Speaker Separation using Visually-derived Binary Masks

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 paper is concerned with the problem of single-channel speaker separation and exploits visual speech information to aid the separation process. Audio from a mixture of speakers is received from a single microphone and to supplement this, video from each speaker in the mixture is also captured. The visual features are used to create a time-frequency binary mask that identifies regions where the target speaker dominates. These regions are retained and form the estimate of the target speaker's speech. Experimental results compare the visually-derived binary masks with ideal binary masks which shows a useful level of accuracy. The effectiveness of the visually-derived binary mask for speaker separation is then evaluated through estimates of speech quality and speech intelligibility and shows substantial gains over the original mixture.

Index Terms: Speaker separation, binary masks, visual features, audio-visual correlation

1. Introduction

This work examines whether the problem of speaker separation can be achieved through the use of visual speech information. When we as humans listen to audio sounds that comprise a mixture of different speakers we are very good at extracting a target speaker from various interfering speakers. Having two ears improves the situation but we also exploit other cues such as observing visual speech information from the speakers. This work considers the scenario of a single audio channel and examines whether visual speech information can provide information to allow extraction of a target speaker from this mixture of sounds.

Traditional audio-only speaker separation is well established when multiple microphones are present. 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 [1]. The problem of speaker separation from just a single audio channel is substantially more difficult. In this situation it is necessary to employ some 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 out other regions. These masks are known as binary masks and each time-frequency component is set to either one or zero depending on whether the region is dominated by the target speaker or is to be masked. The challenge is to estimate accurately the mask and identify time-frequency components to be retained and those which are to be masked. Various approaches have been employed to find the mask 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, common spatial location, amplitude and frequency modulation 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 [2, 3]. An extension of the binary mask is the soft mask, where instead of a binary decision as to whether a time-frequency component is masked a probability of masking is computed which thereby allows some uncertainty to exist in the mask [2, 4].

This work proposes using visual speech information from each speaker in the mixture to estimate the binary mask. Significant correlation exists between audio and visual speech features extracted from a speaker and this can be exploited to enable audio features to be estimated from visual features [5, 6]. Given audio feature estimates for the speakers in the mixture an estimate of the binary mask can be made from which the target speaker can be extracted. 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 by 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 a 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 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.

Other work on speaker separation has also exploited visual speech information from a target speaker's mouth region. For example in a multiple audio channel speaker separation system visual speech information has been used to supplement audio-based methods of extracting a target speaker [9, 10]. In [10] 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 [11] 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 proposed method of visually-derived binary mask estimation for speaker separation is described in Section 2. To compute the mask requires audio estimates of the target and competing speakers and these are estimated from visual speech features which is discussed in Section 3. Experimental results are presented in Section 4 which first examine the accuracy of the visually-derived binary mask and then evaluate the extracted target speaker's speech in terms of speech quality and intelligibility.

2. Visually-derived binary masks

Speaker separation using binary masks involves first the estimation of a time-frequency mask where each component signifies whether that time-frequency component is dominated by either the target speaker or interfering speakers. Areas where the binary mask indicates the region is target-dominated are retained, while regions that are dominated by interfering speakers are masked and discarded. This work exploits audio-visual correlation and proposes a method of estimating the binary mask using visual speech information.

2.1. Mixing Model

In the time-domain it is assumed that a mixed signal, $x(n)$, is made from the addition of speech from a target speaker and an interfering speaker, $s_1(n)$ and $s_2(n)$, where

$$x(n) = s_1(n) + s_2(n) \tag{1}$$

In the power spectrum, assuming the signals are uncorrelated and the analysis window sufficiently long, then

$$|X(f)|^2 = |S_1(f)|^2 + |S_2(f)|^2 \tag{2}$$

where $|X(f)|^2$, $|S_1(f)|^2$ and $|S_2(f)|^2$ are the power spectra of the mixture and the two speech signals respectively, where f represents the spectral bin.

2.2. Estimation of binary mask

The proposal in this work is to use information from visual speech features taken from both the target speaker and interfering speaker to estimate the binary mask. Analysis of audio and visual speech features has shown that significant correlation exists between the two, enabling audio speech features to be estimated from visual speech features [6]. In particular, broad spectral envelope features such as log filterbank or MFCC features can be estimated from 2D-DCT or Active Appearance Model (AAM)visual features with good accuracy. An advantage of such a visually-derived estimate is that the resulting audio features are free from any interference from other speakers or any other sound sources. Estimation of fine spectral detail, such as harmonic frequencies, is not possible from the visual features as they do not contain source information but a smoothed spectral representation is attainable.

From the target speaker and interfering speaker visual features, $\mathbf{v}_1(t)$ and $\mathbf{v}_2(t)$ are extracted at each time frame, t. From the two visual features, estimates of audio features, $\hat{\mathbf{a}}_1(t)$ and $\hat{\mathbf{a}}_2(t)$, are made using MAP estimation

$$\begin{aligned} \hat{\mathbf{a}}_1(t) &= MAP(\mathbf{v}_1(t)) \\ \hat{\mathbf{a}}_2(t) &= MAP(\mathbf{v}_2(t)) \end{aligned} \tag{3}$$

where the estimation is shown by the function $MAP()$. The process of estimating audio features from visual features is explained in Section 3. In this work the visual features are formed from a 2D-DCT of a 100x100 pixel region centered around each speaker's mouth, while the audio features are from a D-dimensional log filterbank.

To compute the binary mask, the D-dimensional log filterbank vector must be interpolated to the dimensionality of the power spectral features which in this work is F=128, and $D < 128$. This is achieved by cubic spline interpolation to give time-frequency spectral representations for the target and interfering speakers, $A_1(t,f)$ and $A_2(t,f)$

$$\begin{aligned} A_1(t,f) &= interp(\hat{\mathbf{a}}_1(t)) \\ A_2(t,f) &= interp(\hat{\mathbf{a}}_2(t)) \qquad 1 \le t \le T \end{aligned} \tag{4}$$

where T is the number of time frames in the utterance. The estimate of the binary mask, $\hat{m}(t,f)$, is now computed in the normal way whereby time-frequency regions are retained when the target speaker's energy is greater than that of the interfering speaker, or in other words when the local signal-to-noise ratio (SNR) is greater than 0dB

$$\hat{m}(t,f) = \begin{cases} 1 & A_1(t,f) \ge A_2(t,f) \\ 0 & A_1(t,f) < A_2(t,f) \end{cases} \tag{5}$$

This is based on the log-max assumption which assumes that in any particular frequency band at any time, the energy contribution of one speaker in the mixture is dominant and masks the other speakers in the mixture [4].

2.3. Time-domain reconstruction

From the time-frequency representation of the mixed signal magnitude spectrum, $|X(t,f)|$, an estimate of the magnitude spectrum of the target speaker, $|\hat{S}_1(t,f)|$, can be made using the estimated binary mask

$$|\hat{S}_1(t,f)| = \hat{m}(t,f)|X(t,f)| \qquad 1 \le t \le T, 1 \le f \le F \tag{6}$$

The sequence of magnitude spectral frames of the extracted target speech must now be transformed into a continuous time-domain speech signal, $\hat{s}_1(n)$. This is achieved by first combining each magnitude spectrum estimate with the phase of the original mixed speech signal, $\angle X(t,f)$, and applying an inverse Fourier transform to obtain a short-duration frame of time-domain samples. These frames are then overlapped by 50% and added together to create the estimate of the target speaker's speech.

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 [6]. The process involves first training a GMM to model the joint density of audio and visual speech features. MAP estimation can then be applied to estimate audio features from visual features.

3.1. Audio and visual features

For visual features 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 [5, 6]. 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 to enable visual speech features to

aid in both robust speech recognition and audio speech enhancement [12, 6]. As such, based on [6], a D-channel mel-scale filterbank is used as the audio feature. These are extracted from 20ms duration frames of audio at 10ms intervals in accordance with the ETSI XAFE standard [13]. Visual features, $\mathbf{v}_t$, are extracted from 100x100 pixel regions centered on a speaker's mouth. A 2D-DCT is applied and the first 20 coefficients retained as the visual vector. The dimensionality of the filterbank, D, is an important parameter in maximising the accuracy of mask estimation and is examined further in Section 4.2.

3.2. MAP estimation of audio features

MAP estimation begins by creating a GMM to model the joint density of audio and visual feature vectors for a speaker. A joint feature vector, $\mathbf{z}_1(t)$, is first created by augmenting audio and visual vectors from speaker 1

$$\mathbf{z}_1(t) = [\mathbf{a}_1(t), \mathbf{v}_1(t)] \tag{7}$$

From a training set of joint feature vectors, 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

$$\Phi_1 = \sum_{c=1}^{C} \alpha_1^c \phi_1^c = \sum_{c=1}^{C} \alpha_1^c \, \mathcal{N}(\mathbf{z}_1; \mu_1^c, \mathbf{\Sigma}_1^c) \tag{8}$$

The GMM comprises C clusters, with the cth cluster represented by prior probability, α_1^c, Gaussian probability density function, ϕ_1^c with mean vector, μ_1^c, and covariance matrix, $\mathbf{\Sigma}_1^c$

Given the model of the joint density of audio-visual vectors, Φ_1, 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, $\mathbf{v}_1(t)$

$$\widehat{\mathbf{a}}_1(t) = \arg\max_{\mathbf{a}} \left(p\left(\mathbf{a} | \mathbf{v}_{1,t}, \Phi_1\right)\right) \tag{9}$$

Similarly, to estimate filterbank vectors for speaker 2, a GMM, Φ_2, is trained on joint feature vectors extracted from speaker 2, i.e.

$$\mathbf{z}_2(t) = [\mathbf{a}_2(t), \mathbf{v}_2(t)] \tag{10}$$

This GMM is used in equation (9), along with visual vectors extracted from speaker 2, $\mathbf{v}_2(t)$, to give an estimate of the audio vector for speaker 2, $\widehat{\mathbf{a}}_{2,t}$. At present the requirement of speaker-specific GMMs is necessary to attain good audio feature estimates as speaker variability is high for visual features [14].

4. Experimental results

An evaluation of the effectiveness of the visually-derived binary mask for speaker separation is made in this section. The audio-visual speech databases used for evaluation are described first. Second, an analysis of the accuracy of the visually-derived binary mask is presented. Finally, experimental results are presented on the quality and the intelligibility of the target speaker's speech following visually-derived speaker separation.

4.1. Audio-visual databases

Two audio-visual speech databases are used in the experiments – one for the target speaker and one for the interfering speaker. One database is extracted from a UK male speaker and the other from a UK female speaker [15, 16], with both comprising 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, visual features were 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 interfering speaker. Of the two example scenarios discussed in Section 1, this corresponds to the second with video from each speaker captured with separate cameras. The mixed audio signal was created by taking speech from the target speaker and adding it to scaled speech from the interfering speaker, where the scaling was adjusted to create a desired signal-to-interference ratio (SIR).

For evaluation purposes, 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.

4.2. Mask accuracy

The accuracy of the visually-derived binary mask is evaluated by comparing with the ideal binary mask that is computed from the actual energy levels in the target and interfering speakers at each time-frequency point. The metric used for evaluation is the percentage of components in the visually-derived mask that were estimated correctly. The experiments examine the effect of different numbers of filterbank channels (from $D = 2$ to $D = 100$) and at at SIRs from -10dB to +20dB, which are reported in Table 1. The results show that mask accuracy improves slightly with increasing numbers of filterbank channels but this increase varies only by at most around 4%.

To investigate further the effect of varying the number of filterbank channels, an artificial test was carried out that took the ideal binary masks calculated from 2, 6, 12, 18, 23, 27, 30, 50 and 100-dimensional ideal filterbank features interpolated to 128 dimensions. Table 2 compares the accuracy of these filtered binary masks to the visually-derived binary masks extracted at an SIR of 0dB. The results for the filtered ideal masks show that the process of filterbank quantisation introduces a substantial reduction in mask accuracy – with quantisation to 2 channels, accuracy is reduced by almost 18%. However, accuracy of the filtered ideal mask does recover rapidly as more filterbank channels are introduced. In comparison, recovery of the visually-derived binary mask is much less – by only 3% in comparison to 11% when moving from 2 to 100 channels. This suggests that there is a fairly low limit on the amount of spectral detail that can be extracted from visual features.

Figure 1 provides further insight into mask estimation and shows the ideal binary mask and then binary masks computed for 2, 23 and 50 channel filterbanks, with each showing the ideal and visually-derived masks. White regions indicate regions that are dominated by the target speaker and are to be retained. Examination reveals that at low numbers of channels the entire time frame is often classed as either target or interfering speaker

due to the lack of spectral detail available. As the number of channels increases, spectral detail improves and so more frequency discrimination is possible. This is certainly evident in the filtered ideal masks, but less discrimination is available from the visually-derived masks as fine spectral detail is not present in the visual features.

SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
D=2	71.57	66.27	67.07	70.37	74.53	82.79
D=6	72.06	67.49	67.60	69.86	74.49	83.20
D=12	73.05	67.43	67.74	70.08	73.95	83.20
D=18	73.76	68.33	67.96	70.39	74.13	83.14
D=23	72.03	66.88	68.30	69.32	74.03	82.04
D=27	73.21	68.44	68.42	70.96	74.80	83.23
D=30	73.19	68.38	68.32	71.54	75.57	83.04
D=50	72.95	68.66	68.96	71.93	75.30	83.09
D=100	**74.38**	**69.70**	**69.95**	**72.59**	**76.13**	**83.61**

Table 1: *Visually-derived mask estimation accuracy (%) at SIRs from -10dB to +20dB and filterbank sizes from 2 to 100 channels.*

Number of channels	Visually-derived	Filtered ideal
D=2	67.07	82.01
D=6	67.60	84.94
D=12	67.74	86.97
D=18	67.96	87.70
D=23	68.30	88.62
D=27	68.42	88.84
D=30	68.32	88.94
D=50	68.96	91.06
D=100	69.95	93.36

Table 2: *Comparison of visually-derived binary mask and ideal binary mask subject to filterbank quantisation, for filterbank sizes from 2 to 100 channels at an SIR of 0dB.*

4.3. Speech quality

To estimate the quality of the target speaker's speech, the signal-to-interference ratio (SIR) is used as defined [17]

$$SIR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{11}$$

where s_{target} and e_{interf} refer to speech from the target speaker and interfering speaker respectively. Tests used the set of 79 mixed sentences and were carried out at initial SIRs of -10dB, -5dB, 0dB, 5dB, 10dB and 20dB. The visually-derived binary masks were applied to the mixtures and the resulting SIRs computed using the BSS toolbox [18] and the results are shown in Table 3. The results show that the visually-derived binary masks are able to extract the target speaker from the mixture and thereby increase the SIRs. Largest gains in SIR occur at the lower input SIRs. The results also show that the number of filterbank channels does not have a large effect on the output SIR which is supported by the findings in Table 1 that showed little differences in mask accuracy for varying number of channels.

The effectiveness of the speaker separation is illustrated in Figure 2 which shows spectrograms of an utterance from the target speaker (Figure 2a), the interfering speaker (Figure 2b),

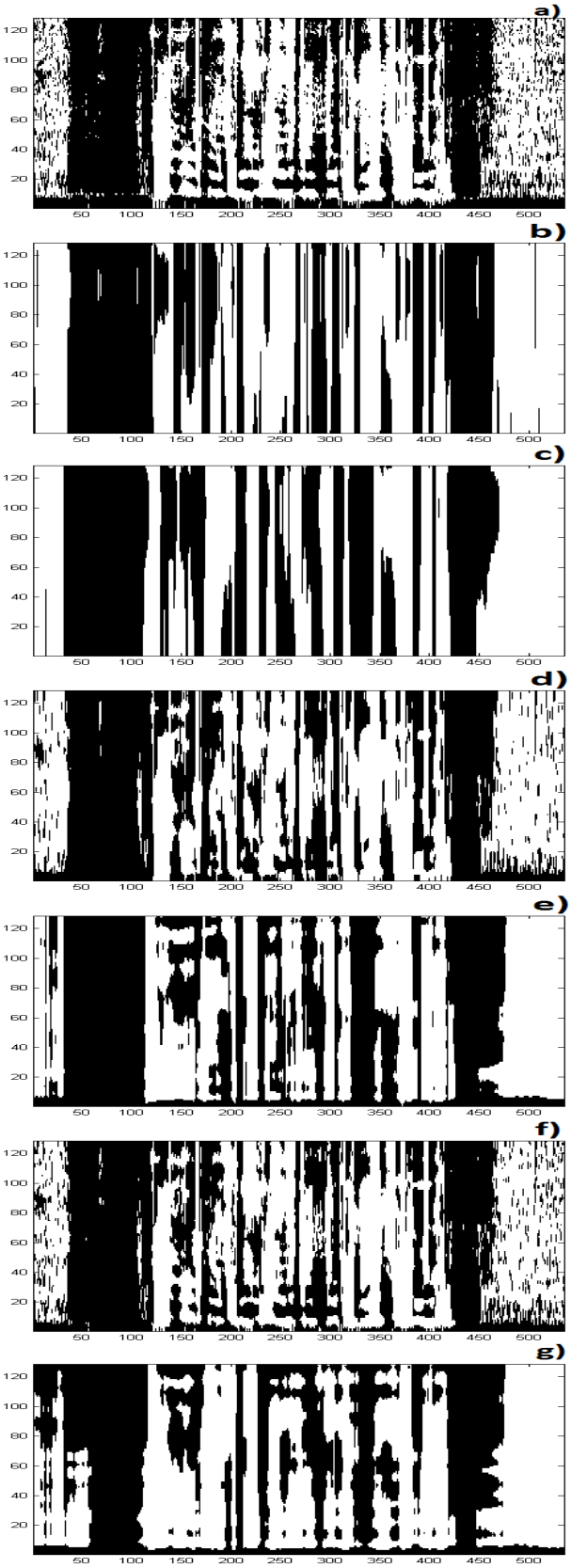

Figure 1: *Binary masks: a) ideal, b) 2-channel ideal, c) 2-channel visually-derived, d) 23-channel ideal, e) 23-channel visually-derived, f) 50-channel ideal, g) 50-channel visually-derived.*

Input SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
D=2	0.06	**2.19**	4.82	8.07	11.73	**20.36**
D=6	-0.11	1.97	**5.03**	**8.13**	**11.91**	20.19
D=12	-0.78	1.16	4.47	7.81	11.53	19.95
D=18	-0.54	1.49	4.29	7.68	11.41	19.82
D=23	-0.19	1.77	3.50	8.03	**11.91**	19.86
D=27	-2.46	-0.03	3.41	7.38	10.94	19.29
D=30	-2.30	-0.26	3.11	7.41	11.34	19.43
D=50	-3.32	-1.02	2.70	6.75	10.83	19.48
D=100	**0.45**	1.44	4.05	7.16	11.10	20.19

Table 3: *Comparison of input and output SIRs for filterbank sizes from 2 to 100 channels.*

the resulting mixture at an SIR of 0dB (Figure 2c) and finally the results of visually-derived binary masking using 2, 23 and 50 filterbank channels. The results show many of the attributes of the target speaker to have been successfully extracted from the mixture.

4.4. Speech intelligibility

This section investigates the effectiveness of speaker separation using the visually-derived binary mask in terms of speech intelligibility. In this work an estimate of speech intelligibility is made using an unconstrained monophone speech recogniser. This comprised a set of 44 monophone HMMs that were arranged in a fully connected grammar. From the masked time-domain estimates of the target speaker's speech, MFCC vectors were extracted in accordance with the ETSI XAFE standard [13]. Table 4 shows recognition accuracy for the target speaker's speech extracted using from 2 to 100 channel filterbanks and at SIRs from -10dB to +20dB. The table also shows baseline performance when no speaker separation (NSS) is applied. Unconstrained monophone accuracy for the original target speaker in clean conditions is 49.22%. 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 [19].

SIR	-10dB	-5dB	0dB	5dB	10dB	20dB
NSS	-7.34	-7.73	-3.30	2.71	8.88	28.84
FB=2	6.81	8.82	11.83	15.10	21.50	35.00
FB=6	7.17	10.79	12.42	15.07	21.68	33.88
FB=12	7.99	9.97	13.18	16.18	21.82	34.95
FB=18	8.20	10.23	13.71	17.16	23.83	35.06
FB=23	9.70	12.53	14.57	18.67	23.27	35.06
FB=27	9.73	12.33	15.92	18.87	**24.59**	35.03
FB=30	9.35	13.24	16.16	**19.43**	24.30	34.97
FB=50	**10.97**	13.74	**16.90**	18.76	24.39	35.21
FB=100	10.91	**14.77**	16.54	17.16	22.56	**35.39**

Table 4: *Target speaker monophone recognition accuracy (%) at SIRs from -10dB to +20dB for filterbank sizes from 2 to 100 channels.*

With no speaker separation (NSS), recognition accuracy falls significantly as SIRs reduce with a sizeable drop observed below 20dB. Applying speaker separation using the visually-derived binary mask improves recognition accuracy for the tar-

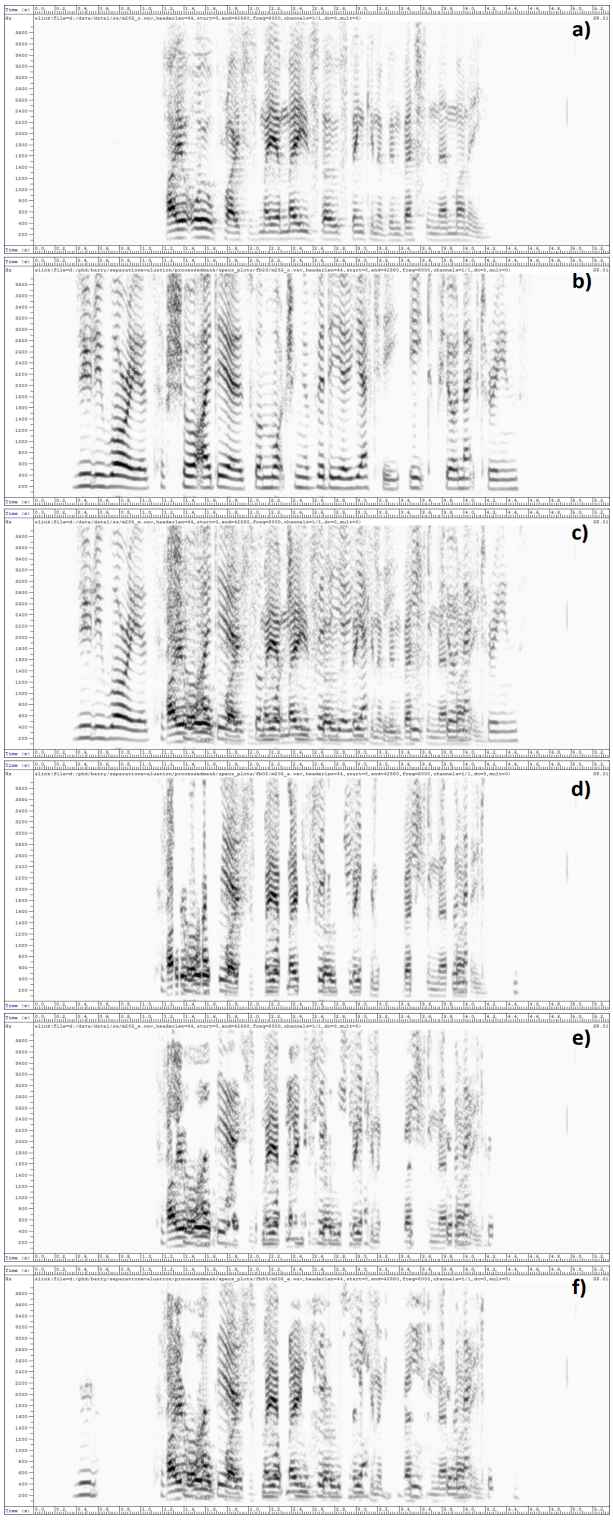

Figure 2: *Spectrograms showing: a) target speaker saying* 'Higher oil prices may amaze those thinking of investing their money*', b) interfering speaker saying '*Zulu warriors have sure ideas when watching a video yeti eat pure nectarines*' c) target speaker mixed with interfering speaker at an SIR of 0dB , d) target speaker extracted using D=2 channels, e) with D=23 channels, f) with D=50 channels.*

get speaker over the uncompensated case. Recognition accuracy consistently increases with increase in numbers of filterbank channels up to 27, but in some cases best recognition accuracy is achieved with 100 channels and in some cases with 50 and 30 channels.

5. Conclusions

This work has shown that visual speech features can provide sufficient spectral information that can be used to create a binary mask for speaker separation purposes. It is observed that the number of filterbank channels does not affect significantly either the mask estimation accuracy or the output SIRs following speaker separation. However, in terms of speech recognition accuracy the method is more sensitive to the number of filterbank channels. At present the proposed method 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 [14].

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

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

[5] 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.

[6] 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.

[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] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] 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.