,t) = \mathbf{P}(f)\mathbf{D}(f)\mathbf{S}(f,t), \qquad (3)$$

these are the so-called permutation ($\mathbf{P}(f)$) and scale ($\mathbf{D}(f)$) indeterminacy problems.

As for the permutation problem, many algorithms have been proposed, with [2, 3] or (most of the available algorithms are) without [5] visual information. [2, 3] both use the audio-visual coherence to the alignment of the spectral components, and [5] applies the combination of the correlation method and direction of arrival (DOA) estimation to address this problem. We have also proposed a method for this problem using the bimodal coherence in a previous work [7].

After the permutation indeterminacy cancellation, we get $\mathbf{Y}(f,t) = \mathbf{D}(f)\mathbf{S}(f,t)$ suppose there is no global permutation, i.e., all components of $Y_k(f,t)$ come from $s_k(n)$.

As for the scale problem, $\mathbf{D}(f_i) \neq \mathbf{D}(f_j)$ when $i \neq j$, $Y_1(f,t)$ is amplified with different scales at different frequency bins. Since we are just interested in the extraction of the target speech $Y_1(f,t)$, and $Y_1(f,t) = \beta(f)S_1(f,t)$, where $\beta(f)$ is the first diagonal entry of $\mathbf{D}(f)$. Therefore, if we reconstruct $y_1(n)$ in the time domain, it would be a FIR filtered version of $s_1(n)$. To address this scale problem, we estimate a set of scale parameters $\{\alpha(f) = 1/\beta(f)\}_f$ to adjust $Y_1^\dagger(f,t) = \alpha(f)Y_1(f,t)$, and the audio-visual coherence can be exploited, as shown in Figure 1.

3. Off-line training process

Our objective at this stage is to statistically characterise the bimodal coherence between audio and visual modalities, as depicted in the lower dashed box in Figure 1 (not including the dashed lines).

3.1. Audio & visual feature extraction

We exploit the non-linear resolution of the human auditory system across an audio spectrum using the Mel-scale filterbank analysis. We denote $\mathcal{F}_l$ as the group of the frequency bins spanned by the l-th filterbank. The spectral power is mapped into these filterbanks to achieve the L-dimensional audio feature $\mathbf{a_T}(t) = [a_{\mathbf{T}1}(t),...,a_{\mathbf{T}L}(t)]^T$ for statistical training, where

$$a_{\mathbf{T}l}(t) = \sum_{f \in \mathcal{F}_l} b_l(f)|S_{\mathbf{T}}(f,t)|^2, \qquad (4)$$

and $b_l(f)$ is the magnitude of the l-th filterbank while $S_{\mathbf{T}}(f,t)$ is the spectral component of the sequence of the training audio. The lower part in Figure 4 shows typical Mel-scale filterbanks.

We use the same front geometric visual features as in [1][3]: the lip width (LW) and height (LH) from the internal labial contour. Figure 2 shows the detailed 2-dimensional visual feature extraction method, $\mathbf{v_T}(t) = [\text{LW}(t), \text{LH}(t)]^T$ is extracted from the testing video of the target speaker.
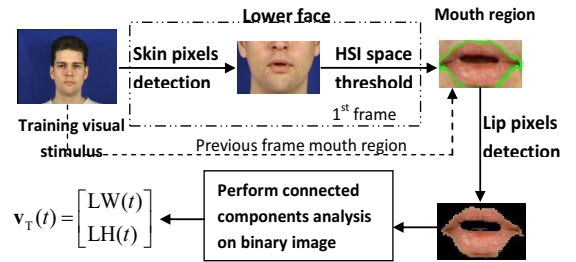

Figure 2: *Visual feature extraction.*

Once all the features are extracted, we concatenate the l-th component of the audio vector $a_{\mathbf{T}l}(t)$ with the visual features $\mathbf{v_T}(t)$ after synchronization, to get a set of 3-dimensional audio-visual vectors $\{\mathbf{u_{T}}_l(t)\}_l$, where $\mathbf{u_{T}}_l(t) = [\mathbf{v_T}(t); a_{\mathbf{T}l}(t)]$ is the audio-visual vector corresponding to the l-th filterbank.

3.2. Feature-level fusion

The audio-visual coherence of each filterbank can be statistically characterized as a GMM model with I kernels:

$$p_{AV}(\mathbf{u}_{\mathbf{T}l}(t)) = \sum_{i=1}^{I} \gamma_{li} p_G(\mathbf{u}_{\mathbf{T}l}(t) \mid \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}), \qquad (5)$$

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 covariance matrix of the i-th kernel for the l-th filterbank. Each kernel of this mixture represents one cluster of the audio-visual data modeled by a joint Gaussian normal distribution:

$$p_G(\mathbf{u}_{\mathbf{T}l}(t) \mid \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}) = \mathcal{N}(\mathbf{u}_{\mathbf{T}l}(t) \mid \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}). \qquad (6)$$

We denote $\lambda_{li} = \{\gamma_{li}, \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li}\}$ as the parameter set, with $\{\lambda_{li}\}_{l,i}$ estimated by the expectation maximization algorithm.

4. Scale ambiguity cancellation algorithm

As $y_1(n)$ is the estimate of $s_1(n)$, $y_1(n)$ will have maximum coherence with the recorded video $v(t)$ associated with the target speaker.

The frequency bin group $f \in \mathcal{F}_l$ is treated as a whole. Suppose $Y_1^\dagger(f,t) = \alpha(\mathcal{F}_l)Y_1(f,t)$ is the exact copy of the source speech $S_1(f,t)$ for $f \in \mathcal{F}_l$ without any scale amplification, i.e., $Y_1^\dagger(f,t) = S_1(f,t)$, for $f \in \mathcal{F}_l$. The audio features extracted from $Y_1^\dagger(f,t)$ will have the maximum coherence with the visual feature $\mathbf{v}(t)$ extracted from $v(t)$. As in [2], we can maximize the following criterion to address the scale ambiguity:

$$\hat{\alpha}(\mathcal{F}_l) = \arg\max_{\alpha(\mathcal{F}_l)} \sum_t p_{AV}(\mathbf{u}_l^\dagger(t)), \qquad (7)$$

where $\mathbf{u}_l^\dagger(t) = [\mathbf{v}(t); a_l^\dagger(t)]$ and $a_l^\dagger(t)$ (resp. $a_l(t)$) is the audio feature extracted from $Y_1^\dagger(f,t)$ (resp. $Y_1(f,t)$). Combining equation (4) gives

$$\sum_t a_l^\dagger(t) = |\alpha(\mathcal{F}_l)|^2 \sum_t a_l(t). \qquad (8)$$

Since the covariance matrix in equation (6) is diagonal, the joint distribution probability density of each kernel is the product of the marginal probability densities. Therefore,

$$p_{AV}(\mathbf{u}_l^\dagger(t)) = \sum_{i=1}^{I} \gamma_{li} \mathcal{N}(\mathbf{u}_l^\dagger(t) \mid \boldsymbol{\mu}_{li}, \boldsymbol{\Sigma}_{li})$$
$$= \sum_{i=1}^{I} \gamma_{li} p_V(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV}) p_A(a_l^\dagger(t) \mid \boldsymbol{\mu}_{liA}, \boldsymbol{\Sigma}_{liA}) \qquad (9)$$

where $\boldsymbol{\mu}_{liV} = [\mu_{li1}, \mu_{li2}]^T$, $\boldsymbol{\Sigma}_{liV} = diag([\sigma_{li1}, \sigma_{li2}])$ are the mean vector and covariance matrix of the visual features of the i-th kernel for the l-th filterbank; $\boldsymbol{\mu}_{liA} = \mu_{li3}$ and $\boldsymbol{\Sigma}_{liA} = \sigma_{li3}$ corresponds to the audio feature. In equation (9),

$$p_V(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV}) = \mathcal{N}(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV})$$
$$p_A(a_l^\dagger(t) \mid \boldsymbol{\mu}_{liA}, \boldsymbol{\Sigma}_{liA}) = \mathcal{N}(a_l^\dagger(t) \mid \mu_{li3}, \sigma_{li3}). \qquad (10)$$

Maximizing the objective function in equation (7) is equivalent to finding an audio feature $a_l^\dagger(t)$ that maximizes $p_{AV}(\mathbf{u}_l^\dagger(t))$ in equation (9). Since the visual feature at time frame t is determined, $p_V(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV})$ is determined by the parameter sets that have been estimated in the off-line training

process. If there is only one kernel, i.e. $I = 1$, we can just let $a_l^\dagger(t) = \mu_{l13}$ since $p_A(a_l^\dagger(t) \mid \boldsymbol{\mu}_{liA}, \boldsymbol{\Sigma}_{liA})$ is a Gaussian distribution. However, there are generally multiple kernels, so $a_l^\dagger(t) = \mu_{l13}$ is a weighted average over those kernels:

$$a_l^\dagger(t) = \sum_{i=1}^{I} c_{li}(t)\mu_{li3}, \qquad (11)$$

where $\quad c_{li}(t) = \dfrac{\gamma_{li} p_V(\mathbf{v}(t) \mid \boldsymbol{\mu}_{liV}, \boldsymbol{\Sigma}_{liV})}{\sum_j \gamma_{lj} p_V(\mathbf{v}(t) \mid \boldsymbol{\mu}_{ljV}, \boldsymbol{\Sigma}_{ljV})}.$

We combine equations (8) and (11) to give the scale parameter:

$$\alpha(\mathcal{F}_l) = \sqrt{\sum_t a_l^\dagger(t) / \sum_t a_l(t)}. \qquad (12)$$

In such a way, we get L scale parameters, and each one affects the frequency bins spanned by a filterband. This scheme can reach a high resolution, which is determined by the number of filters: and the larger the number, the higher the resolution. If finally we use M filters to analyse M bins, then we address the scale ambiguity for nearly each frequency bin (consider the non-linearity of the Mel filterbanks), and thus the highest resolution is achieved.

However, adjacent $\mathcal{F}_l$ overlap with each other, and of course we cannot define two scale parameters for an overlapped frequency bin. To solve this problem: at each centre frequency bin of $\mathcal{F}_l$, the scale parameter $\alpha(\mathcal{F}_l)$ is fixed; as for the other bins, we smooth between those scale parameters.

5. Experiments

5.1. Data

We tested the proposed algorithm on the XM2VTS [8] multi-modal database, in which the speech data were recorded 4 times at approximately one month intervals, with continuous sentences of words and digits in mono, 16 bit, 32 kHz, PCM wave files, and the frontal face videos captured at 25 fps.

In the off-line training process, we trained the audio-visual coherence model of the target speaker with 40 s of audio and visual speech. The audio was downsampled to 8 kHz, and a 32 ms (256 point) Hamming window with 12 ms (96 point) overlap was applied in STFT. The number of Mel-scale filters was 12, ($L = 12$). Thus we extracted twelve sets of 3-dimensional audio-visual features $\{\mathbf{u}_{\mathbf{T}l}(t)\}_l$ for training. The visual features were upsampled to 50 Hz to be synchronized with the audio features. For simplicity, we only used 5 ($I = 5$) kernels to approximate the audio-visual coherence. Therefore, $L \times I = 60$ parameter sets γ_{li} were estimated.

The algorithm was tested on convolutive mixtures synthesized on computer. The filters $\{h_{pk}\}$ were generated by the system utilizing the head related transfer functions (HRTFs) of a dummy head [9], and the length of each mixing filter is 64. For a 2×2 mixing process, we specified the two azimuth angles of the sources in relation to a human head to determine the $\{h_{pk}\}$. Two audio signals with each lasting 4 s were convolved with the filters to generate the mixtures, and Gaussian white noise (GWN) was added to both mixtures at different signal to noise ratios (SNRs).

5.2. Experimental results

After the training process, we got the GMM model, where each kernel approximated one class of utterance in a filterbank. Figure 3 shows the scatter plot of the first two components of the audio-visual vector (i.e. the visual vector $LW(t)$ and $LH(t)$).

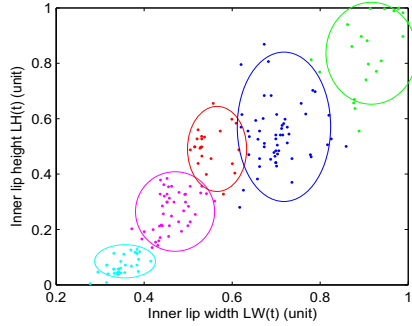Figure 3: *Clustering of the first two components of the audio-visual vector $\mathbf{u}_{\mathbf{T}1}(t)$ after GMM modeling ($I = 5$).*

We denote $G_{11}(f)$ as the global filter from target speech $s_1(n)$ to $y_1(n)$ in the frequency domain to evaluate the performance, and $G_{11}(f) = DFT(\sum_{p=1}^{P}\{w_{1p} * h_{p1}\})$. If $y_1(n)$ were the exact copy of $s_1(n)$, then the time domain transformation of $G_{11}(f)$ would be the unit impulse. Thus, for $\forall f, G_{11}(f) = 1$, i.e., components of $Y_1(f,t)$ coming from $S_1(f,t)$ would be equally amplified. The blue solid curve in Figure 4 is the filter $G_{11}(f)$ without adjusting scales. We found that in some frequency bins, source components were hugely amplified, while in some other bins especially in the high frequency bin, they were greatly attenuated. However, after the scale parameters were applied to the source estimate $Y_1(f,t)$, the new global filter $G_{11AV}(f)$ as shown in the magenta dashed curve, was much flatter in the frequency domain. Therefore, the scale ambiguity was reduced and the extracted target speech was more coherent with the original speech.

The denoising effect of the proposed algorithm is also proved much better after applying the scale ambiguity cancellation. The speech enhancement is more obvious in high noise environment. Table 1 shows the output signal to interference ratios (SIRs) at low input SNRs. The input SIRs are also calculated. The audio-visual in the table denotes the proposed algorithm using the audio-visual coherence, while the audio-only represents the traditional speech extraction algorithm using only audio signals. The results were averages of 20 experiments with different mixing processes.
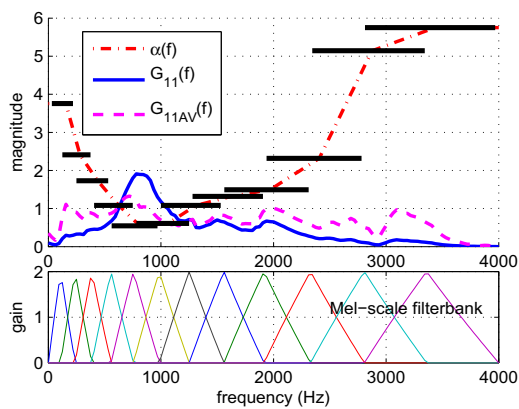

Figure 4: *Global filter comparison (blue solid and magenta dashed curves) before and after the scale cancellation method.*

Table 1: *Output SIR* (dB) *comparison.*

Input SNR	4	6	8	10	12
Input SIR	-3.13	-2.38	-1.82	-1.42	-1.15
audio-only	0.66	2.07	3.83	5.29	6.77
audio-visual	2.07	3.71	4.55	5.82	6.80

6. Conclusions

We have presented a target speech extraction system using bi-modal coherence. In this system, filterbank analysis was exploited to extract the audio features, which were combined with geometric visual features to form an audio-visual feature space. The GMM model was then trained on the audio-visual data set to characterise the bimodal coherence statistically. A new scale cancellation scheme exploiting the audio-visual coherence was applied to the ICA-separated components to enhance the extracted speech. Our algorithm was tested on the XM2VTS multi-modal database and showed improved performance. In the future, we will consider using some dynamic features in video, instead of just static features. In addition, we will increase the number of kernels to improve the accuracy of the audio-visual model.

7. Acknowledgements

8. References

[1] Sodoyer, D., Schwartz, J.L., Girin, L., Klinkisch, J. and Jutten, C., "Separation of Audio-Visual Speech Sources: a New Approach Exploiting the Audio-Visual Coherence of Speech Stimuli", EURASIP J. Appl. Signal Process., 11:1165–1173, 2002.

[2] Wang, W., Cosker, D., Hicks, Y., Sanei, S. and Chambers, J., "Video Assisted Speech Source Separation", in Proc. IEEE ICASSP, 425–428, 2005.

[3] Rivet, B., Girin, L. and Jutten, C., "Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals from Convolutive Mixtures", IEEE Trans. Audio Speech Lang. Process., 15(1):96–108, 2007.

[4] Casanovas, A.L., Monaci, G., Vandergheynst, P. and Gribonval, R., "Blind Audiovisual Source Separation Using Overcomplete Dictionaries", in Proc. IEEE ICASSP, 2008.

[5] Sawada, H., Mukai, R., Araki, S., and Makino, S., "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation", IEEE Trans. Speech Audio Process., 12(5): 530–538, 2004.

[6] Comon, P., "Independent Component Analysis, a New Concept?", Signal Process., 36(3):287–314, 1994.

[7] Liu, Q., Wang, W. and Jackson, P., "Use of Bimodal Coherence to Resolve Spectral Indeterminacy in Convolutive BSS", submitted to LVA/ICA 2010.

[8] Messer, K., Matas, J., Kittler, J., Luettin, J. and Maitre, G., "XM2VTSDB: The Extended M2VTS Database", In: AVBPA, 1999. Online: http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/.

[9] Gardner, B. and Martin, K., "Head Related Transfer Functions of a Dummy Head", 1994. Online: http://sound.media.mit.edu/ica-bench/.