REVERBERANT SPEECH SEPARATION BASED ON AUDIO-VISUAL DICTIONARY LEARNING AND BINAURAL CUES

Qingju Liu, Wenwu Wang, Philip Jackson, Mark Barnard

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

ABSTRACT

Probabilistic models of binaural cues, such as the interaural phase difference (IPD) and the interaural level difference (ILD), can be used to obtain the audio mask in the time-frequency (TF) domain, for source separation of binaural mixtures. Those models are, however, often degraded by acoustic noise. In contrast, the video stream contains relevant information about the synchronous audio stream that is not affected by acoustic noise. In this paper, we present a novel method for modeling the audio-visual (AV) coherence based on dictionary learning. A visual mask is constructed from the video signal based on the learnt AV dictionary, and incorporated with the audio mask to obtain a noise-robust audio-visual mask, which is then applied to the binaural signal for source separation. We tested our algorithm on the XM2VTS database, and observed considerable performance improvement for noise corrupted signals.

Index Terms— Binaural source separation, interaural difference, audio-visual dictionary learning, matching pursuit, noise reduction.

1. INTRODUCTION

For binaural signals, exploiting the interaural cues IPD and ILD [1], we can statistically evaluate the probability of each time-frequency (TF) point of the audio mixture that belongs to each source, and therefore obtain TF-domain *audio masks* for source separation. However, the parameter estimation of an interaural statistical model is degraded by noise and long reverberation. To overcome this limitation, we propose a novel method exploiting both binaural and visual cues.

Visual cues have the potential to improve the intelligibility of noise corrupted speech, since they are not affected by acoustic noise. Video focused on the mouth region has proved useful in separation matrix estimation for blind source separation (BSS) [2], mitigating the ambiguities of convolutive BSS [3, 4], and providing other information such as the activity information to assist the audio domain separation [5].

To exploit the additional information provided by the visual cues, we propose a novel audio-visual (AV) dictionary learning method. Each AV atom of the dictionary contains a short audio segment and a concurrent video segment that are bimodal-coherent [5, 6]. In other words, the occurrence of one modality (e.g., visual lip movements) often indicates the existence of the other (e.g., the utterance of words).

In this paper, we develop a dictionary learning method to capture the audio-visual coherence, which is then used as an additional cue to refine the TF mask obtained by the binaural cues. The source signal is reconstructed by applying the TF mask to the mixture spectrogram, followed by an inverse short time Fourier transform (STFT).

2. BINAURAL SOURCE SEPARATION

A source signal arrives at the left ear $l(n)$ and the right ear $r(n)$ (where n is the discrete time index) with different time delays τ and attenuations, that can be obtained from the following equation:

$$L(m,w)/R(m,w) = 10^{\frac{\alpha(m,w)}{20}} e^{j\phi(m,w)}, \tag{1}$$

where $L(m,w)$ and $R(m,w)$ are the STFTs of $l(n)$ and $r(n)$ respectively, at the TF point (m,w). $\alpha(m,w)$ is the ILD and $\phi(m,w)$ is the IPD, which can be statistically modeled with two Gaussian distributions [1]:

$$\begin{cases} p_{\text{IPD}}(m,w|i,\tau) \sim \mathcal{N}(\xi_{i\tau}(w); \sigma^2_{i\tau}(w)) \\ p_{\text{ILD}}(m,w|i) \sim \mathcal{N}(\mu_i(w); \eta^2_i(w)) \end{cases}, \tag{2}$$

where $p_{\text{IPD}}(m,w|i,\tau)$ is the likelihood of $\phi(m,w)$ being originated from source i at delay τ, while $p_{\text{ILD}}(m,w|i)$ is the likelihood of $\alpha(m,w)$ being originated from source i, if the binaural signals are mixtures of several sources. The parameter set $\left\{\xi_{i\tau}(w), \sigma^2_{i\tau}(w),\ \mu_i(w), \eta^2_i(w)\right\}$ can be estimated via the expectation maximization (EM) method:

- E step. Calculate the posterior probability of a TF point (m,w) coming from source i at delay τ:

  $$p(i,\tau|m,w) = \frac{p_{\text{IPD}}(m,w|i,\tau)p_{\text{ILD}}(m,w|i)\varphi_{i\tau}}{\sum_j p_{\text{IPD}}(m,w|j,\tau)p_{\text{ILD}}(m,w|j)\varphi_{j\tau}},$$

  where $\varphi_{i\tau}$ is the overall probability of a TF point coming from source i at delay τ.

- M step. Update the parameters.

  $\xi_{i\tau}(w)$ and $\sigma^2_{i\tau}(w)$ are updated by the expectations over m, while $\mu_i(w)$ and $\eta^2_i(w)$ are over m and τ, and $\varphi_{i\tau}$ is the expectation of $p(i,\tau|m,w)$ over m and w.

Once converged after several iterations of EM, we can obtain the TF mask for source i: $\mathcal{M}_i(m,w) = \sum_\tau p(i,\tau|m,w)$. Using this mask, we can estimate source i from either $L(m,w)$ or $R(m,w)$, by $\hat{S}_i(m,w) = \mathcal{M}_i(m,w)L(m,w)$. More details about this technique can be found in [1]. We

978-1-4673-0183-1/12/$31.00 ©2012 IEEE

denote TF mask that contributes to the reconstruction of the target speech as $\mathcal{M}_a(m, w)$.

However, the binaural cues of IPD and ILD are seriously affected by acoustic noise. To address this limitation, we incorporate the visual information through bimodal coherence modeling based on AV dictionary learning.

3. AUDIO-VISUAL DICTIONARY LEARNING

We aim to capture *the bimodal-coherent parts* of an AV sequence, but not to code the whole sequence. Using a similar bimodal dictionary learning framework described in [6], we develop a new AV dictionary learning method. We denote an AV sequence as follows

$$\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$$

where a and v denote audio and visual modalities, and

$$\boldsymbol{\psi}^a = (\psi^a(m)) \in \mathbb{R}^{\tilde{M}}, \boldsymbol{\psi}^v = (\psi^v(y, x, l)) \in \mathbb{R}^{\tilde{Y} \times \tilde{X} \times \tilde{L}}$$

in which m is the time frame index of the short-term energy function $\psi^a(m)$ derived from the audio stream, l is the image frame index, and y, x denote the pixel coordinates. Similarly, we define the AV atom in the redundant [5, 6] dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}, k = 1, 2, ..., K$ as $\boldsymbol{\phi}_k = (\boldsymbol{\phi}_k^a; \boldsymbol{\phi}_k^v)$ where $\boldsymbol{\phi}_k^a = (\phi_k^a(m)) \in \mathbb{R}^M$ and $\boldsymbol{\phi}_k^v = (\phi_k^v(y, x, l)) \in \mathbb{R}^{Y \times X \times L}$. We also define an AV segment taken from $\boldsymbol{\psi}$ as $\bar{\boldsymbol{\psi}}_{\hat{m}} = (\bar{\boldsymbol{\psi}}_{\hat{m}}^a; \bar{\boldsymbol{\psi}}_{\hat{y}\hat{x}\hat{l}}^v)$ which has the same dimension as $\boldsymbol{\phi}_k$, and $\bar{\boldsymbol{\psi}}_{\hat{m}}^a = [\psi^a(\hat{m} + 1), ..., \psi^a(\hat{m}+M)]^T \in \mathbb{R}^M$ and $\bar{\boldsymbol{\psi}}_{\hat{y}\hat{x}\hat{l}}^v = \boldsymbol{\psi}^v(\hat{y}+1 : \hat{y}+Y, \hat{x}+1 : \hat{x} + X, \hat{l} + 1 : \hat{l} + L) \in \mathbb{R}^{Y \times X \times L}$, where $\hat{m}$ and $\hat{y}, \hat{x}, \hat{l}$ indicate the locations of the segment on the AV sequence, and the superscript T means transpose.

Using the atoms chosen from the dictionary $\mathcal{D}$ and their translations[1], the AV sequence can be coded as:

$$\begin{pmatrix} \psi^a(m) \\ \psi^v(y, x, l) \end{pmatrix} \approx \sum_{i=1}^{N} \begin{pmatrix} c_i \phi_{b_i}^a (m - m_i) \\ \phi_{b_i}^v (y - y_i, x - x_i, l - l_i) \end{pmatrix} \quad (3)$$

where $\boldsymbol{\psi}$ is approximated by the combination of multiple atoms indexed by b_i and their translations parameterized by m_i, y_i, x_i, l_i, and c_i is a scaling factor for approximating the audio sequence. To synchronize audio and visual sequence, $|m_i/f_s^a - l_i/f_s^v| < 1/f_s^v$ is enforced, assuming $f_s^v < f_s^a$, where f_s^a and f_s^v are respectively the frame rates of the audio and visual sequences. The parameter set $\Omega = \{b_i, c_i, m_i, y_i, x_i, l_i\}, i = 1, ..., N$ can be found by the matching pursuit (MP) technique [7]. In the i-th iteration of MP, b_i-th atom $\boldsymbol{\phi}_{b_i}$ is chosen to fit $\boldsymbol{\psi}$ the best.

To find these parameters, we define the matching criterion as follows, which measures how good atoms $\boldsymbol{\phi}_k$s are to fit $\boldsymbol{\psi}$:

$$[b_i, m_i, y_i, x_i, l_i] = \\ \underset{[k,\hat{m},\hat{y},\hat{x},\hat{l}]}{\arg\max} \left(\left|< \bar{\boldsymbol{\psi}}_{\hat{m}}^a, \boldsymbol{\phi}_k^a >\right| \cdot e^{\left(-\overline{(\bar{\boldsymbol{\psi}}_{\hat{y}\hat{x}\hat{l}}^v - \boldsymbol{\phi}_k^v)^2}/\sigma_v^2\right)} \cdot V_{\hat{y}\hat{x}\hat{l}} \right), \quad (4)$$

where $\hat{m}, \hat{y}, \hat{x}, \hat{l}$ are all the possible temporal and spacial shifts of $\boldsymbol{\phi}_k$ over $\boldsymbol{\psi}$. σ_v is a weighting constant. $V_{\hat{y}\hat{x}\hat{l}}$ is the temporal variance of $\bar{\boldsymbol{\psi}}_{\hat{y}\hat{x}\hat{l}}^v$, i.e., the mean value of the variances of $\bar{\boldsymbol{\psi}}_{\hat{y}\hat{x}\hat{l}}^v$ over the third (temporal) dimension. $|< \cdot, \cdot >|$ is the inner product modulus and the overline calculates the mean value over all elements. After the i-th iteration, contributions of $\boldsymbol{\phi}_{b_i}^a$ to $\boldsymbol{\psi}^a$ will be removed:

$$\boldsymbol{\psi}^a(m_i + 1 : m_i + M) \leftarrow \bar{\boldsymbol{\psi}}_{m_i}^a - c_i \boldsymbol{\phi}_{b_i}^a, \quad (5)$$

where c_i is the audio coefficient calculated as $\left|< \bar{\boldsymbol{\psi}}_{m_i}^a, \boldsymbol{\phi}_{b_i}^a >\right|$.

In equation (4), the first term evaluates how $\boldsymbol{\phi}_k^a$ fits $\boldsymbol{\psi}^a$ segment; the second term exponentially computes the similarity between $\boldsymbol{\phi}_k^v$ and $\boldsymbol{\psi}^v$ segment; the third term promotes dynamic atoms, to avoid visual atoms converging to static background. The dictionary can be learnt by two steps iteratively:

```
Input: A training AV sequence $\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$
Output: An AV dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}, k = 1, 2, ..., K$
foreach $iter = 1 : MaxIter$ do
    - Coding step
    foreach $i = 1 : N$ do
        • Find $\{b_i, c_i, m_i, y_i, x_i, l_i\}$ to maximize the criterion
        (4) using MP.
        • Calculate audio residue with equation (5).
    end
    - Learning step
    foreach $k$ = random permute$(1 : K)$ do
        $\mathcal{I}_k = \{i\}$, subject to $b_i = k$ and $i = 1, 2, ..., N$.
        • $\boldsymbol{\phi}_k^a$ update.
        * foreach $i \in \mathcal{I}_k$ do
            $\boldsymbol{\psi}^a(m_i + 1 : m_i + M) \leftarrow \bar{\boldsymbol{\psi}}_{m_i}^a + c_i \boldsymbol{\phi}_k^a$.
        end
        * Apply singular value decomposition (SVD) to $\bar{\boldsymbol{\Psi}}_k^a$
        whose columns are $\bar{\boldsymbol{\psi}}_{m_i}^a, i \in \mathcal{I}_k$.
        * Update $\boldsymbol{\phi}_k^a$ with the first left singular vector.
        * Update $c_i, i \in \mathcal{I}_k$ with elements in the first right
        singular vector multiplied by the first singular value.
        * foreach $i \in \mathcal{I}_k$ do
            $\boldsymbol{\psi}^a(m_i + 1 : m_i + M) \leftarrow \bar{\boldsymbol{\psi}}_{m_i}^a - c_i \boldsymbol{\phi}_k^a$.
        end
        • $\boldsymbol{\phi}_k^v$ update.
        $\boldsymbol{\phi}_k^v = \sum_{i \in \mathcal{I}_k} \bar{\boldsymbol{\psi}}_{y_i x_i l_i}^v / \sum_{i \in \mathcal{I}_k} 1$.
    end
end
```

The average scaling factor for the k-th atom is:

$$\bar{c}_k = \sum\nolimits_{i \in \mathcal{I}_k} c_{b_i} / \sum\nolimits_{i \in \mathcal{I}_k} 1, \quad (6)$$

which is for the visual mask generation in the next section.

4. VISUALLY CONSTRAINED TF MASK

Once $\mathcal{D}$ is obtained, the testing sequence $\boldsymbol{\psi}$ can be mapped onto $\mathcal{D}$ to obtain their coding coefficients. As the audio test sequence $\boldsymbol{\psi}^a$ is obtained from mixtures of speech sources contaminated by noise, we will only map the visual sequence $\boldsymbol{\psi}^v$, and then use the coherence and synchrony between audio and visual sequence to predict the audio sequence $\hat{\boldsymbol{\psi}}^a$ which is then used to generate a visual mask $\mathcal{M}_v(m)$. The visual mask is further integrated with the audio mask $\mathcal{M}_a(m, w)$

[1] N is the number of used atoms and the translated versions of these atoms. In practice, $N > K$.

based on a non-linear function.

First, we decompose $\boldsymbol{\psi}^v$ using visual atoms from $\mathcal{D}$ via maximizing the following criterion using MP:

$$[b_i, y_i, x_i, l_i] = \underset{[k,\hat{y},\hat{x},\hat{l}]}{\arg\max} \left(e^{\left(-\overline{(\bar{\boldsymbol{\psi}}^v_{\hat{y}\hat{x}\hat{l}} - \boldsymbol{\phi}^v_k)^2}/\sigma_v^2\right)} \cdot V_{\hat{y}\hat{x}\hat{l}} \right), \quad (7)$$

which is the product of the last two terms in equation (4). $\hat{y}, \hat{x}, \hat{l}$ are all the possible localization parameters of $\boldsymbol{\phi}^v_k$ over $\boldsymbol{\psi}^v$. Using N translated atoms, the bimodal-coherent part $\hat{\boldsymbol{\psi}}^v$ of $\boldsymbol{\psi}^v$ can be approximated as

$$\hat{\psi}^v(y, x, l) = \sum\nolimits_{i=1}^{N} \phi^v_{b_i}(y - y_i, x - x_i, l - l_i), \quad (8)$$

Due to the synchrony and coherence of the AV atoms, the audio sequence $\hat{\boldsymbol{\psi}}^a$ for the target speaker can be predicted by

$$\hat{\psi}^a(m) = \sum\nolimits_{i=1}^{N} \bar{c}_{b_i} \phi^a_{b_i}(m - m_i), \quad (9)$$

where $m_i = \text{round}(l_i * f^a_s / f^v_s)$. By comparing $\psi^a(m)$ (taken directly from the binaural mixtures) and $\hat{\psi}^a(m)$, we can generate a frequency-independent visual mask:

$$\mathcal{M}_v(m) = \begin{cases} 1 & \text{if } \hat{\psi}^a(m) > \psi^a(m) \\ \sqrt{\hat{\psi}^a(m)/\psi^a(m)} & \text{otherwise,} \end{cases} \quad (10)$$

which is then integrated with the audio mask $\mathcal{M}_a(m, w)$ as follows.

In the E step of the first iteration of EM for binaural parameter estimation, we update the posterior probability assuming the first output is the target signal:

$$p(1, \tau | m, w) \leftarrow p(1, \tau | m, w)\mathcal{M}_v(m). \quad (11)$$

This essentially removes components not coming from the target speech. As a result, in the M step, the estimation of the parameters associated with the target speech becomes more accurate.

After the audio mask $\mathcal{M}_a(m, w)$ and the visual mask $\mathcal{M}_v(m)$ are both obtained, we apply the power law transformation to $\mathcal{M}_a(m, w)$, where the frequency-independent power coefficients r are determined on the basis of $\mathcal{M}_v(m)$, as shown in Fig. 1:

$$\mathcal{M}_{av}(m, w) = \mathcal{M}_a(m, w)^{r(\mathcal{M}_v(m))}. \quad (12)$$

Several of the power coefficients (e.g., $r(1) = 4, r(0.25) = 2...$) are fixed, and the rest $r(\mathcal{M}_v(m))$ can be obtained via curve fitting techniques, e.g., spline interpolation. With this power law, $\mathcal{M}_{av}(m, w)$ will be amplified when the $\mathcal{M}_v(m)$ is high (> 0.5), otherwise, it will be attenuated. Finally, we apply the AV mask to either $L(m, w)$ or $R(m, w)$, to obtain the separated source in the TF domain, e.g. $\hat{S}_1(m, w) = \mathcal{M}_{av}(m, w)L(m, w)$. which can be transformed back to the time domain to obtain the audio-visual separated signal.

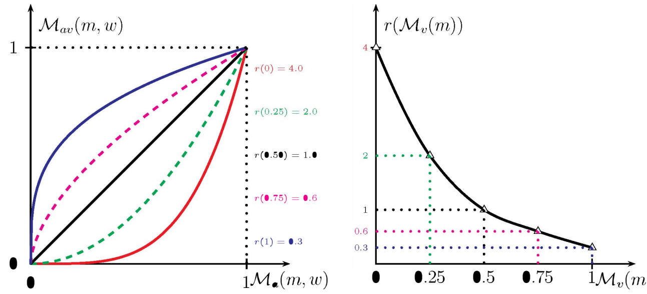

Fig. 1. Visually constrained TF mask generation.

5. EXPERIMENTAL RESULTS

5.1. Data setup

The audio-visual data used in our experiments were from the XM2VTS database [8]. We selected 4 sequences of a target speaker (subject ID: 38) reading digits in two different sessions. Three of them were concatenated for training the AV dictionary, lasting about 56 seconds with $\tilde{L} = 1442$ frames in total. The remaining one was used for testing. The sampling rates for audio and video were 16 KHz and 25 Hz, respectively. We first manually cropped a rough mouth region of $86 \times 140(\tilde{Y} = 86, \tilde{X} = 140)$ pixels for each frame as $\boldsymbol{\psi}^v$. The audio energy vector $\boldsymbol{\psi}^a$ was extracted from 400 ms overlapping Hamming window with 300 ms overlap. The audio resolution became $f^a_s = 100$, while $f^v_s = 25$.

5.1.1. *AV dictionary learning and visual mask*

We set the dictionary size $K = 10$, the audio atom length $M = 48$ and the video atom size $Y \times X \times L = 60 \times 120 \times 12$. $IterMax = 100$ iterations were run. At most $N = 96 \approx (0.8 \times 1442)/L$ atoms and their translations were used to represent the audio-visual signal, where 0.8 denotes the sparsity. After the AV dictionary learning, the bimodal-coherent parts were learnt. Three of the learnt atoms are shown in Fig. 2.

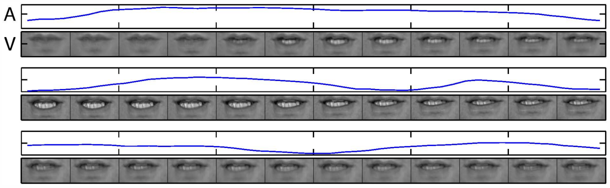

Fig. 2. The AV atoms obtained by the dictionary learning.

Then we can decompose the test video $\boldsymbol{\psi}^v$ using the MP method with equation (7) and reconstruct $\hat{\boldsymbol{\psi}}^a$ using equation (9). Comparison between $\hat{\boldsymbol{\psi}}^a$ and the ground truth is shown in Fig. 3, for an 8-second signal.

5.1.2. *Binaural signals*

We used the Aachen Impulse Response (AIR) database [9] to generate the binaural mixtures. We chose the 'stairway' environment with a dummy head. The target speaker was in front of the dummy head, and we gradually changed the azimuth α of the competing speaker on the right side from 0° to 75° with an angle increment of 15°. We varied the distance d between the speakers and the dummy head (1 m, 2 m and 3 m), introducing different direct-to-reverberant ratios (DRRs).

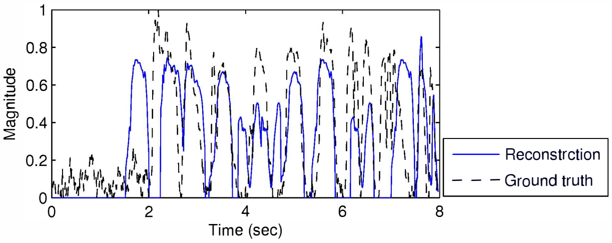

Fig. 3. Reconstructed audio sequence (solid) against the reference from the clean audio sequence (dashed).

The competing speech was randomly chosen from the other audio sequences of the XM2VTS database, composed of digits or other continuous speech. Gaussian white noise (GWN) was added at different signal to noise ratios (SNRs).

5.2. Performance comparison

We compared our method with the benchmark method proposed in [1]. The signal to distortion ratios (SDRs) were used as the performance metric. To investigate how the reverberations influence the performance, we first evaluated the results by varying d with respect to α. For each α, we randomly chose 5 different competing speech signals for separation, and produced the average result for the target speech signal. No noise was added and we used 8-second long signals for the evaluation. With the increase of the distance, the performance improvement increased as well, but still only very modest improvement was achieved, about 0.5 dB on average.

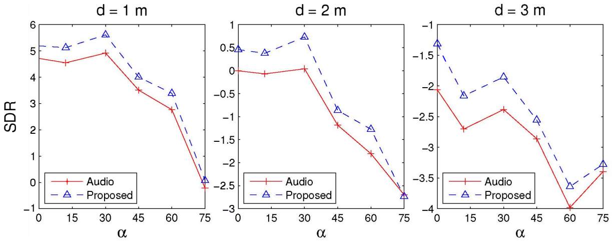

Fig. 4. The SDRs at different source-microphone distance.

We then tested the influence of the noise levels on the performance. GWN was added at [-5 0 5 10] dB, and the distance was fixed at 1 m. Results were still averages over 5 randomly chosen interfering speakers. The robustness to acoustic noise was much more obvious in Fig. 5, especially in a high noise environment. We found that when the speech signal was embedded in noise (SNR$= -5$ dB), our method showed an 1.71 dB improvement over all angles. However, when the noise level is low, e.g. SNR$=$ 10 dB and noise free cases, our method shows only 0.42 dB and 0.53 dB respectively.

6. CONCLUSIONS

A binaural source separation method based on AV dictionary learning is proposed, where visual information is used for the initialization of binaural parameter estimation, as well as tuning the audio masks. The visual information is obtained via the bimodal-coherent AV atoms, learnt with a novel AV dictionary learning method. The proposed algorithm has been tested on the XM2VTS database, and an average of 1.7 dB improvement is achieved in high noise levels, which demonstrates the potential use for noise reduction. However, this dictionary is speaker-dependent, and to learn a more general dictionary from various speakers, we need much more training data, and a more robust visual feature might be extracted to replace the high-dimensional visual data.

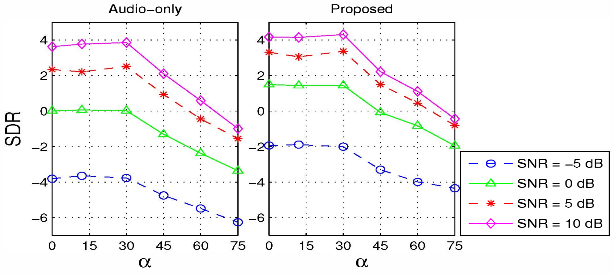

Fig. 5. The SDRs at different SNR levels.

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).

7. REFERENCES

[1] M.I. Mandel, R.J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, and Language Process. (ASLP)*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[2] 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 J. Appl. Signal Process.*, vol. 2002, no. 11, pp. 1165–1173, Jan. 2002.

[3] 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 Trans. ASLP*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[4] Q. Liu, W. Wang, and P.J.B. Jackson, "Use of bimodal coherence to resolve spectral indeterminacy in convolutive BSS," in *Proc. LVA/ICA*, 2010, pp. 131–139.

[5] A.L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation using overcomplete dictionaries," in *Proc. ICASSP*, 2008, pp. 1841–1844.

[6] G. Monaci, P. Vandergheynst, and F.T. Sommer, "Learning bimodal structure in audio-visual data," *IEEE Trans. Neural Netw.*, vol. 20, no. 12, pp. 1898–1910, Dec. 2009.

[7] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Proc.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[8] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "Xm2vtsdb: The extended m2vts database," in *Proc. AVBPA, http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/*, 1999, pp. 72–77.

[9] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. ICDSP*, 2009, pp. 1–5.