Using audio and visual information for single channel speaker separation

Faheem Khan[1,2], *Ben Milner*[1]

[1]School of Computing Sciences, University of East Anglia, Norwich, UK
[2] Department of Software Engineering, University of Science and Technology, Bannu, Pakistan
f.khan@uea.ac.uk, b.milner@uea.ac.uk

10.21437/Interspeech.2015-360

Abstract

This work proposes a method to exploit both audio and visual speech information to extract a target speaker from a mixture of competing speakers. The work begins by taking an effective audio-only method of speaker separation, namely the soft mask method, and modifying its operation to allow visual speech information to improve the separation process. The audio input is taken from a single channel and includes the mixture of speakers, and a separate set of visual features is extracted from each speaker. This allows modification of the separation process to include not only the audio speech but also visual speech from each speaker in the mixture. Experimental results are presented that compare the proposed audio-visual speaker separation with audio-only and visual-only methods using both speech quality and speech intelligibility metrics.

Index Terms: Speaker separation, soft mask, visual features, audio-visual correlation

1. Introduction

The aim of this work is to address the problem of single channel speaker separation by using both audio and visual information. Humans are very good at extracting a target speaker from a mixture of interfering speakers. Having two ears is beneficial but humans also exploit visual speech information from a target speaker. Many audio-only methods of speaker separation have been proposed and have varying levels of success [1, 2, 3, 4]. A smaller number of visual-only methods of speaker separation have also been proposed [5, 6, 7]. However, few approaches have examined whether the audio and visual information can be combined to further improve separation of speakers.

Audio-only speaker separation can be very effective when multiple microphones are used. Techniques such as deconvolution and blind source separation (BSS) make assumptions that the signals in the mixture are independent and exploit the input signals to extract the individual audio sources [8, 1]. Speaker separation from just a single audio channel is substantially more difficult making it necessary to employ knowledge of the way humans perceive speech and to make various assumptions about the speech signals. Most methods exploit the masking property of human speech perception and aim to identify and extract time-frequency regions of the speech mixture that are dominated by the target speaker and mask or attenuate other regions. Binary masking involves determining whether each time-frequency component represents the target speaker or not and is subsequently retained or removed [9, 10]. Soft masking can be better as uncertainty in the mask is allowed, where rather than retaining or removing a time-frequency component, a fraction of the component is retained, generally in proportion to the local signal-to-noise ratio (SNR) [2, 3]. With both methods a major challenge is to estimate accurately the mask and identify time-frequency components to be retained and those which are to be masked. Many approaches have been employed and these typically operate by grouping time-frequency regions according to various criteria. One of the most effective is computational auditory scene analysis (CASA) which groups regions perceptually, making use of cues such as harmonicity and onset and offset times [1]. Alternative approaches have used statistical approaches whereby dependencies between time-frequency regions are established and used to form the mask [4].

There are substantially fewer visual-only methods of speaker separation. These rely on correlation existing between the visual and audio speech features to provide an estimate of the audio feature given a visual feature [11, 12]. Visually-derived audio feature estimates have been used to form a perceptually motivated filter that can extract a target speaker from the mixture [5]. An alternative method uses visually-derived audio features from both speakers in a mixture to estimate a binary mask that extracts the target speaker from the audio mixture [6]. In other applications visual features have been used to improve hidden Markov model (HMM) decoding of input speech signals where the HMMs provide statistics on the speech to be separated [7].

Some work on using both audio and visual speech information for speaker separation has been reported although this is applied to multiple audio channels rather than to a single channel which is the focus of this work. In [13] a target speaker is first extracted from a speech mixture using audio BSS. Visual information from speakers is then used to address permutation and scaling ambiguities present after BSS.

This work proposes combining the audio-only soft mask method with visual speech information to improve speaker separation. A review of the soft-mask method is presented in Section 2. The combination of this with visual speech information is presented in Section 3. Section 4 explains how the necessary audio features are estimated from visual features. Experimental results evaluating the quality and intelligibility of the processed speech are presented in Section 5.

2. Audio-only speaker separation

The soft mask method of speaker separation has been shown to outperform both binary masking and Wiener filtering methods for single channel speaker separation [4]. Consequently this forms the basis for the proposed combined audio-visual method of speaker separation.

In the time-domain, speech from the target speaker, $x_1(n)$, and competing speaker, $x_2(n)$, are assumed to be additive to create the time-domain mixture, $y(n)$. From the time-domain signals, short-time log spectral vectors are extracted, where (adopting the same notation as in [4]) x_{1d} and x_{2d} are the dth elements in the D=128 dimensional vectors extracted from speak-

ers 1 and 2 respectively, and y_d is the dth element extracted from the mixture of the two speakers.

The soft mask method makes an element-wise mixture-maximisation assumption of the log spectral vectors from the speakers in the mixture [14]

$$y_d = \max(x_{1d}, x_{2d}) + e_d \quad d = 1, ..., D \tag{1}$$

where e_d is the error in the mixmax approximation.

An MMSE estimate of each element of the target speaker's log spectral vector, $\hat{x}_{1d}$, is made from the conditional expectation given $\boldsymbol{y}$

$$\hat{x}_{1d} = E(x_{1d}|\boldsymbol{y}) = \int_{x_{1d}} x_{1d} p(x_{1d}|\boldsymbol{y}) dx_{1d} \quad d = 1, ..., D \tag{2}$$

The log spectral features of each speaker are modelled using a Gaussian mixture model (GMM) that comprises I Gaussian subsources for speaker 1 and J subsources for speaker 2. Each subsource from the target speaker has a prior probability, $p_{s_1}(s_1 = i|i = 1, 2, \ldots, I)$ and for the competing speaker $p_{s_2}(s_2 = j|j = 1, 2, \ldots, J)$. The subsources are modelled using Gaussian distributions as

$$p_{\boldsymbol{x}_1|s_1}(\boldsymbol{x}_1|s_1 = i) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{x}_1, \mu_{1d}^i, \Sigma_{1d}^i\right) \tag{3}$$

$$p_{\boldsymbol{x}_2|s_2}(\boldsymbol{x}_2|s_2 = j) = \prod_{d=1}^{D} \mathcal{N}\left(\boldsymbol{x}_2, \mu_{2d}^j, \Sigma_{2d}^j\right) \tag{4}$$

where μ_{1d}^i, μ_{2d}^j, Σ_{1d}^i and Σ_{2d}^j are the means and variances of speakers 1 and 2 and subsources i and j respectively.

Modelling the subsources allows the MMSE estimate of equation 2 to be conditioned on each combination of the subsources, i and j

$$\hat{x}_{1d} = \sum_{i,j} \int_{x_{1d}} x_{1d} p(x_{1d}|\boldsymbol{y}, s_1 = i, s_2 = j) d_{x_{1d}} \times p(s_1 = i, s_2 = j|\boldsymbol{y}) \tag{5}$$

This comprises two factors. The first is a MMSE estimate of x_{1d} given $\boldsymbol{y}$ for a particular combination, i and j, of the subsources. The second factor is the posterior probability of the two subsources given $\boldsymbol{y}$. This can be viewed as a weighted summation, according to the probability of each pair of subsources, of the conditional estimate of x_{1d} from $\boldsymbol{y}$ according to the subsources i and j which, following [4] is evaluated as

$$\begin{aligned} \hat{x}_{1d} &= E(x_{1d}|\boldsymbol{y}) \\ &= \sum_{i,j} p(s_1 = i, s_2 = j|\boldsymbol{y}) \\ &\times \begin{cases} \frac{\sigma_{1d}^{2i}}{\sigma_{1d}^{2i} + \sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i} + \sigma_d^2} \mu_{1d}^i & \text{if } \mu_{1d}^i \geq \mu_{2d}^j \\ \mu_{1d}^i & \text{if } \mu_{1d}^i < \mu_{2d}^j \end{cases} \end{aligned} \tag{6}$$

where σ_{1d}^{2i} and σ_d^2 are the variances of speakers 1 for subsource i and the mixture respectively. For the reduction of computational complexity, it was further shown in [4] following [15] that instead of using the weighted summation of all the subsources, the MMSE estimate can be made from the two most probable subsources that maximize $p(s_1 = i, s_2 = j|\boldsymbol{y})$ and is computed as

$$\hat{x}_{1d} = \begin{cases} \frac{\sigma_{1d}^{2i^*}}{\sigma_{1d}^{2i^*} + \sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i^*} + \sigma_d^2} \mu_{1d}^{i^*} & \text{if } \mu_{1d}^{i^*} \geq \mu_{2d}^{j^*} \\ \mu_{1d}^{i^*} & \text{if } \mu_{1d}^{i^*} < \mu_{2d}^{j^*} \end{cases} \tag{7}$$

where i^* and j^* are representing the two most probable subsources that maximize $p(s_1 = i, s_2 = j|\boldsymbol{y})$.

$$\{i^*, j^*\} = \arg\max_{i,j} p(s_1 = i, s_2 = j|\boldsymbol{y}) \tag{8}$$

The conditional estimate is computed in two ways depending on whether the mean component of the target speaker from the i^*th subsource, $\mu^{i^*}_{1d}$, is greater or less than the mean of the competing speaker from the j^*th subsource, $\mu^{j^*}_{2d}$. This can be likened to binary masking which would set the output to y_d or zero.

3. Audio-visual speaker separation

The audio-only soft mask method can be extended to utilise visual information with the aim of improving estimation of the target speaker's spectral component, $\hat{x}_{1d}$, from the mixture. Following on from equation 7 this has been done in two stages where the first considers just the situation where the target mean component is less than the competing speaker component (i.e. $\mu^{i^*}_{1d} < \mu^{j^*}_{2d}$). Secondly, the situation where the target mean component is greater than the competing speaker component (i.e. $\mu^{i^*}_{1d} \geq \mu^{j^*}_{2d}$) is considered.

3.1. Target mean less than competing mean : $\mu^{i^*}_{1d} < \mu^{j^*}_{2d}$

In binary masking when the target mean is less than the competing mean it is assumed that no information about the target can be obtained from the audio mixture and the estimate is set to zero. The soft mask improves on this by setting the estimate to the target mean, $\mu^{i^*}_{1d}$. The inclusion of visual information allows a further modification to the estimate. In this situation equation 7 is modified by making the estimate a weighted combination of the target mean and an estimate of the target, x^V_{1d}, that is derived from a visual speech feature, $\boldsymbol{v}_1$, extracted from video of the target speaker's mouth

$$\hat{x}_{1d} = \begin{cases} \frac{\sigma_{1d}^{2i^*}}{\sigma_{1d}^{2i^*} + \sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i^*} + \sigma_d^2} \mu_{1d}^{i^*} & \text{if } \mu_{1d}^{i^*} \geq \mu_{2d}^{j^*} \\ \alpha \mu_{1d}^{i^*} + (1 - \alpha) x_{1d}^V & \text{if } \mu_{1d}^{i^*} < \mu_{2d}^{j^*} \end{cases} \tag{9}$$

The weighting term, α, adjusts the contributions made by the target mean and visual component in the estimate, $\hat{x}_{1d}$. The procedure for obtaining the audio estimate x^V_{1d} from a visual speech feature, $\boldsymbol{v}_1$, extracted from video of the target speaker's mouth is explained in Section 4.

3.2. Target mean greater than competing mean : $\mu^{i^*}_{1d} \geq \mu^{j^*}_{2d}$

When the target mean is greater than the competing mean in equation 7 the estimate of the target is made from a Wiener-type weighting of the target mean and input mixture of speakers, y_d.

Similar to equation 9 the visually-derived estimate of the target, x^V_{1d}, can be introduced as

$$\hat{x}_{1d} = \begin{cases} \beta\left(\frac{\sigma_{1d}^{2i^*}}{\sigma_{1d}^{2i^*} + \sigma_d^2} y_d + \frac{\sigma_d^2}{\sigma_{1d}^{2i^*} + \sigma_d^2} \mu_{1d}^{i^*}\right) + (1 - \beta) x_{1d}^V & \text{if } \mu_{1d}^{i^*} \geq \mu_{2d}^{j^*} \\ \alpha \mu_{1d}^{i^*} + (1 - \alpha) x_{1d}^V & \text{if } \mu_{1d}^{i^*} < \mu_{2d}^{j^*} \end{cases} \tag{10}$$

A second weighting term, β, is introduced to adjust the contribution made by the visual information.

4. Audio estimation from visual features

Several studies have shown high levels of correlation to exist between audio and visual speech features [11, 12]. With a combination of log filterbank audio features and 2D-DCT visual

features an audio-visual correlation of R=0.8 is reported. The existence of this correlation has been exploited in both robust speech recognition and audio speech enhancement [12, 16].

For the purpose of audio-visual speaker separation, equations 9 and 10 require an estimate of the dth component of the log spectral vector from the target speaker, x_{1d}^V, that is to be provided by a visual feature vector, $\boldsymbol{v}_1$. Many different visual features have been proposed that have high correlation to audio spectral features and include active appearance models, 2-D DCT and cross-DCT [17, 12]. The 2-D DCT visual feature has been chosen in this work given its high correlation to log spectral features [12]. The 2D-DCT features are computed from 100×100 pixel blocks that are centred around the speaker's mouth and truncated to 15 components in a zig-zag pattern [18].

4.1. Estimation of audio features

Estimation of the log spectral vector begins by creating a GMM to model the joint density of the audio and visual feature vectors from a speaker. A joint feature vector, $\boldsymbol{z}_1$, is first created by augmenting log spectral audio vectors and 2D-DCT visual vectors, $\boldsymbol{a}_1$ and $\boldsymbol{v}_1$, from speaker 1

$$\mathbf{z}_1(t) = [\boldsymbol{a}_1, \boldsymbol{v}_1] \tag{11}$$

Using a training set of joint feature vectors extracted from speaker 1, expectation maximisation (EM) clustering is applied to create a GMM, Φ_1, that models the joint density of the audio and visual features for speaker 1

$$\Phi_1 = \sum_{c=1}^{C} \gamma_1^c \phi_1^c = \sum_{c=1}^{C} \gamma_1^c \, \mathcal{N}(\mathbf{z}_1; \mu_1^c, \boldsymbol{\Sigma}_1^c) \tag{12}$$

The GMM comprises $C = 32$ clusters with the cth cluster having a prior probability, γ_1^c, Gaussian probability density function, ϕ_1^c with mean vector, $\boldsymbol{\mu}_1^c$, and covariance matrix, $\boldsymbol{\Sigma}_1^c$

Given the model of the joint density of audio-visual vectors, Φ_1, an estimate of the log spectral audio vector, $\hat{\boldsymbol{a}}_1$, can be made from a 2D-DCT visual vector extracted from speaker 1's mouth region, $\boldsymbol{v}_1$

$$\hat{\boldsymbol{a}}_1 = \arg\max_{\boldsymbol{a}} \left(p\left(\boldsymbol{a}_1 | \boldsymbol{v}_{1,t}, \Phi_1\right)\right) \tag{13}$$

The log spectral component, x_{1d}^V, is extracted from the dth element of the estimated vector, $\hat{\boldsymbol{a}}_1$.

5. Experimental results

The performance of audio-visual speaker separation is evaluated in this section. The audio-visual speech database used for evaluation is described first. Second, three metrics to measure the quality and intelligibility of the separated speech are defined. Speaker separation results are then presented.

5.1. Audio-visual database

The GRID audio-visual speech database is used in these experiments [19]. A male speaker (speaker 1) is used as the target and a female speaker (speaker 4) is the competing speaker. Of the 1000 utterances spoken by each speaker, 800 are used for training and the remaining 200 for testing. The audio for both speakers was downsampled to a sampling frequency of 8kHz and log spectral vectors extracted at 10ms intervals. The video was upsampled to 100 frames per second to match the audio frame rate.

The test scenario assumes that the two speakers are talking simultaneously and are located close together. Video is captured from each speaker with a separate camera. The mixed audio is created by taking speech from the target speaker and mixing it with speech from the competing speaker that is scaled to create the desired SNR. Other SIRs were also evaluated with similar results obtained. For the tests reported, the male speaker is the target and the female the competing speaker. Each of the 200 test utterances from the male speaker was mixed with a randomly selected utterance from the female speaker with the restriction that no mixture used the same two sentences. Experiments were also undertaken with the speakers reversed so the female became the target and no significant differences were observed in the results.

5.2. Speech quality and intelligibility

Three measures are used to examine the effectiveness of speaker separation. The quality of the target speaker's speech is estimated using the signal-to-interference ratio (SIR) [20]

$$SIR = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{comp}\|^2} \tag{14}$$

where s_{target} and e_{comp} refer to speech from the target speaker and competing speaker respectively. The level of distortion in the target speech is measured using the speech distortion ratio (SDR) as

$$SDR = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{15}$$

where e_{interf}, e_{noise} and e_{artif} are the interference, noise and distortion artefacts present in the speech [20].

An estimate of speech intelligibility is made using a whole word speech recogniser trained on the GRID database [19]. Each utterance follows a grammar containing six words of the following structure *command→colour→preposition→letter→digit→adverb*. From the estimates of the target speaker's speech, MFCC vectors were extracted in accordance with the ETSI XAFE standard [21] and the resulting word accuracy used as an estimate of intelligibility. It should be noted that these recognition tests are used to provide an indication of intelligibility. The methods presented in this work are not a proposed method of speaker separation for speech recognition. For this task, effective methods have been developed that operate on the features themselves without the need to reconstruct an audio signal [22].

5.3. Target mean less than interfering mean

This section examines the effect of introducing visual information into the audio soft mask when the target mean is less than the interfering mean as described by equation 9. The target and competing speakers were mixed to give an initial, uncompensated SNR of 0dB. The variable α controls the ratio of target mean, μ_{1d}^{i*}, to visual information, x_{1d}^V. With $\alpha = 1$ no visual information is used and so the estimate is purely the soft mask result, while with $\alpha = 0$ the output is purely the visual estimate.

Figure 1 shows the SIR, SDR and recognition accuracy when varying α from 0 to 1. A-only in the figure is representing the audio only soft-mask method of equation 7 while AV is representing the audio-visual method of equation 9. The A-only method is not affected by α or in other words $\alpha = 1$

all the time. For AV method, SIR peaks with $\alpha = 0.2$ and begins to drop with $\alpha > 0.3$ which equates to highest quality when the estimate is based largely (80%) on the visual estimate of the target speech. The SDR peaks with α around 0.4 although varying the contributions of the visual estimate and target mean has less effect than observed with the SIR. Recognition performance peaks at α=0.35 with an accuracy of 72% – in clean conditions recognition accuracy for the target speaker is 94%. For all three metrics, including visual information in the AV method, has improved performance over the audio-only case ($\alpha = 1$).

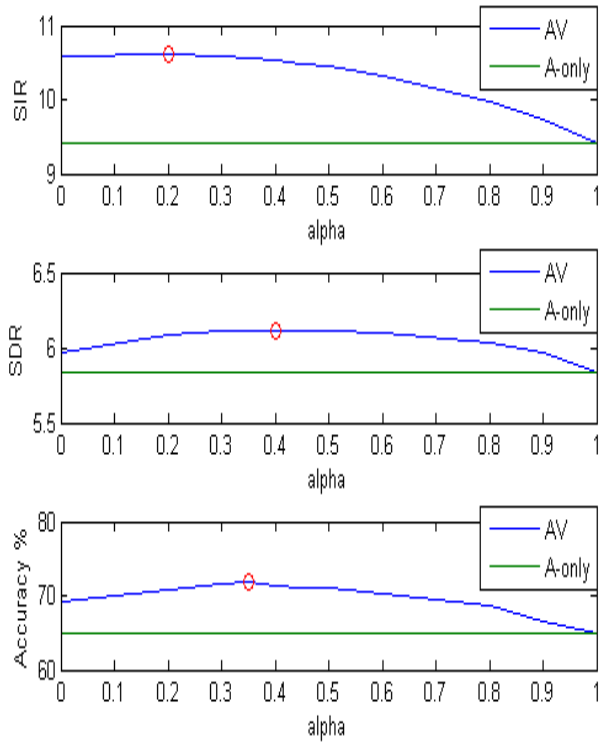

Figure 1: *Comparison of SIR, SDR and recognition accuracy for AV and A-only methods. The circles are indicating the peak values.*

5.4. Target mean greater than competing mean

The effect of adjusting the contribution made from visual information to the target estimate when the target mean is greater than the competing mean is now investigated using the configuration described in equation 10. Figure 2 shows the SIR, SDR and recognition accuracy when varying the visual contribution, β, from 0 to 1.

In situations when the target mean is less than the competing mean, the value α=0.35 is used as determined by the previous experiments. For reference, the performance at $\beta = 1$ corresponds to the situation when no visual information is included in the estimate when target mean is greater than the competing mean and the target spectral estimate is made from audio only which is the original soft mask. At this point ($\beta = 1$ and $\alpha = 0.35$), performance is equal to the best obtained in the results shown in Figure 1. As β reduces the visual information makes more contribution to the estimate. For SIR, SDR and recognition accuracy as more visual information is included,

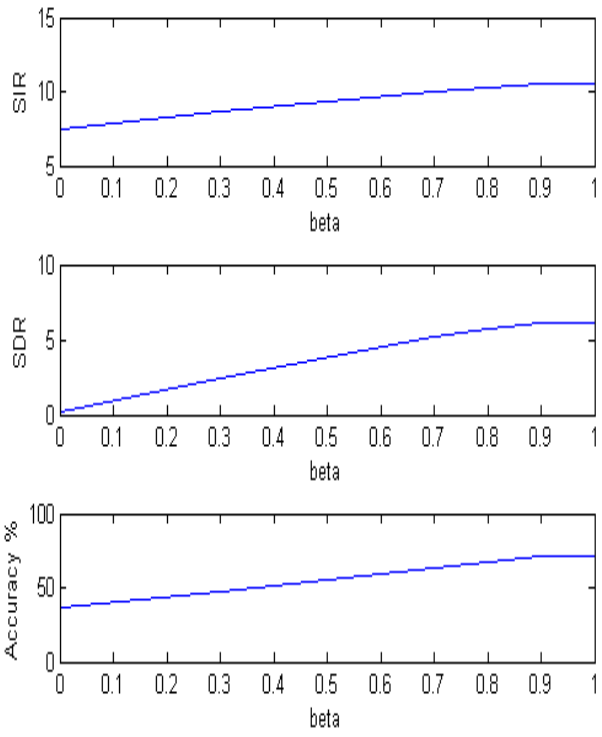

Figure 2: *SIR, SDR and recognition accuracy when $\alpha = 0.35$ and varying β from 0 to 1 in equation 10.*

and thereby audio information reduced, performance falls. All three metrics reach minimum levels when the target estimate is based only on visual information, i.e. $\beta = 0$. Therefore an optimal value of β is one. Hence it is concluded that the introduction of β does not give any improvements and is dropped from further investigation. This suggest that in times when the the target speaker is dominant then the audio information is more useful than the the visual information.

6. Conclusions

This work has shown that the performance, both in terms of quality and intelligibility, of the audio-only method can be improved by including visual speech information. For both the sets of experiments, the performance of the audio-only soft mask is improved by including visual information in the condition when the target mean is less than the competing mean. However, when the target mean is greater than the competing mean, visual information has no positive effect and instead reduces performance. This suggest that in times when the soft mask is confident that the target speaker is dominant then it is preferable to make full use of the audio information through the mixed audio, y_d, and target mean, μ^i_{1d}. Conversely, when the soft mask indicates that the competing speaker is dominant, then a better estimate of the target speaker in that situation is to use a substantial proportion of the visual information and a smaller part of the target mean – a ratio of 2 to 1 was found to be optimal. Further analysis found that the target mean was less than the competing mean for approximately 43% of time-frequency regions meaning that the improvement gained by including visual information was widespread across the target speech.

7. References

[1] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press, 2006.

[2] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1766–1776, August 2007.

[3] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Eurospeech*, 2003, pp. 1009–1012.

[4] M. Radfar and R. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[5] F. Khan and B. Milner, "Speaker separation using visual speech features and single-channel audio," in *Interspeech*, 2013.

[6] ——, "Speaker separation using visually-derived binary masks," in *Proc. AVSP*, 2013.

[7] J. Hershey and M. Casey, "Audio-visual sound separation via hidden Markov models," in *Proc. Neural Information Processing Systems*, 2001.

[8] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2007.

[9] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/114/4/10.1121/1.1610463

[10] Y. Li and D. Wang, "On the optimality of ideal binary time frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230 – 239, 2009.

[11] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal tract and facial behavior," *Speech Communication*, vol. 26, no. 1, pp. 23–43, Oct. 1998.

[12] I. Almajai and B. Milner, "Visually-derived Wiener filters for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1642–1651, Aug. 2011.

[13] Q. Liu, W. Wang, and P. Jackson, "Audio-visual convolutive blind source separation," in *Sensor Signal Processing for Defence (SSPD 2010)*, 2010.

[14] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," $IEEE_{JS}AP, vol.\ 10, no.\ 6, pp.341 - -351, September 2002$.

[15] Y. Ephraim and N. Merhav, "Lower and upper bounds on the minimum mean-square error in composite source signal estimation," *Information Theory, IEEE Transactions on*, vol. 38, no. 6, pp. 1709–1724, Nov 1992.

[16] X. Shao and J. Barker, "Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment," *Speech Communication*, vol. 50, no. 4, pp. 337–17353, Apr 2008.

[17] T.F.Cootes, G. Edwards, and C.J.Taylor, "Active appearance models," *IEEE Trans. PAMI*, vol. 23, no. 6, pp. 691–685, June 2001.

[18] K. Sayood, *Introduction to Data Compression*. Morgan-Kaufmann, 2000.

[19] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *JASA*, vol. 150, no. 5, pp. 2421–2424, nov 2006.

[20] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[21] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," ETSI STQ-Aurora DSR Working Group, ES 202 212 version 1.1.1, Nov. 2003.

[22] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. Green, "The PASCAL CHiME speech separation and recognition challenge," *Computer Speech and Language*, vol. 27, no. 3, pp. 621–633, 2013.