Audio-visual Segmentation and "The Cocktail Party Effect"

Trevor Darrell[1], John W. Fisher III, Paul Viola, MIT AI Lab
William Freeman , Mitsubishi Electric Research Lab

Abstract

Audio-based interfaces usually suffer when noise or other acoustic sources are present in the environment. For robust audio recognition, a single source must first be isolated. Existing solutions to this problem generally require special microphone configurations, and often assume prior knowledge of the spurious sources. We have developed new algorithms for segmenting streams of audio-visual information into their constituent sources by exploiting the mutual information present between audio and visual tracks. Automatic face recognition and image motion analysis methods are used to generate visual features for a particular user; empirically these features have high mutual information with audio recorded from that user. We show how utterances from several speakers recorded with a single microphone and video camera can be separated into constituent streams; we also show how the method can help reduce the effect of noise in automatic speech recognition.

Introduction

Interfaces to computer systems generally are tethered to users, e.g., via a keyboard and mouse in the case of personal computers, through a touchscreen when dealing with automatic tellers or kiosks, or with a headset microphone or telephone when using automatic speech recognition systems. In contrast, humans interact at a distance and are remarkably adept at understanding the utterance of remote speakers, even when other noise sources or speakers are present. The "cocktail party effect"—the ability to focus in on a meaningful sub-stream of audio-visual information—is an important and poorly understood aspect of perception [1].

In this paper we show how multi-modal segmentation can be used to solve a version of the cocktail party problem, separating the speech of multiple speakers recorded with a single microphone and video camera.

[1] MIT AI Lab Room NE43-829, 545 Technology Square, Cambridge MA 02139 USA. Phone:617 253 8966, Fax:617 253 5060, Email: trevor@ai.mit.edu

Our technique is based on an analysis of joint audio-visual statistics, and can identify portions of the audio signal that correspond to a particular region of the video signal, and vice-versa. Automatic face recognition is used to identify locations of speakers in the video, and mutual information analysis finds the portions of the audio signal that are likely to have come from that image region. We can thus attenuate the energy in the audio signal due to noise sources or other speakers, to aid automatic speech recognition and teleconferencing applications.

Source Separation

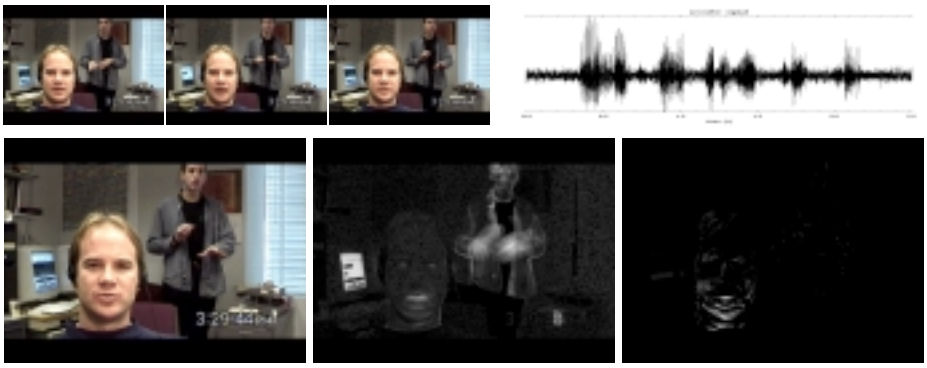
Most approaches to the problem of observing and listening to a speaker in a noisy and crowded room rely on active methods, either physically steering a narrow field microphone or adjusting the delay parameters of a beam-forming array [4,9,10].

These approaches are valuable, but require special sensors and sophisticated calibration techniques. We have developed passive methods that work on broadly tuned, monaural audio signals and which exploit time-synchronized video information. Our approach works using readily available PC teleconferencing camera and microphone components, as well as on video from broadcast and archival sources.

We base our method on the statistical analysis of signals with multiple independent sources. Prior statistical approaches to source separation often took audio-only approaches, and were successful when multiple microphones were available and the number of sources were known. The "blind source separation" problem has been studied extensively in the machine learning literature, and has been shown to be solvable using the Independent Components Analysis technique [2,11,13] and related methods. However these methods required knowledge of the number of sources and microphones, and could not solve the general source separation problem with a single microphone (as humans do).

Multi-modal Mutual Information

Mutual Information Analysis is a powerful technique that has been shown to be able to accurately register signals despite a wide range of visual sensor types (e.g., intensity, range) [14]. Here we apply the technique to multi-modal data types, including both audio and visual channels, and show how it can identify pixels in a video sequence which are moving in synchrony with an audio source.



(a) (b) (c)

Figure 1 Top: Video (left) and associated audio (right) signals. Bottom: Video frame (left), pixel-wise standard deviation image (middle), and mutual information image (right).

Classically, the mutual information between two random vectors can be written as $I(X,Y) = H(X) + H(Y) - H(X,Y)$ where $H(X)$ is the entropy of vector X and $H(X,Y)$ is the joint entropy of X and Y. In the case where X and Y are normally distributed with mean u_X, u_Y and covariance Σ_X, Σ_Y and jointly distributed with mean u_{XY} and covariance Σ_{XY}, then this is simply

$$I(X,Y) = \tfrac{1}{2}\log(2\pi e)^n|\Sigma_X| + \tfrac{1}{2}\log(2\pi e)^m|\Sigma_Y| - \tfrac{1}{2}\log(2\pi e)^{n+m}|\Sigma_{XY}| = \tfrac{1}{2}\log\frac{|\Sigma_X||\Sigma_Y|}{|\Sigma_{XY}|}$$

where m, n are the length of X, Y. This formalism has been applied with a scalar but time-varying X representing the audio source, and a multi-dimensional time-varying Y representing the video source [8].

The Gaussian assumption in unrealistic in many environments; a more general method is to use non-parametric density models. In order to make the problem tractable high dimensional audio and video measurements are projected to low dimensional subspaces. The parameters of the sub-space are learned by maximizing the mutual information between the derived features [6]

These methods have been used to identify which locations in a video signal correspond to a single audio source [6,8]. For example, Figure 1(top) shows example joint video and audio signals. Figure 2(bottom) shows a frame from the video, the pixels identified as high variance

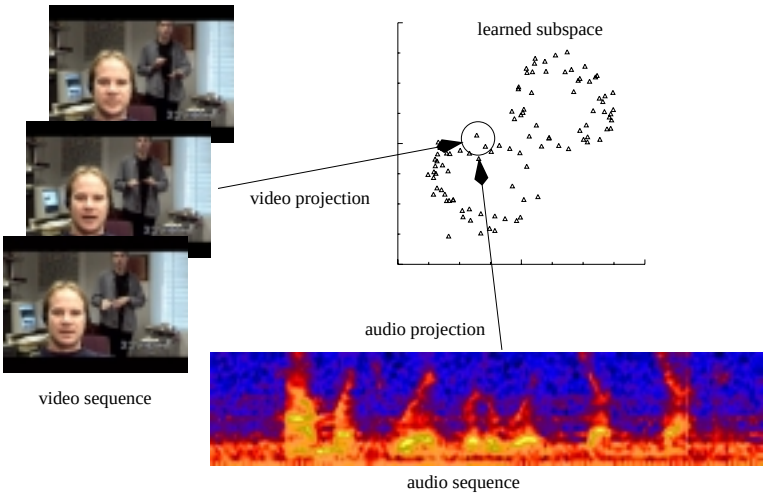


Figure 2: Projection of audio and video vectors.

from video information alone, and the pixels identified as having high mutual information with the audio source. One can see that analysis of image variance or motion alone fails to distinguish the pixels moving in synchrony with the audio source. Analyzing the joint mutual information over both video and audio signals approach easily identifies the pixels moving in the mouth region, corresponding to the speech sound source.

Spectral Separation

We are interested in the inverse problem: we wish to enhance or attenuate audio components given a corresponding video region. Unfortunately, an instantaneous scalar audio model makes it difficult to divide the audio signal into constituent sources. To overcome this problem, we extend the model to use a multi-dimensional time-frequency sound representation, which makes it easier to segment different audio sources.

It has long been known that multidimensional representations of acoustic signals can be useful for recognition and segmentation. Most typical is a representation which represents the signal in terms of frequency vs. time. Many acoustic events can be segregated in terms of pitch, such as classical musical instruments and human vowel sounds.

While the cepstrum, spectrogram, and correlogram are all possible representations, we use a periodogram-based representation.

In our implementation audio is sampled at 11.025 KHz, and then transformed into periodogram coefficients using hamming windows of 5.4ms duration sampled at 30 Hz (commensurate with the video rate). At each point in time there are 513 periodogram coefficients. We use a non-parametric density estimation algorithm, applied to multi-dimensional, time-varying audio and image features. Specifically, let $v_i \sim V \in \Re^{N_V}$ and $a_i \sim V \in \Re^{N_a}$ be video and audio measurements, respectively, taken at time i. Let $f_v : \Re^{N_v} \mapsto \Re^{M_v}$ and $f_a : \Re^{N_a} \mapsto \Re^{M_a}$ be mappings parameterized by the vectors α_v and α_a, respectively. In our experiments f_v and f_a are single-layer perceptrons and $M_a = M_v = 1$. However, the adaptation method extends to any differentiable mapping and output dimensionality [5]. During adaptation the parameter vectors α_a, α_v (the perceptron weights) are chosen such that

$$\{\alpha_a, \alpha_v\} = \arg\max_{\alpha_a, \alpha_v} I(f_v(V, \alpha_v), f_a(A, \alpha_a)) \ .$$

This process is illustrated in figure 2 in which video frames and sequences of periodogram coefficients are projected to scalar values. A clear advantage of learning a projection is that rather than requiring pixels of the video frames or spectral coefficients to be inspected *individually* the projection summarizes the entire set efficiently.

In [7] demonstrate how this framework can be utilized to segment the audio information based on raw pixel data in a user specified window of a video sequence. To create a fully automatic system for human-computer interface applications, we need pre-processing steps which transform the pixel data into a more suitable representation.

Automatic Face and Motion Detection

We use face detection and motion estimation as preprocessing steps to the audio-visual mutual information analysis. These provide an analysis of the video data to extract visual features that will correspond to the acoustic energy of an utterance in real-world conditions. A face detection module provides the location of pixels in a video stream which belong to an individual face. We restrict the adaptation algorithm to only consider these pixels, and thus only find components

(a) (b) (c)

Figure 3: (a) Example image with detected face regions, (b) image of α_v for left speaker, (c) α_v for right speaker.

of the audio stream which have high mutual information to that individual's face. Our implementation used the CMU facedetector library, which is based on a neural-network algorithm trained with both positive and negative examples of face images [12].

In addition to intensity information, image motion features are estimated and used as the video input term in the mutual information analysis. Computing mutual information with optic flow rather than raw pixel intensity change alone can help in cases where there is contrast normalized motion, such as with random dot patterns. We used the well-known robust optic flow implementation detailed in [3], which combines outlier rejection, multi-resolution processing, segmentation, and regularization. In practice we set the parameters of this algorithm to strongly regularize the motion estimates, since precise motion boundary localization was not important for our task.

Results

We first tested our system on an image sequence with two speakers recorded with a single microphone (the speakers were recorded with stereo microphones so as to obtain a reference, but the experiments used a single audio source). Figure 3(a) shows an example frame from the video sequence with detected faces.

By selecting data from one of the two detected face regions we can enhance the voice of the speaker on the left or right. As the original data was collected with stereo microphones we can compare our result to an approximation to an ideal Wiener filter (neglecting cross channel leakage). Since the speakers are male and female, the signals have better spectral separation and the Wiener filter can separate them. For

the male speaker the Wiener filter improves the SNR by 10.43 dB, while for the female speaker the improvement is 10.5 dB. Our technique achieves a 9.2 dB SNR gain for the male speaker, and a 5.6 dB SNR gain for the female speaker.

It is not clear why performance is not as good for the female speaker, but figures 3(b) and (c) are provided by way of partial explanation. Having recovered the audio in the user-assisted fashion described we used the recovered audio signal for video attribution (pixel-based) of the entire scene. Figures 3(b) and (c) are the images of the resulting α_v when using the male (b) and female (c) recovered voice signals. The attribution of the male speaker (b) appears to be clearer than that of (c); this may be an indication that the video cues were not as detectable for the female speaker as they were for the male in this experiment.

Our second test evaluated the ability of our method to improve accuracy in speech recognition where other noise sources were present. A single user spoke into a handheld video camera with built-in microphone, at a distance of approx. 4 feet. A second noise source was synthetically added to the audio stream at varying SNR levels (8-13db). Recognition was performed using a commercially available transcription package; the system was trained as specified without added noise.

Preliminary results from this test are encouraging; in recognition tests involving a combination of digits and spoken phrases, we obtained approximately a 33% reduction in error rate with our method. At higher noise levels (8db), the observed error rate was 55% unfiltered, and 38% with our technique applied as a preprocessing step. At lower noise levels (13db), the observed error rate was 33% unfiltered, and 22% after our technique was applied. Averaged over all SNR levels, the unfiltered error rate was 49% (106/216 words), while the filtered error rate was 30% (65/216). Note that the absolute error rates are unnaturally high, since the system was not trained with the added noise source. Nonetheless we believe the relative error rate improvement is a promising sign.

We hope that this technique can aid in the eventual goal of enabling audio interface without an attached microphone for casual users.

Bibliography

B. Arons. A Review of The Cocktail Party Effect. Journal of the American Voice I/O Society 12, 35-50, 1992.

A. Bell and T. Sejnowski. An information maximisation approach to blind separation and blind deconvolution, Neural Computation, 7, 1129-1159, 1995.

M. Black and P. Anandan, A framework for the robust estimation of optical flow, Fourth International Conf. on Computer Vision, ICCV-93, Berlin, Germany, May, 1993, pp. 231-236

M. A. Casey, W. G. Gardner, and S. Basu "Vision Steered Beam-forming and Transaural Rendering for the Artificial Life Interactive Video Environment (ALIVE)", Proceedings of the 99th Convention of the Aud. Eng. Soc. (AES), 1995.

J. Fisher and J. Principe. Unsupervised learning for nonlinear synthetic discriminant functions. In D. Casasent and T. Chao, eds., SPIE Optical Pattern Recognition VII, vol 2752, p 2-13, 1996.

J. W. Fisher III, A. T. Ihler, and P. Viola, "Learning Informative Statistics: A Nonparametric Approach," in Advances in Neural Information Processing Systems 12, Denver, Colorado, 1999.

J. W. Fisher III, T. Darrell, W. Freeman, and P. Viola, Learning Joint Statistical Models for Audio-Visual Fusion and Segregation, to appear in Advances in Neural Information Processing Systems 13.

J.~Hershey and J.~Movellan.Using audio-visual synchrony to locate sounds. in Advances in Neural Information Processing Systems 12, 1999.

Q. Lin, E. Jan, and J. Flanagan. "Microphone Arrays and Speaker Identification." IEEE Transactions on Speech and Audio Processing,Vol. 2, No. 4, pp. 622-629, 1994.

T. Nakamura, S. Shikano, and K. Nara. "An Effect of Adaptive Beamforming on Hands-free Speech Recognition Based on 3-D Viterbi Search", ICSLP'98 Proceedings Australian Speech Science and Technology Association, Incorporated (ASSTA), vol. 2, p. 381, 1998.

B. Pearlmutter and L. Parra. "A context-sensitive generalization of ICA", Proc. ICONIP '96, Japan, 1996.

H. Rowley, S. Baluja, and T. Kanade, Neural Network-Based Face Detection, Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR-96, pp. 203-207,. IEEE Computer Society Press. 1996.

P. Smaragdis, "Blind Separation of Convolved Mixtures in the Frequency Domain." International Workshop on Independence \& Artificial Neural Networks University of La Laguna, Tenerife, Spain, February 9 - 10, 1998.

P.~Viola, N.~Schraudolph, and T.~Sejnowski. Empirical entropy manipulation for real-world problems. in Advances in Neural Information Processing Systems 8, pages 851--7, 1996.