Audio-Visual Speech Recognition With Background Music Using Single-Channel Source Separation

Emad M. Grais, Ibrahim Saygin Topkaya, Hakan Erdogan

Faculty of Engineering and Natural Sciences
Sabanci University, Orhanli, Tuzla, 34956, Istanbul.
{grais,isaygint,haerdogan}@sabanciuniv.edu

ABSTRACT

In this paper, we consider audio-visual speech recognition with background music. The proposed algorithm is an integration of audio-visual speech recognition and single channel source separation (SCSS). We apply the proposed algorithm to recognize spoken speech that is mixed with music signals. First, the SCSS algorithm based on nonnegative matrix factorization (NMF) and spectral masks is used to separate the audio speech signal from the background music in magnitude spectral domain. After speech audio is separated from music, regular audio-visual speech recognition (AVSR) is employed using multi-stream hidden Markov models. Employing two approaches together, we try to improve recognition accuracy by both processing the audio signal with SCSS and supporting the recognition task with visual information. Experimental results show that combining audio-visual speech recognition with source separation gives remarkable improvements in the accuracy of the speech recognition system.

1. INTRODUCTION

One of the challenging problems of automatic speech recognition (ASR) systems is recognizing speech signals when they are mixed with background music or any other signals. The performance of a speech recognition system quickly degrades when there is music in the background. To improve speech recognition performance it would be better to remove music from the speech signal before applying ASR. Augmenting audio information with visual information that is not affected by the background signals will also improve the recognition performance. The need to recognize speech signals that are mixed with background music signals is encountered in many applications such as broadcasting news, songs, documentary programs, and other shows on TV.

Single channel source separation (SSCS) aims to separate the original source signals from a single observed mixture of these source signals. Nonnegative matrix factorization [1] models are trained using training data for each source signal and these models are employed in separating source signals in the observed mixed signal [2, 3].

This research is partially supported by The Scientific and Technological Research Council of Turkey (TUBITAK) under the scientific and technological research support program (code 1001), project number 107E015 entitled "Novel Approaches in Audio Visual Speech Recognition".

In this paper, we combine SCSS techniques with AVSR to recognize speech signals that are mixed with music signals. The aim of the proposed algorithm is to make use of the advantages of combining visual information in the speech recognition process, and also make use of the advantages of separating the speech signal from the mixed signal. We use the NMF algorithm and spectral masks to separate speech signals from the background music signals. We use NMF for SCSS because it yields a fast, efficient, and simple algorithm. Combining NMF with spectral masks gives better separation results than using NMF only [2]. We assume that training audio signals for each source are available. NMF and the training audio data are used to train a set of basis vectors for each source in magnitude spectral domain. After observing the mixed signal, NMF is used to decompose the magnitude spectrogram of the mixed signal with the trained basis vectors for both sources. The decomposition results are used to build a spectral mask. The spectral mask computes the spectrogram of the estimated speech signal by scaling the mixed signal spectrogram according to the contribution of the speech signal in the mixed signal.

Speech recognition from audio with hidden Markov models (HMM) [4] employs hidden states with Gaussian mixture model emissions and Markovian transitions between the states. When there is noise in the audio source and in the presence of visual information, audio-visual speech recognition (AVSR) may be also used which relies on supporting the audio information with visual information [5]. In AVSR the visual features are handled in a separate stream of states thus resulting a multi-stream HMM (MSHMM) [5]. We use the separated speech signal together with visual data as different streams in an MSHMM.

The remainder of this paper is organized as follows: In section 2, we describe the speech-music separation algorithm. In section 3, we show the main procedures of the audio-visual speech recognition which we employ during recognition of the separated speech signal. In the remaining sections, we represent our observations and the results of our experiments.

2. SPEECH-MUSIC SIGNAL SEPARATION

Given an observed mixed signal $y(t)$ which is a mixture of speech $x(t)$ and music signals $m(t)$, we aim to find an estimate for $x(t)$ from $y(t)$. We solve this problem in the short time Fourier transform (STFT) domain. Let $Y(t, f)$ be the STFT of

$y(t)$, where t represents the frame index and f is the frequency-index. Due to linearity of the STFT, we have:

$$Y(t,f) = X(t,f) + M(t,f), \tag{1}$$

$$|Y(t,f)|\, e^{j\phi_Y(t,f)} = |X(t,f)|\, e^{j\phi_X(t,f)} + |M(t,f)|\, e^{j\phi_M(t,f)}. \tag{2}$$

The phase angles are usually ignored in this framework [3]. Hence, we can write the magnitude spectrogram of the measured audio signal as the sum of source signals' magnitude spectrograms as follows:

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{M}. \tag{3}$$

Here $\boldsymbol{X}$ and $\boldsymbol{M}$ are unknown magnitude spectrograms, and need to be estimated using observed data and training speech and music spectra. The magnitude spectrogram for the observed signal $y(t)$ is obtained by taking the magnitude of the DFT of the windowed signal for each column of the spectrogram.

To solve this problem, we use NMF with the magnitude spectra of the training data to train a set of basis vectors for each source as shown in section 2.2. Then NMF is used to decompose the spectrogram of the mixed signal into a weighted linear combination of these trained basis vectors for both sources as shown in section 2.3. The weighted sum of the decomposion terms that include the trained speech basis vectors is used as an initial estimate of the magnitude spectra of the speech signal. The weighted sum of the remaining decomposion terms is used as an initial estimate of the magnitude spectra of the music signal. The initial estimates of both sources are used to build a spectral mask as shown in section 2.4. The spectral mask calculates the spectrogram of the estimated speech signal by scaling every entry of the mixed signal spectrogram according to the contribution of the speech signal in the mixture.

2.1. Non-negative matrix factorization

Non-negative matrix factorization is used to decompose any nonnegative matrix $\boldsymbol{V}$ into a nonnegative basis vectors matrix $\boldsymbol{B}$ and a nonnegative weights matrix $\boldsymbol{W}$.

$$\boldsymbol{V} \approx \boldsymbol{B}\boldsymbol{W}. \tag{4}$$

The matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ can be found by solving the following generalized Kullback-Leibler divergence cost function [1]:

$$\min_{\boldsymbol{B},\boldsymbol{W}} D\left(\boldsymbol{V} \,||\, \boldsymbol{B}\boldsymbol{W}\right), \tag{5}$$

where

$$D\left(\boldsymbol{V} \,||\, \boldsymbol{B}\boldsymbol{W}\right) = \sum_{i,j}\left(\boldsymbol{V}_{i,j}\log\frac{\boldsymbol{V}_{i,j}}{(\boldsymbol{B}\boldsymbol{W})_{i,j}} - \boldsymbol{V}_{i,j} + (\boldsymbol{B}\boldsymbol{W})_{i,j}\right),$$

subject to elements of $\boldsymbol{B}, \boldsymbol{W} \geq 0$. The solution for equation (5) can be computed by alternating updates of $\boldsymbol{B}$ and $\boldsymbol{W}$ as follows:

$$\boldsymbol{B} \leftarrow \boldsymbol{B} \otimes \frac{\frac{\boldsymbol{V}}{\boldsymbol{B}\boldsymbol{W}}\boldsymbol{W}^T}{\mathbf{1}\boldsymbol{W}^T}, \tag{6}$$

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{B}^T\frac{\boldsymbol{V}}{\boldsymbol{B}\boldsymbol{W}}}{\boldsymbol{B}^T\mathbf{1}}, \tag{7}$$

where $\mathbf{1}$ is a matrix of ones with the same size of $\boldsymbol{V}$, the operations $\otimes$ and all divisions are element-wise multiplication and division respectively.

2.2. Training the bases

Given a set of training data of speech and music signals, the magnitude spectrogram $\boldsymbol{X}_{\text{train}}$ and $\boldsymbol{M}_{\text{train}}$ of the training speech and music signals are calculated respectively. NMF uses the two spectrograms to train a set of basis vectors as a model for each source signal. The update rules in equations (6, 7) are used to decompose the magnitude spectrograms into bases and weights positive matrices as follows:

$$\begin{aligned}\boldsymbol{X}_{\text{train}} &\approx \boldsymbol{B}_{\text{speech}}\boldsymbol{W}_{\text{speech}},\\ \boldsymbol{M}_{\text{train}} &\approx \boldsymbol{B}_{\text{music}}\boldsymbol{W}_{\text{music}},\end{aligned} \tag{8}$$

after each iteration, we normalize the columns of $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$. All the matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ are initialized by positive random noise. The bases matrices $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$ are used as trained models for speech and music signals.

2.3. Decomposition of the mixed signal

After observing the mixed signal $y(t)$, the magnitude spectrogram $\boldsymbol{Y}$ of the mixed signal is computed. To find the contribution of every source signal in the mixed signal, NMF is used to decompose the magnitude spectrogram $\boldsymbol{Y}$ of the mixed signal as a linear combination with the trained basis vectors in $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$ as follows:

$$\boldsymbol{Y} \approx \left[\boldsymbol{B}_{\text{speech}}\ ,\ \boldsymbol{B}_{\text{music}}\right]\boldsymbol{W}, \tag{9}$$

where $\boldsymbol{B}_{\text{speech}}$ and $\boldsymbol{B}_{\text{music}}$ are obtained from solving equations in (8). Here we only solve for $\boldsymbol{W}$ in equation (9) using the update rule in equation (7), and the bases matrix is fixed. $\boldsymbol{W}$ is initialized by positive random noise. The initial estimate of the separated speech signal magnitude spectrogram is found by multiplying the bases matrix $\boldsymbol{B}_{\text{speech}}$ with its corresponding weights in matrix $\boldsymbol{W}$ in equation (9). Also the initial estimate of the separated music signal magnitude spectrogram is found by multiplying the bases matrix $\boldsymbol{B}_{\text{music}}$ with its corresponding weights in matrix $\boldsymbol{W}$ in equation (9). The initial magnitude spectrogram estimates for speech and music signals are respectively calculated as follows:

$$\tilde{\boldsymbol{X}} = \boldsymbol{B}_{\text{speech}}\boldsymbol{W}_S, \quad \tilde{\boldsymbol{M}} = \boldsymbol{B}_{\text{music}}\boldsymbol{W}_M. \tag{10}$$

Where $\boldsymbol{W}_S$ and $\boldsymbol{W}_M$ are submatrices in matrix $\boldsymbol{W}$ that correspond to the speech and music components respectively in equation (9).

2.4. Spectral mask and speech signal reconstruction

As we can see from equations (9, 10) the two matrices $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{M}}$ may not sum up to the matrix $\boldsymbol{Y}$. We usually get nonzero decomposion error since NMF usually gives an approximation as follows:

$$\boldsymbol{Y} \approx \tilde{\boldsymbol{X}} + \tilde{\boldsymbol{M}}. \tag{11}$$

Assuming noise is negligible in the mixed signal, the estimated spectrograms of speech and music should sum up to the mixed signal spectrogram. To make the error zero, we use the initial estimated magnitude spectrograms $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{M}}$ to build a spectral mask [2] as follows:

$$\boldsymbol{H} = \frac{\tilde{\boldsymbol{X}}^p}{\tilde{\boldsymbol{X}}^p + \tilde{\boldsymbol{M}}^p}. \tag{12}$$

Where $p > 0$ is a parameter, $(.)^p$, and division are element-wise operations. Notice that elements of $\boldsymbol{H} \in [0, 1]$. These masks will scale every time-frequency bin in the observed mixed signal magnitude spectrogram with a ratio that explains how much each signal contributes in the mixed signal as follows:

$$\hat{\boldsymbol{X}} = \boldsymbol{H} \otimes \boldsymbol{Y}, \quad \hat{\boldsymbol{M}} = (\boldsymbol{1} - \boldsymbol{H}) \otimes \boldsymbol{Y}. \tag{13}$$

where $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{M}}$ are the final estimates of the magnitude spectrograms of the speech and music signals respectively, $\boldsymbol{1}$ is a matrix of ones, and $\otimes$ is element-wise multiplication. Spectral mask works as a soft mask for the observed mixed signal. Every entry of the separated speech signal spectrogram is a scaled version of its corresponding entry of the spectrogram of the mixed signal. The scale values are defined in the spectral mask matrix $\boldsymbol{H}$. Using different values for p leads to different kinds of masks. When $p = 2$ the mask $\boldsymbol{H}$ can be considered as a Wiener filter. At $p = \infty$, we achieve a binary mask (hard mask), which will choose the larger source component at each entry as the only component.

After finding the contribution of the speech signal in the mixed signal, the estimated speech signal $\hat{x}(t)$ can be found by using inverse STFT on the estimated magnitude spectrogram $\hat{\boldsymbol{X}}$ combined with the phase of the mixed signal.

After separating the speech signal from the music background, the audio-visual speech recognition system is used with the separated signal $\hat{x}(t)$ rather than dealing with the observed mixed signal $y(t)$. In the next sections, we show the main procedures for audio-visual speech recognition for the separated speech signal.

3. AUDIO-VISUAL SPEECH RECOGNITION SYSTEM

As mentioned in the introduction, the recognition system proposed in this work relies on performing speech recognition using both a speech signal separated from background music and its corresponding visual information. Because of the multi-channel nature of the system, separate feature extraction processes for each channel are performed for training and recognition. Extracted features for different channels are then handled in an MSHMM, where these multiple streams of observations are used in calculating the emission probabilities of the HMM model. In the MSHMM, given a multi-stream observation sequence $(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$, it is assumed that each observation is a concatenation of multiple vectors $\boldsymbol{o}_t^T = [\boldsymbol{o}_t^{1\,T}, \ldots, \boldsymbol{o}_t^{S\,T}]$, where S is the number of streams–which is two in our case. The emission probability for a state q_t is:

$$p(\boldsymbol{o}_t|q_t) = \prod_{i=1}^{S} p(\boldsymbol{o}_t^i|q_t)^{\alpha_i}, \tag{14}$$

where α_i are the stream weights. The streams are usually separately modeled with a Gaussian mixture model. In the following subsection, we give information about the features that we use in our recognition experiments.

3.1. Audio-Visual Features for the MSHMM

Audio features are extracted as Mel frequency cepstral coefficients (MFCC) [6] with 13 static features as well as Δ and $\Delta\Delta$ features, making a total of 39 features. For the separated speech, feature extraction is performed after the proposed methods extract speech from the mixed signal.

For the visual data, a square region of interest (ROI) is extracted and tracked between consecutive frames. The ROI is determined by using landmark points, which are extracted using Active Shape Models (ASM) [7, 8]. After all the landmark points are extracted on the face area, weight center of the lip is taken as ROI center. The size of the ROI is calculated by taking as one and a half times of the distance between the eye centers.

To extract the visual features that are used in the MSHMM, Principle Component Analysis (PCA) is applied on ROI frames. For PCA top 30 principle components for each frame are extracted. To represent a visual frame, first derivatives of principle components are also added resulting in a vector of 60 dimensions. So, an audio-visual frame is represented by two streams, having 39 and 60 features respectively.

3.2. Training the HMMs and MSHMMs

Initially we model the phones with three states, and train an audio-only HMM. Then, to obtain an audio-visual MSHMM from audio only HMM, we concatenate visual features to audio features and train the multi-stream model using single-pass retraining from the audio-only HMM with only one iteration which gives better results than jointly training all streams. Since in the training we use only clean audio which is much more reliable than the video data and instead of performing a combined training, this single-pass retraining approach relying mostly on audio data using visual data only to calculate emission probabilities of the visual stream gives better results.

3.3. Recognition with the MSHMM

After training and obtaining the MSHMM from clean audio and visual data, we test the recognition accuracies on different types of audio and visual data. Since the main objective of our work is investigating recognition on mixed and separated speech and these conditions have no effect on visual data, visual stream is always the same in recognition. On the contrary audio stream changes with respect to the speech information being used. To test the accuracies, we give different stream weights (α values) to each stream to perform audio only, visual only or audio visual recognition. For audio only recognition we give one to audio stream weight and zero to visual stream weight and vice versa for visual only recognition. For audio-visual speech recognition, we perform different weight combinations on validation/held-out data at each given signal to music ratio (SMR) and take the combination that gives the best results.

4. EXPERIMENTS AND DISCUSSION

We tested the proposed system with the M2VTS video database [9], which consists of videos of 37 different people recorded in five sessions and arranged in five tapes. On the videos, the speakers say ten French digits, which are modeled as ten words and 19 phonemes. We have used first four tapes as training data and the fifth tape (excluding one video due to occlusion on the chin) as testing data. The reason for our choice is that, first four tapes are recorded under similar conditions however, fifth tape

has some visual differences like glasses or hat that add extra challenge to the data. This type of testing with one tape is different from jack-knife testing in previous works [5] and may result in slight relative decrease in visual recognition accuracy. For music data, piano music from piano society web site [10] was downloaded. We used 38 pieces from different composers but from a single artist for training and left out one piece for the testing stage. The test data was formed by adding random portions of the test music file to the 36 speech utterance files at different speech to music ratio (SMR) values in dB. The audio power levels of each file were found using the "audio voltmeter" program from the G.191 ITU-T STL software suite [11].

For the speech music separation algorithm, we used the training speech signals from the first four tapes. The magnitude spectrograms for the training speech and music data were calculated by using the STFT, a Hamming window was used, and the FFT was taken at 512 points, the first 257 FFT points only were used since the remaining points are the conjugate of the first 257 points. The sampling rate is 16KHz. We trained different numbers of basis vectors N_s for the speech signal and N_m for music signal such that $N_s, N_m \in \{32, 64, 128\}$. In order to get better source separation results, we applied the proposed SCSS algorithm on male and female speakers separately by builiding different bases for them.

The parameters $\{N_s, N_m, p\}$ of the source separation and the audio-visual stream weights $\{\alpha_a, \alpha_v\}$ of the audio-visual speech recognition were searched for every SMR on validation data by trying out several values. We used the first tape from the same database as validation data. We recorded the values of the parameters that gave the best results for different experiments as shown in Table 1.

After finding the parameter values, we applied the proposed algorithm to the test set in tape 5 and the trained models that are trained on the first four tapes as shown before. Table 2 shows the results that correspond to the parameter values that are given in Table 1 for every SMR value. The table shows the performance of using only speech recognition system (*Audio* column) without using either visual information or source separation. It also shows the results of using only visual information (*Visual* column), using Audio-Visual automatic speech recognition without source separation (*Audio Visual* column), using automatic speech recognition after applying source separation without using any visual information (*SCSS Audio* column). In the last column of that table, we show the results of our proposed algorithm, which combines the single channel source separation with Audio-Visual automatic speech recognition (*SCSS A-V* column). The table shows that incorporating visual information only or SCSS only improves the performance of ASR. Incorporating both SCSS and visual information to ASR gives remarkable improvements in the accuracy of ASR.

We can see from Table 1 that using SCSS in audio-visual speech recognition makes the AVSR system rely more on the audio data even for low SMR.

Tablo 1: Best choice of the parameters for different methods and different SMR values.

SMR	Audio Visual		SCSS Audio			SCSS A-V $N_m = 32, p = 1$		
dB	α_a	α_v	N_s	N_m	p	N_s	α_a	α_v
-5	0.1	0.9	128	128	2	32	0.5	0.5
0	0.3	0.7	32	32	1	32	0.5	0.5
5	0.5	0.5	32	32	1	32	0.7	0.3
10	0.6	0.4	32	32	1	32	0.8	0.2
15	0.7	0.3	128	32	1	128	0.8	0.2
20	0.9	0.1	128	32	1	128	0.9	0.1
Clean	1.0	0.0	n/a	n/a	n/a	n/a	1.0	0.0

Tablo 2: Recognition accuracies % for different methods.

SMR dB	Audio	Visual	Audio Visual	SCSS Audio	SCSS A-V
-5	15.83%	43.89%	46.39%	45.83%	**66.94%**
0	25.83%	43.89%	57.78%	62.50%	**80.28%**
5	55.00%	43.89%	75.83%	84.72%	**90.83%**
10	81.94%	43.89%	87.78%	91.39%	**95.56%**
15	92.22%	43.89%	95.28%	97.22%	**98.06%**
20	97.78%	43.89%	98.06%	99.17%	**99.44%**
Clean	100%	43.89%	100%	100%	**100%**

5. CONCLUSION

In this paper, we introduced an algorithm for audio-visual speech recognition using source separation to get better recognition accuracy. We separated the speech signal from the mixed signal, then we applied the audio-visual speech recognition on the separated speech signal. In our future work, we plan to integrate the visual information to improve source separation as proposed in [12]. Furthermore, the MSHMM model used in this work handles stream transitions in a synchronous fashion, however there exists extended models [13] in the literature that can handle state level asynchrony and can improve recognition rate, implementation of which are left as future work.

6. REFERENCES

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[2] Emad M. Grais and Hakan Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *17th International Conference on Digital Signal Processing*, 2011.

[3] Mikkel N. Schmidt and Rasmus K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *InterSpeech*, 2006.

[4] L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, Jan 1986.

[5] S. Dupont and J. Luettin, "Audio-visual speech modelling for continuous speech recognition," *IEEE Transactions on Multimedia*, 2000.

[6] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[7] T.F. Cootes and C.J.Taylor, "Active shape models - smart snakes," in *In British Machine Vision Conference*. 1992, pp. 266–275, Springer-Verlag.

[8] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," *ECCV*, 2008.

[9] Stephane Pigeon and Luc Vandendorpe, "The m2vts multimodal face database (release 1.00).," in *AVBPA*. 1997, vol. 1206 of *Lecture Notes in Computer Science*, pp. 403–409, Springer.

[10] URL, "http://pianosociety.com," 2009.

[11] URL, "http://www.itu.int/rec/T-REC-G.191/en," 2009.

[12] Llagostera Casanovas A., Monaci G., Vandergheynst P., and Gribonval R., "Blind Audiovisual source separation based on sparse redundant representations," *Multimedia, IEEE Transactions*, vol. 12, no. 5, pp. 358–371, 2010.

[13] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1274–1288, 2002.