Speaker Separation Using Visual Speech Features and Single-channel Audio

Faheem Khan, Ben Milner

School of Computing Sciences, University of East Anglia, Norwich, UK

f.khan@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work proposes a method of single-channel speaker separation that uses visual speech information to extract a target speaker's speech from a mixture of speakers. The method requires a single audio input and visual features extracted from the mouth region of each speaker in the mixture. The visual information from speakers is used to create a visually-derived Wiener filter. The Wiener filter gains are then non-linearly adjusted by a perceptual gain transform to improve the quality and intelligibility of the target speech. Experimental results are presented that estimate the quality and intelligibility of the extracted target speaker and a comparison is made of different perceptual gain transforms. These show that significant gains are achieved by the application of the perceptual gain function.

Index Terms: Speaker separation, Wiener filter, visual features, audio-visual correlation

1. Introduction

Speaker separation is the process of extracting a target speaker from a mixture of sounds that comprise other speakers and acoustic noise. Audio only speaker separation is well established when multiple microphone channels are available. Techniques such as deconvolution and blind source separation (BSS) make the assumption that the various signals in the mixture are independent and exploit the set of input signals to extract individual audio sources [1]. Other work has considered the more difficult problem of speaker separation from a single audio channel. In this instance prior speaker-dependent statistical knowledge of the speakers is utilised to enable extraction of the target speaker. Methods using spectral masking have been effective at solving this problem and use either hard or soft masks to identify time-frequency regions that belong to a target speaker [2, 3].

Visual speech information from a target speaker's mouth region has also been used in multiple audio channel speaker separation to supplement audio-based methods of extracting a target speaker [4, 5]. For example, in [5] 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. The method still uses multiple audio channels but supplements this information with visual information that increases the quality of the extracted target speech. Visual speech has also been used to aid single channel speaker separation [6] by improving the accuracy of hidden Markov model (HMM) decoding of input speech signals, with the HMMs providing statistics on the speech to be separated.

The proposal in this work places more emphasis on visual speech information from speakers and dispenses with the need for multiple audio channels. The proposed system uses a single microphone as the audio input which receives the mixture of speech from the speakers. Information to enable separation of speakers is provided from visual speech features that are extracted from the mouth region of each speaker in the mixture. Several example scenarios can be envisaged with such a system. A first scenario uses a single microphone and single camera, possibly located together, to extract audio and video. The video captured by the camera contains all the speakers present in the mixture, from which each speaker would need to be identified and tracked, such as in [7, 8]. Visual features for each speaker can then be extracted. A second scenario again uses a single microphone, but now uses a series of individual cameras with each capturing video from each speaker in the mixture. These cameras could again be located centrally and be pre-positioned to capture video at positions where speakers would be located. In comparison to the above scenarios, in audio-only speaker separation a 'zooming in' to a speaker is only possible when multiple microphones are distributed within the environment which is a more complex configuration.

For visual features to be able to provide audio information it is necessary to find audio-visual features that are correlated. Several studies have shown that high levels of correlation exist between audio and visual features extracted from a speaker [9, 10]. For mel-filterbank audio features and 2D-DCT visual features, audio-visual correlation of R=0.8 is reported. This correlation has subsequently been exploited successfully to enable visual speech features to aid in both robust speech recognition and audio speech enhancement [11, 10]. Previous work on visually-derived Wiener filtering has used visual speech information to formulate an audio Wiener filter to enhance noisy speech [10]. The work proposed here extends this framework to speaker separation and also introduces a perceptual transformation of the Wiener filter gains with the aim of improving both the quality and intelligibility of the target speaker.

The proposed method of visually-derived speaker separation is described in Section 2. This requires audio estimates of the target and competing speakers which are estimated from visual speech features and this is discussed in Section 3. Details of the implementation in terms of creating the time-domain target speaker's speech are explained in Section 4. Experimental results are presented in Section 5 to evaluate the proposed method in terms of speech quality and intelligibility.

2. Visually-derived speaker separation

The proposed method of speaker separation exploits the audio-visual correlation between a speaker's mouth shape and the resulting audio signal. Consequently the method requires only a single audio input rather than two or more audio channels as with conventional speaker separation approaches. Information to aid extraction of the target speaker is taken from video inputs for each speaker in the mixture. As a further processing stage, a perceptual transform is applied to the Wiener filter gains to improve speech quality and intelligibility.

2.1. Wiener filtering for speaker separation

In the discrete Fourier transform (DFT)-domain the Wiener filter, $W(k)$, is defined

$$W(k) = \frac{P_{SS}(k)}{P_{SS}(k) + P_{NN}(k)} \quad (1)$$

where $P_{SS}(k)$ and $P_{NN}(k)$ represent the clean speech and noise power spectra, respectively, and k represents the discrete frequency bin. For application to speaker separation the Wiener filter is modified so that the clean speech is replaced by the target speaker, S_1, and the contaminating noise is replaced by the competing speaker, S_2, for a two speaker problem. To obtain the power spectra statistics for the target and competing speakers it is proposed to estimate these from visual speech features taken from the two speakers. Analysis into the correlation of audio and visual speech features has shown that significant levels of correlation exist which make estimation possible [10]. However, the analysis also revealed that insufficient audio-visual correlation is present to make a fine resolution estimate, although estimation of a less spectrally detailed filterbank vector is possible. As a consequence the speaker separation Wiener filter to extract speaker 1, $W_{1,t}^{FB}$, is modified to operate in the filterbank domain and can be defined

$$W_{1,t}^{FB}(i) \quad = \quad \frac{\widehat{a}_{1,t}(i)}{\widehat{a}_{1,t}(i) + \widehat{a}_{2,t}(i)} \quad (2)$$

where $\widehat{a}_{1,t}(i)$ and $\widehat{a}_{2,t}(i)$ are filterbank estimates for the target speaker and competing speaker, i indicates the filterbank channel and t represents the time frame.

2.2. Perceptual gain transformation

A series of perceptually-motivated transformations of the Wiener gains are now considered. These aim to reduce distortion of the target speaker and improve suppression of the competing speaker and are implemented as a perceptual gain transform, $\Pi(.)$. This can be considered a non-linear transformation of the Wiener filter gains and gives a perceptual gain $H(i)$ (Note, subscripts have been dropped for clarity).

$$H(i) = \Pi\big(W^{FB}(i)\big) \quad (3)$$

Four different perceptual gain transformations have been investigated and these can broadly be described as piecewise or parametric. Equations (4) to (7) define the resulting gain functions, H^1 to H^4, and these are plotted in Figure 1.

$$H^1(i) \quad = \quad W^{FB}(i) \quad (4)$$

$$H^2(i) \quad = \quad \begin{cases} W^{FB}(i) & W^{FB}(i) > \alpha \\ 0 & W^{FB}(i) \leq \alpha \end{cases} \quad (5)$$

$$H^3(i) \quad = \quad \big(W^{FB}(i)\big)^3 \quad (6)$$

$$H^4(i) \quad = \quad \begin{cases} 0 & W^{FB}(i) < \beta_1 \\ \big(W^{FB}(i)\big)^2 & \beta_1 \leq W^{FB}(i) \leq \beta_2 \\ W^{FB}(i) & W^{FB}(i) > \beta_2 \end{cases} \quad (7)$$

Gain function H^1 serves as a baseline and is set equal to the Wiener filter gain, W^{FB}. The second function, H^2, restricts the gain so that if it falls below a threshold, α, then it is set to zero. This removes time-frequency regions where the SNR falls below a certain threshold and can be likened to the binary mask method of speech enhancement [2], but now with the mask estimated from visual features. Instead of removing regions with local SNRs below 0dB (corresponding to a gain of 0.5), gain cut-off values of $\alpha = 0.2$, 0.4 and 0.6 have been tested in this work. Gain function H^3 is the cube of the Wiener gain and this has the effect of non-linearly reducing the Wiener gain. Lower gain values experience a considerable downscaling while higher gains are reduced by a smaller factor. The fourth gain function, H^4, is a piecewise function that aims to capture properties of the previous gain functions by dividing the gain into three regions with zero gain, a squared Wiener gain and linear Wiener gain, respectively. Two variables, β_1 and β_2, define these regions and have been set to 0.2 and 0.5 for this work, based on preliminary test results.

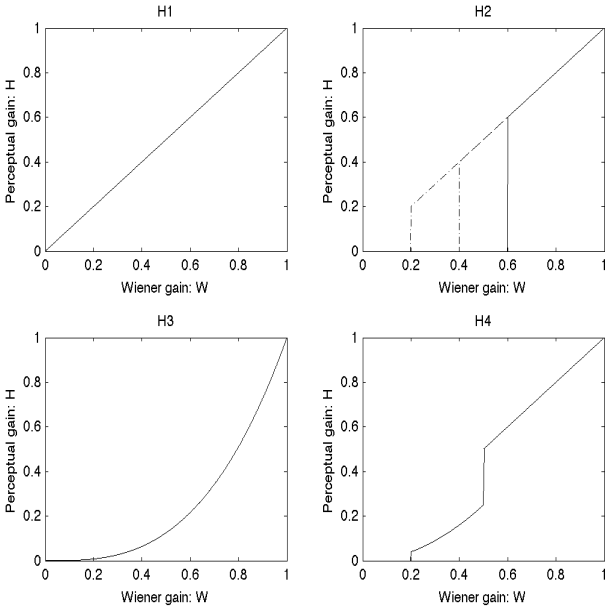

Figure 1: *Perceptual gain functions.*

3. Estimation of audio features from video

The relatively high level of correlation between audio and visual features has led to effective methods of estimating audio features from visual features within a MAP framework [10]. The process involves first training a GMM to model the joint density of audio and visual features. MAP estimation can then be applied to estimate audio features from visual features.

3.1. Audio and visual features

Many different audio and visual features have been developed and in combination have been shown to have varying levels of audio-visual correlation. To accurately estimate audio features from visual features it is necessary to select features that exhibit high levels of audio-visual correlation. As such, based on [10], 23 channel mel-scale filterbank vectors, $\mathbf{a}_t$, are used as the audio features. These are extracted from 20ms duration frames of audio at 10ms intervals in accordance with the ETSI XAFE standard [12]. Visual features, $\mathbf{v}_t$, are extracted from 100x100 pixel regions centered on a speaker's mouth. A 2D-DCT is then

applied and the first 20 coefficients in a zig-zag manner are retained as the visual vector.

3.2. MAP estimation of audio features

The MAP estimation begins by creating a GMM to model the joint density of audio and visual vectors for a speaker. A joint feature vector, $\mathbf{z}_t$, is first created by augmenting audio and visual vectors

$$\mathbf{z}_t = [\mathbf{a}_t, \mathbf{v}_t] \tag{8}$$

From a training set of joint feature vectors, expectation maximisation (EM) clustering is applied to create a GMM, Φ^z, that models the joint density of the audio and visual features

$$\Phi^z = \sum_{c=1}^{C} \alpha_c \phi_c^z = \sum_{c=1}^{C} \alpha_c \, \mathcal{N}(\mathbf{z}; \mu_c^z, \mathbf{\Sigma}_c^z) \tag{9}$$

The GMM comprises C clusters, with the cth cluster represented by prior probability, α_c, Gaussian probability density function, ϕ_c^z with mean vector, μ_c^z, and covariance matrix, $\mathbf{\Sigma}_c^z$

Given the model of the joint density of audio-visual vectors, Φ^z, a MAP estimate of the audio vector for the target speaker, $\widehat{\mathbf{a}}_{1,t}$, can be made from a visual vector extracted from speaker 1's mouth region, $\mathbf{v}_{1,t}$

$$\widehat{\mathbf{a}}_{1,t} = \arg\max_{\mathbf{a}} \left(p\left(\mathbf{a} | \mathbf{v}_{1,t}, \Phi^z\right)\right) \tag{10}$$

Similarly, a visual vector extracted from speaker 2's mouth region, $\mathbf{v}_{2,t}$, can be applied to equation (10) to give an estimate of the audio vector for speaker 2, $\widehat{\mathbf{a}}_{2,t}$. Together, these filterbank estimates define the filterbank-domain Wiener filter.

4. Implementation

This section outlines briefly the stages involved in applying visually-derived speaker separation to extract a target speaker.

4.1. Perceptual gain calculation

The first stage involves utilising the visual speech features to calculate the perceptual gain, $H(i)$, and is summarised below:

1. Extract visual vectors, $\mathbf{v}_{1,t}$ and $\mathbf{v}_{2,t}$, from the video sequences from the target and competing speakers.
2. Estimate audio filterbank vectors, $\hat{\mathbf{a}}_{1,t}$ and $\hat{\mathbf{a}}_{2,t}$, for the two speakers from visual features using equation (10).
3. Construct visually-derived Wiener filter of equation (2).
4. Apply perceptual gain transforms to the Wiener filter from equations (4) to (7) to give perceptual gain, $H_t(i)$.

This gives a 23-D filterbank-domain perceptual gain function.

4.2. Speaker separation

From the single channel audio input that comprises the mixed speech, short duration frames of speech are extracted and the magnitude spectrum, $|X_t(k)|$ and phase, $\angle X_t(k)$, computed. The perceptual gain can now be applied to the magnitude spectrum of the mixed speech to extract the target speaker. However, before this can be applied the 23-D filterbank-domain perceptual gain must be interpolated to the dimensionality of the magnitude spectrum, which in this work is K=128 spectral bins. This is achieved using cubic spline interpolation. The magnitude spectrum estimate of the target speaker, $|\widehat{S_{1,t}}(k)|$, can now be computed

$$|\widehat{S_{1,t}}(k)| = H_t(k)|X_t(k)| \tag{11}$$

The magnitude spectrum estimate of the target speaker is now combined with the phase of the mixed speech, $\angle X_t(k)$, and an inverse Fourier transform applied to obtain a short-duration frame of time-domain samples. Overlap and adding of frames gives the final estimate of the target speaker's speech.

5. Experimental results

This section evaluates the effectiveness of the proposed method of visually-derived speaker separation. First the audio-visual databases used for evaluation are described. Secondly, two sets of experimental results are presented that estimate the quality and the intelligibility of the target speaker's speech following visually-derived speaker separation.

5.1. Audio-visual databases

The audio-visual data used in the experiments is taken from two audio-visual speech databases, one extracted from a UK male speaker and the other from a UK female speaker [13, 14]. Both databases comprises a set of 279 phonetically rich sentences that were typically 3 to 5 seconds in duration. For both speakers the first 200 utterances were used for training with the remaining 79 utterances used for testing. The audio in both databases was downsampled to a sampling frequency of 8kHz and filterbank vectors extracted at 10ms intervals. The video was upsampled to 100 frames per second to match the audio frame rate. For both speakers, video was captured from the front of the face using a 100×100 pixel region centered on the speaker's mouth.

The experimental scenario investigated is of two speakers talking simultaneously and being located close together in space, with the male speaker the target and the female the competing speaker. Of the two example scenarios discussed in Section 1 this corresponds to the second with video from each speaker captured from separate cameras. The audio was created by taking speech from the target speaker and adding it to scaled speech from the competing speaker, where the scaling was adjusted to create a desired signal-to-interference ratio (SIR). Each of the 79 test utterances from the male speaker were mixed with a randomly selected utterance from the female speaker with the proviso that no mixture used the same two sentences. Unreported experiments were also carried out with the speakers reversed with no significant differences in performance observed. MAP estimation of audio features from visual features used speaker-dependent GMMs that were trained on each speaker.

5.2. Speech quality

To estimate the quality of the target speaker's speech the signal-to-interference ratio (SIR) is used and defined as [15]

$$SIR = 10 \log_{10} \frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{12}$$

where s_{target} and e_{interf} refer to speech from the target speaker and competing speaker respectively. Tests were carried out at initial SIRs of -10dB, -5dB, 0dB, 10dB and 20dB. Visually-derived speaker separation was applied to the mixtures using the four perceptual gain functions introduced in Section 2.2 and the resulting SIRs computed using the BSS toolbox [16] and the results shown in Table 1. The results show that using the Wiener gain, H^1, gives a good increase in quality, particularly at the lower SIRs. Applying a perceptual gain transform gives further increases in the output SIR. With the exception of

Input SIR	-10dB	-5dB	0dB	10dB	20dB
H^1	-3.35	-0.14	3.59	11.87	20.34
H^2, α=0.2	-2.30	0.56	4.05	12.05	20.39
H^2, α=0.4	-0.51	1.78	5.07	12.58	20.57
H^2, α=0.6	**1.77**	3.49	6.51	13.65	21.10
H^3	1.20	**3.63**	**6.88**	**14.44**	**21.94**
H^4	-0.73	1.76	5.04	12.65	20.65

Table 1: *Comparison of input and output SIRs for different perceptual gain functions.*

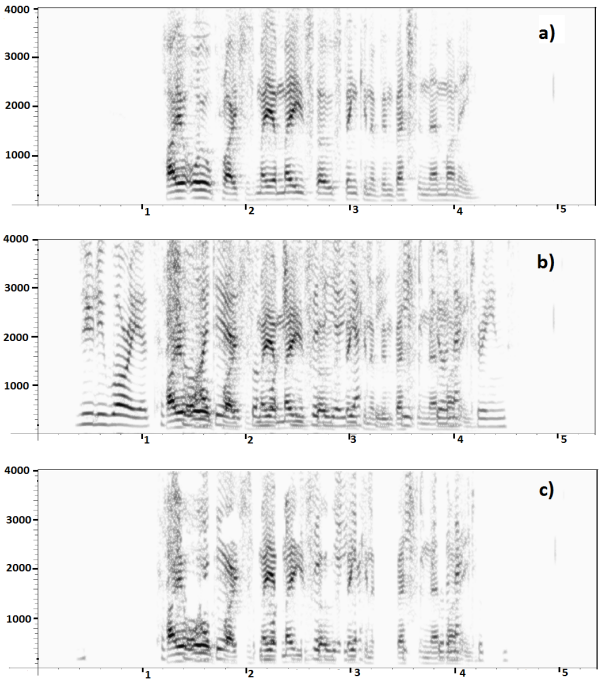

Figure 2: *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) target speaker mixed with a competing speaker at an SIR of 0dB saying* 'Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*', c) visually-derived speaker separation using perceptual gain function* H^2.

-10dB, the cube gain function H^3 gives best performance, with gain function H^2 (with α=0.6) being very close. This corresponds to the Wiener gain with spectral masking below an SIR of 1.8dB (α=0.6). Lowering the point of spectral masking reduces speech quality in terms of the output SIR.

The effectiveness of speaker separation is illustrated in Figure 2 which shows spectrograms of an utterance from the target speaker (Figure 2a), the resulting mixture with a competing speaker at an SIR of 0dB (Figure 2b) and finally the result of visually-derived speaker separation using H^2 (Figure 2c). This shows many of the attributes of the target speaker to have been successfully extracted from the mixture.

5.3. Speech intelligibility

In addition to speech quality it is useful to know whether the proposed visually-derived speaker separation is able to increase the intelligibility of the target speaker's speech. To provide an estimate of speech intelligibility an unconstrained monophone speech recogniser was employed. This comprised a set of 44 monophone HMMs that were arranged in a fully connected grammar. From the time-domain estimates of the target speaker's speech, MFCC vectors were extracted in accordance with the ETSI XAFE standard [12]. Table 2 shows recognition accuracy for the target speaker using various perceptual gain functions at SIRs from -10dB to +20dB. The entry named NSS shows results when no speaker separation has been applied. It should be noted that these speech recognition tests are included to provide an indication of intelligibility and not as a proposed method of speaker separation for speech recognition. For this task, effective methods have been developed that operate on the features themselves without reconstructing an audio signal [17].

SIR	-10dB	-5dB	0dB	10dB	20dB
NSS	-7.34	-7.73	-3.30	8.88	28.84
H^1	7.90	9.23	12.77	20.58	**33.82**
H^2, α=0.2	**14.77**	**15.13**	14.77	22.29	33.77
H^2, α=0.4	10.70	13.48	**16.63**	**22.74**	33.80
H^2, α=0.6	7.90	10.50	12.53	20.47	32.20
H^3	10.44	13.39	15.63	17.61	29.05
H^4	14.27	14.30	14.54	22.53	33.65

Table 2: *Target speaker monophone recognition accuracy (%) at SIRs from -10dB to +20dB.*

The unconstrained monophone accuracy for the original target speaker in clean conditions is 49.22%. The results show that with no speaker separation, recognition accuracy falls significantly as SIRs reduce with a sizeable drop observed below 20dB. Applying speaker separation using the Wiener gain (i.e. H^1) gives a good increase in recognition accuracy for the target speaker over the uncompensated case. The perceptual gain functions give further increases in recognition accuracy. Consistently best performance is given by H^2 which removes any signal when the Wiener gain is below α. At the very low SNRs of -10dB and -5dB a value of α=0.2 gives best performance, while at higher SNRs a value α=0.4 is better. The piecewise gain function of H^4 also performs well and has highest recognition accuracy when averaged across all SNRs. However, the cubic gain function of H^3 performs less well.

6. Conclusions

This work has shown that visual speech features provide important speaker information that can be used effectively within a single-channel audio speaker separation task. The visual features provide an initial estimate of Wiener gain to extract a target speaker from a speech mixture. Applying a perceptual transform aids the extraction of the target speaker in terms of both speech quality and intelligibility. Of the various perceptual gain function investigated, the most effective across both quality and intelligibility is similar to spectral masking where time-frequency regions are removed when the Wiener gain falls below about 1.8dB. The method proposed uses speaker-dependent models, and while this seems typical of single channel speaker separation methods, it would be desirable to have a speaker-independent system. The high levels of speaker variability in the visual domain make this challenging, but methods of speaker adaptation and speaker-independent visual features are currently being investigated [18].

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, 2007.

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

[4] 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. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 96 –108, Jan 2007.

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

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

[7] E. Ong, Y. Lan, B. Theobald, R. Harvey, and R. Bowden, "Robust facial feature tracking using selected multi-resolution linear predictors," in *In Proc. Int. Conference Computer Vision ICCV09*, 2009, pp. 1483–1490.

[8] F. Huang and T. Chen, "Tracking of multiple faces for human-computer interfaces and virtual environments," in *International Conference on Multimedia and Expo*, vol. 3, 2000, pp. 1563–1566.

[9] 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.

[10] 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.

[11] 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. 337–353, Apr 2008.

[12] 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.

[13] B. Theobald, J. Bangham, I. Matthews, and G. Cawley, "Near-videorealistic synthetic talking faces: Implementation and evaluation," *Speech Communication*, vol. 44, pp. 127–140, Oct. 2004.

[14] B. Theobald, S. F. F. Elisei, and G. Bailly, "LIPS2008: Visual speech synthesis challenge," in *Interspeech*, 2008, pp. 2310–2313.

[15] 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, 2006.

[16] C. Fevotte, R. Gribonval, and E. Vincent, "BSS EVAL toolbox user guide," 2005, available from http://www.irisa.fr/metiss/bss eval/.

[17] 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.

[18] Y. Lan, B. Theobald, R. Harvey, and E. Ong, "Improving visual features for lip-reading," in *International Conference on Auditory-visual Speech Processing (AVSP)*, 2010.