CO-CHANNEL AUDIOVISUAL SPEECH SEPARATION USING SPECTRAL MATCHING CONSTRAINTS

R. M. Dansereau

Carleton University, Department of Systems & Computer Engineering
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada

ABSTRACT

In this paper the problem of co-channel speech separation for convolutive mixtures is considered where visual cues from one of the speakers is available as side information. The visual cues from the one speaker in the two speaker speech separation are used to estimate the spectral content of the speech and this spectral estimate is in turn used to constrain the solution of the coupling reconstruction filters in the convolutive mixture. The preliminary experimental results show that good performance in speech separation is obtained for our limited case study of visual cues obtained from the spoken numbers of "one" thru "four".

1. INTRODUCTION

Speech is bimodal with an intrinsic coherence between the audition of speech (what we hear) and what is visually observed in the position/motion of the speaker's lips [1]. For the problem of speech separation, we can take advantage of this coherence of acoustic and visual speech by incorporating visual cues such as lip features as side information for the speech separation [2, 3].

In this paper we consider the 2×2 speech separation problem for convolutive coupling mixtures where the difficulty is in the choice of decorrelation filters. In general, the choice of decorrelation filter is not unique so its selection for an appropriate signal separation must be constrained to suit the characteristics of the observed signals [4]. We show an approach of using spectral matching [4] by estimating the acoustic spectrum of the speakers from their lip motion.

2. SPEECH SEPARATION PRELIMINARIES

Consider the 2×2 speech separation problem where two co-existent and independent speech sources are recorded by two spatially separate microphones. The observed signals $y_1(t)$ and $y_2(t)$ by the two microphones are given as [4, 5]

$$y_1(t) = s_1(t) + H_{12}(s_2(t)) \quad (1)$$

$$y_2(t) = s_2(t) + H_{21}(s_1(t)) \quad (2)$$

where $s_1(t)$ and $s_2(t)$ are the speech signals observed at the corresponding microphone in the absence of the competing speech signal, and the systems H_{12} and H_{21} represent the coupling effects between the two channels.

To eliminate the coupling effect between the channels, we can reconstruct the source signals using the follwoing decoupling filter [4]

$$r_1(t) = y_1(t) - G_{12}(r_2(t)) \quad (3)$$

$$r_2(t) = y_2(t) - G_{21}(r_1(t)) \quad (4)$$

where it can easily be verified that if

$$G_{12} = H_{12}, \qquad G_{21} = H_{21} \quad (5)$$

then $r_1(t) = s_1(t)$ and $r_2(t) = s_2(t)$. Hence, the signals $s_1(t)$ and $s_2(t)$ are separated. Similarly, if

$$G_{12} = 1/H_{21}, \qquad G_{21} = 1/H_{12} \quad (6)$$

and H_{12} and H_{21} are invertible, then $r_1(t) = H_{12}(s_2(t))$ and $r_2(t) = H_{21}(s_1(t))$. In this case, $s_1(t) = y_1(t) - r_1(t)$ and $s_2(t) = y_2(t) - r_2(t)$. For either case, the observed source signals $s_1(t)$ and $s_2(t)$ are completely recovered.

The difficulty in recovering $s_1(t)$ and $s_2(t)$ from the convolutive coupling mixture in (1) and (2) is the H_{12} and H_{21} are typically unknown so G_{12} and G_{21} must be blindly estimated. The approach suggested in the following sections makes this estimation of G_{12} and G_{21} only partially blind by using lip position/motion information of at least one speaker to constrain the solution for determining G_{12} and G_{21}. This constraint will also include that the speech sources are independent and statistically uncorrelated.

3. LIP FEATURE EXTRACTION

As indicated in [4], if detailed spectral properties of $s_1(t)$ and $s_2(t)$ are known, then constraining the solution for the frequency response $G_{12}(\omega)$ of G_{12} and $G_{21}(\omega)$ of G_{21} by using spectral matching can be done. To obtain an estimate of the spectral properties of the speech, we propose using the coherence between lip position/motion visual cues as side information.

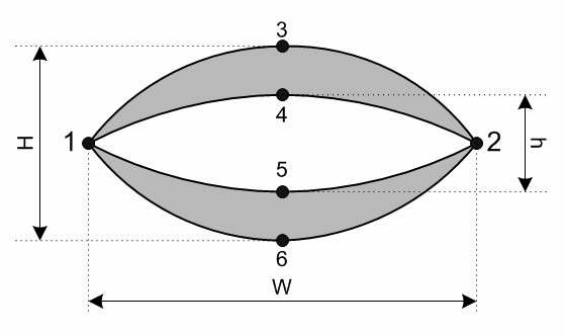

Fig. 1. Geometric feature points on the lips.

To capture the visual cues $v(t)$, the lip feature extraction approach presented in [6] was used where motion, colour, and edge information are combined in a Markov random field (MRF). Then Bayesian labeling within the MRF is used to segment lip/non-lip regions. In this initial study of audiovisual speech separation, after the lip/non-lip region segmentation three lip features are extracted as illustrated in Fig. 1, which include the outer lip height H, the inner lip height h, and the corner-to-corner mouth width W. Future studies will look more extensively at lip contours and active shape modeling such as in [7].

The visual speech used for this study was small and consists of only one speaker repeating the numbers from "one" to "ten" over a set of 30 iterations. It should be noted that only one speaker was used in this initial study so that reasonably reliable mapping of visual cues to phonemes could be done. In pratice, with larger audiovisual databases for different speakers mean error rates for visual speech recognition have been quoted as poor as 40.3% [7].

To model the visual cues, a left-right hidden Markov model (HMM) with continuous density was used that mapped visual cues to word structures. The HMM was designed to have 21 states which includes a state for silence periods. For training the HMM, 25 iterations from the data set were used and the last five kept for testing. The HMM was trained using difference vectors from the extracted lip features in Fig. 1 for sequential frames spaced 25 ms apart. The identified word structures resulting from the HMM can then be used to index into a database of prototypical spectra $P_{v_1}(\omega)$ for that particular phoneme/word structure for the speaker. Of course, since we have limited our study to one speaker we do not have the wider range of spectra normally associated with different speakers. These spectral estimates can now be used for the spectral matching speech separation technique discussed next.

4. SPECTRAL MATCHING

After detailed spectral estimates of the acoustic speech are obtained using the coherence between lip features and speech, the next step proposed is to perform the speech separation by spectral matching as outline in [4]. In the two-speaker case, if visual cues $v_1(t)$ and $v_2(t)$ are available for both competing speakers, then the spectral matching method of Weinstein *et al.* [4] can be applied directly using the spectral estimates of $s_1(t)$ and $s_2(t)$. The more likely scenario though would be to have the visual cues $v_1(t)$ for only one speaker while $v_2(t)$ are unknown. Handling these cases is addressed in this section.

The spectral matching approach to speech separation [4] can be described as follows. To start, the frequency response for the system described by (1) and (2) can be written as

$$\mathcal{H}(\omega) = \begin{bmatrix} 1 & H_{12}(\omega) \\ H_{21}(\omega) & 1 \end{bmatrix} \tag{7}$$

where $H_{12}(\omega)$ and $H_{21}(\omega)$ are the frequency responses for the systems H_{12} and H_{21}, respectively. The reconstruction filter described by (3) and (4) is the inverse of $\mathcal{H}(\omega)$, which can be written as

$$\mathcal{H}^{-1}(\omega) = \frac{1}{1 - G_{12}(\omega)G_{21}(\omega)} \begin{bmatrix} 1 & -G_{12}(\omega) \\ -G_{21}(\omega) & 1 \end{bmatrix} \tag{8}$$

where $G_{12}(\omega)$ and $G_{21}(\omega)$ are estimates of $H_{12}(\omega)$ and $H_{21}(\omega)$, respectively, and for the inverse of $\mathcal{H}(\omega)$ to exist we must have

$$1 - G_{12}(\omega)G_{21}(\omega) \neq 0 \qquad \forall \omega. \tag{9}$$

If $\mathcal{H}(\omega)$ is invertible, (*i.e.*, (9) must hold) then in terms of power spectra the reconstruction filter can be described by

$$\begin{aligned} \begin{bmatrix} P_{r_1r_1}(\omega) & P_{r_1r_2}(\omega) \\ P_{r_2r_1}(\omega) & P_{r_2r_2}(\omega) \end{bmatrix} &= \frac{1}{|1 - G_{12}(\omega)G_{21}(\omega)|^2} \\ \cdot \begin{bmatrix} 1 & -G_{12}(\omega) \\ -G_{21}(\omega) & 1 \end{bmatrix} &\begin{bmatrix} P_{y_1y_1}(\omega) & P_{y_1y_2}(\omega) \\ P_{y_2y_1}(\omega) & P_{y_2y_2}(\omega) \end{bmatrix} \\ \cdot \begin{bmatrix} 1 & -G_{21}^*(\omega) \\ -G_{12}^*(\omega) & 1 \end{bmatrix} \end{aligned} \tag{10}$$

where $P_{r_ir_j}(\omega) i,j = 1,2$ is the auto- and cross-spectra of $r_1(t)$ and $r_2(t)$, and $P_{y_iy_j}(\omega) i,j = 1,2$ are the auto- and cross-spectra of $y_1(t)$ and $y_2(t)$. Multiplying out (10) gives

$$\begin{aligned} |1 - G_{12}(\omega)G_{21}(\omega)|^2 P_{r_1r_1}(\omega) = \\ P_{y_1y_1}(\omega) - G_{12}(\omega)P_{y_2y_1}(\omega) - G_{12}^*(\omega)P_{y_1y_2}(\omega) \\ + |G_{12}(\omega)|^2 P_{y_2y_2}(\omega) \end{aligned} \tag{11}$$

$$\begin{aligned} |1 - G_{12}(\omega)G_{21}(\omega)|^2 P_{r_2r_2}(\omega) = \\ P_{y_2y_2}(\omega) - G_{21}(\omega)P_{y_1y_2}(\omega) - G_{21}^*(\omega)P_{y_2y_1}(\omega) \\ + |G_{21}(\omega)|^2 P_{y_1y_1}(\omega) \end{aligned} \tag{12}$$

$$|1 - G_{12}(\omega)G_{21}(\omega)|^2 P_{r_2r_1}(\omega) = P_{y_2y_1}(\omega) - G_{21}(\omega)P_{y_1y_1}(\omega) - G_{12}^*(\omega)P_{y_2y_2}(\omega) + G_{21}(\omega)G_{12}^*(\omega)P_{y_1y_2}(\omega) \quad (13)$$

where the unknowns are $G_{12}(\omega)$, $G_{21}(\omega)$, and $P_{r_ir_j}(\omega)$.

If visual cues $v_1(t)$ and $v_2(t)$ are observed for the two speakers, then we can estimate the reconstruction auto-spectra as

$$P_{r_1r_1}(\omega) = \alpha_1^2 P_{v_1v_1}(\omega) \approx P_{s_1s_1}(\omega) \quad (14)$$

$$P_{r_2r_2}(\omega) = \alpha_2^2 P_{v_2v_2}(\omega) \approx P_{s_2s_2}(\omega) \quad (15)$$

directly from the visual cues as well as derive estimates for the cross-spectra as

$$P_{r_1r_2}(\omega) = \alpha_1\alpha_2 P_{v_1v_2}(\omega) \approx P_{s_1s_2}(\omega) \quad (16)$$

$$P_{r_2r_1}(\omega) = \alpha_1\alpha_2 P_{v_2v_1}(\omega) \approx P_{s_2s_1}(\omega). \quad (17)$$

The scaling factors α_1 and α_2 are required since the spectral estimates from the visual cues $v_1(t)$ and $v_2(t)$ do not include the static gain/attenuation of the real sources $s_1(t)$ and $s_2(t)$ (*i.e.*, we cannot know how loud the speakers are talking from their lip motion). Using the estimates for $P_{r_1r_1}(\omega)$ and $P_{r_2r_2}(\omega)$, we solve for $G_{12}(\omega)$ and $G_{21}(\omega)$ using (11)-(13) as described in [4].

There may be scenarios where visual cues are available for only one speaker, say for source $s_1(t)$ with visual cues $v_1(t)$, and not for the competing speaker. In this scenario, we can still make the spectral estimate in (14) but the visual cues $v_2(t)$ to estimate $P_{r_2r_2}(\omega)$ (along with the cross-spectra $P_{r_1r_2}(\omega)$ and $P_{r_2r_1}(\omega)$) are not available.

To handle the case when visual cues are available for only one speaker, we consider the power spectra representation for (1) and (2) which can be expressed as

$$\begin{bmatrix} P_{y_1y_1}(\omega) & P_{y_1y_2}(\omega) \\ P_{y_2y_1}(\omega) & P_{y_2y_2}(\omega) \end{bmatrix} = \begin{bmatrix} 1 & H_{12}(\omega) \\ H_{21}(\omega) & 1 \end{bmatrix} \cdot \begin{bmatrix} P_{s_1s_1}(\omega) & P_{s_1s_2}(\omega) \\ P_{s_2s_1}(\omega) & P_{s_2s_2}(\omega) \end{bmatrix} \begin{bmatrix} 1 & H_{21}^*(\omega) \\ H_{12}^*(\omega) & 1 \end{bmatrix} \quad (18)$$

where $P_{s_is_j}(\omega) i,j = 1,2$ is the auto- and cross-spectra of $s_1(t)$ and $s_2(t)$. To simplify (18), we make the assumption that the sources $s_1(t)$ and $s_2(t)$ are zero-mean, statistically uncorrelated wide-sense stationary random processes such that the cross-correlation is

$$\mathcal{E}\{s_1(t)s_2^*(t-\tau)\} = 0 \qquad \forall\tau \quad (19)$$

where $\mathcal{E}\{\cdot\}$ stands for expectation. Using this assumption, the cross-spectra of the speech sources are $P_{s_1s_2}(\omega) = 0$ and $P_{s_2s_1}(\omega) = 0$.

Multiplying out (18) with $P_{s_1s_2}(\omega) = 0$ and $P_{s_2s_1}(\omega) = 0$, we obtain

$$P_{y_1y_1}(\omega) = P_{s_1s_1}(\omega) + H_{12}(\omega)H_{12}^*(\omega)P_{s_2s_2}(\omega) \quad (20)$$

and solving for $P_{s_2s_2}(\omega)$ gives

$$P_{s_2s_2}(\omega) = \frac{P_{y_1y_1}(\omega) - P_{s_1s_1}(\omega)}{H_{12}(\omega)H_{12}^*(\omega)} \quad (21)$$

Substituting (14) into (21), and letting $\beta(\omega) \approx H_{12}(\omega)$ since $H_{12}(\omega)$ is unknown gives

$$P_{r_2r_2}(\omega) = P_{s_2s_2}(\omega) \approx \frac{P_{y_1y_1}(\omega) - \alpha_1^2 P_{v_1v_1}(\omega)}{\beta^2(\omega)} \quad (22)$$

For strictly positive auto-spectra, since $P_{v_1v_1}(\omega)$ must necessarily lie under $P_{y_1y_1}(\omega)$ we can bound α_1 with

$$0 < \alpha_1 \leq \min_{\forall\omega} \sqrt{\frac{P_{y_1y_1}(\omega)}{P_{v_1v_1}(\omega)}}. \quad (23)$$

The function $\beta(\omega)$ is chosen to best estimate the unknown frequency response $H_{12}(\omega)$ the couples $s_2(t)$ to $y_1(t)$ and may be chosen as a piece-wise function or as a constant. If $\beta(\omega)$ is chosen as a constant, then the estimate of $P_{s_2s_2}(\omega)$ in (22) reduces to an instantaneous coupling. When $P_{s_2s_2}(\omega)$ is used in conjunction with $P_{s_1s_1}(\omega)$ in (11)-(13), then the convolutive coupling mixture of (1) and (2) is still being addressed regardless that the estimate for $P_{s_2s_2}(\omega)$ was made for an instantaneous coupling.

5. EXPERIMENTS

To test the proposed audiovisual speech separation system, we simulated a scenario with two microphones and one video camera by separately observing one speaker with one microphone and the camera, and then another speaker with just the microphone. Once the *clean* data was captured, the scenario simulated was for two microphones separated by 120 cm with the two speakers placed at a distance of 30 cm perpendicular to the plane of the microphones. The cross-coupling transfer functions where simulated as unity gain single pole systems with poles at $z = 0.7$ for the transfer function $H_{12}(z)$ and $z = 0.6$ for $H_{21}(z)$.

Using the simulation setup described, example input acoustic speech plots are shown in Fig. 2 where Fig. 2(a) and Fig. 2(b) show the *clean* speech signals as observed for $s_1(t)$ and $s_2(t)$, respectively, and Fig. 2(c) shows the signal mixture $y_1(t)$ for microphone #1. The speech in Fig. 2(a) is for the speaker of interest $s_1(t)$, with associated video signal $v_1(t)$, speaking the numbers "one" to "four". The speech shown in Fig. 2(b) is for the competing speaker $s_2(t)$ and is a sentence from the TIMIT speech database. Note that no visual cues are available for $s_2(t)$.

With the simulation setup complete, the speech separation was performed by first estimating the auto-spectra $P_{r_1r_1}(\omega) = P_{v_1v_1}(\omega)$ from the visual cues $v_1(t)$ and using this estimate as described in Sec. 4 to solve for $G_{12}(\omega)$ and

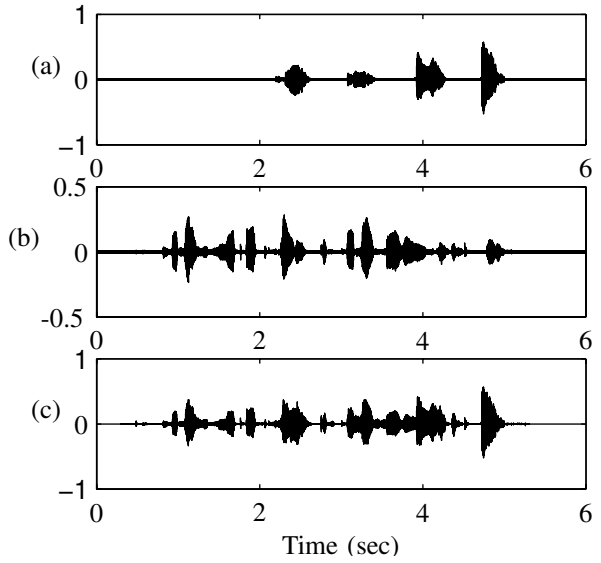

Fig. 2. (a) Speech source $s_1(t)$, (b) speech source $s_2(t)$ of competing speaker, and (c) mixed signal $y_1(t)$.

$G_{21}(\omega)$. Performing the speech separation with the filters $G_{12}(\omega)$ and $G_{21}(\omega)$ resulted in $r_1(t) \approx s_1(t)$ as shown in Fig. 3(a) and $r_2(t) \approx s_2(t)$ as shown in Fig. 3(b). We can see that a reasonably good speech separation has occurred under these simulation conditions.

The speech separation shown in Fig. 3 is quite promising, but note that these results were obtained when a very limited set of word structures were used for the visual cues $v_1(t)$ (namely just the word structures for the spoken numbers "one" to "ten" from one speaker). As indicated, visual speech recognition is generally much poorer with mean error rates of current techniques of around 40.3% [7]. Even with poor visual speech recognition in general, our experimental results still prove promising since if we have the scenario where the speakers are stationary, then the decoupling filters $G_{12}(\omega)$ and $G_{21}(\omega)$ can be estimated over time with increasing accuracies as time progresses.

6. CONCLUSIONS

We have shown how visual cues of a speaker can be used as side information for acoustic speech separation by spectral matching. The 2×2 speech separation problem was considered where visual cues for only one speaker are available. While this initial study considered training a HMM on the visual cues of only a single speaker, it is shown that good separation results are obtained and that if an HMM trained with a more general set of visual cues from a cross section of the population, that as long as the speakers are stationary that the reconstruction filters can be obtained with improved performance over time.

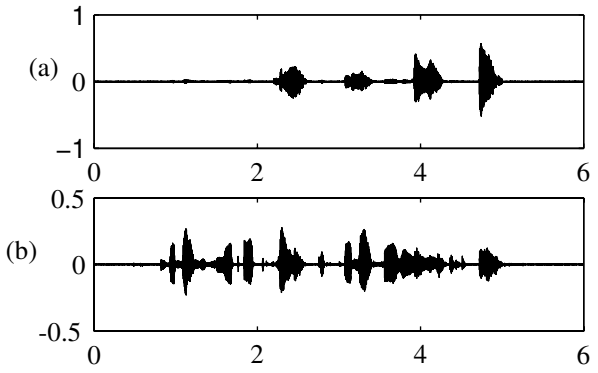

Fig. 3. Separated speech for (a) $r_1(t)$ and (b) $r_2(t)$.

7. REFERENCES

[1] J. Robert-Ribes, J.-L. Schwartz, T. Lallouache and P. Escudier, "Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise," *J. Acoust. Soc. Am.*, vol. 103, no. 6, pp. 3677-3689, 1998.

[2] T. Darrell, J. W. Fisher, P. Viola and W. Freeman, "Audio-visual segmentation and "The cocktail party effect",", *Proc. of the Intern. Conf. on Multimodal Interfaces (Bejing), Oct. 2000.*

[3] L. Girin, A. Allard and J.-L. Schwartz, "Speech signals separation: a new approach exploiting the coherence of audio and visual speech," *IEEE 4th Workshop on Multimedia Signal Process.*, pp. 631–636, 2001.

[4] E. Weinstein, M. Feder and A. V. Oppenheim, "Multichannel signal separation by decorrelation," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 4, pp. 405-413, Oct. 1993.

[5] D. Yellin and E. Weinstein, "Multichannel signal separation: Methods and analysis," *IEEE Trans. Signal Processing*, vol. 44, no. 1, pp. 106–118, Jan. 1996.

[6] R. M. Dansereau, C. Li and R. A. Goubran, "Lip feature extraction using motion, color, and edge information," *IEEE Intern. Workshop on Haptic, Audio and Visual Environ. and Apps. (HAVE'2003)*, pp. 1–6, Sept. 2003.

[7] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, Sept. 2000.