AUDIO-VISUAL GRAPHICAL MODELS FOR SPEECH PROCESSING

John Hershey,

University of California San Diego
Machine Perception Laboratory

Hagai Attias, Nebojsa Jojic,Trausti Kristjansson

Microsoft Research
Redmond Washington

ABSTRACT

Perceiving sounds in a noisy environment is a challenging problem. Visual lip-reading can provide relevant information but is also challenging because lips are moving and a tracker must deal with a variety of conditions. Typically audio-visual systems have been assembled from individually engineered modules. We propose to fuse audio and video in a probabilistic generative model that implements cross-model self-supervised learning, enabling adaptation to audio-visual data. The video model features a Gaussian mixture model embedded in a linear subspace of a sprite which translates in the video. The system can learn to detect and enhance speech in noise given only a short (30 second) sequence of audio-visual data. We show some results for speech detection and enhancement, and discuss extensions to the model that are under investigation.

1. INTRODUCTION

We often take for granted the ease with which we can carry on a conversation in the midst of noise. This auditory scene analysis problem confounds current automatic speech recognition systems, which can fail to recognize speech in the presence of very small amounts of interfering noise. It is well known that in humans, vision often plays a crucial role, because we often have an unobstructed view of the lips that modulate the sound. This fact has motivated efforts to use video information for tasks of audio-visual scene analysis, such as speech recognition and speaker detection [1].

Such systems have typically been built using separate modules for tasks such as tracking the lips, extracting features, and detecting speech components, where each module is independently designed to be invariant to different speaker characteristics, lighting conditions, and noise conditions. However a system that can adapt to one's face under the current lighting condition may perform better than one trained for a variety of conditions without adaptation.

We address the integration and the adaptation problems of audio-visual scene analysis by using a probabilistic generative model to combine video tracking, feature extraction, and tracking of the phonetic content of audio-visual speech. A generative model offers several advantages. It allows us to capture and exploit dependencies between modalities. It gives us principled methods of inference and learning across modalities that ensure the Bayes optimality of the system. It allows us to extend the model, for instance by adding temporal dynamics, in a principled way while maintaining optimality properties. It also allows us to use the same model for a variety of inference tasks, such as enhancing speech by reading lips, detecting whether a person is speaking, or predicting the lips using audio.

(a) Audio Model (b) Video Model

Fig. 1. Audio and Video Models

In previous work it has been shown that a generative model could capture dependencies between time delays of the speech signal in two microphone signals and motion in a camera of the image of the speaker [2]. In that system the cross-modal calibration parameters were automatically discovered during unsupervised learning, and the audio time delay signal was able to bootstrap learning of the visual tracking, yielding much better tracking when the multi-modal system was adapted jointly than when the models were adapted independently in each modality. Audio-visual speech recognition has been explored in a variety of papers [1]. Speaker localization has been handled in other systems such as [3]. Unsupervised learning of video tracking has been developed for example in [4]. Adaptation to noise conditions has been demonstrated in for example [5].

Here we develop a generative model that accomplishes aspects of all of these works. It fuses audio and video by learning the dependencies between the noisy speech signal from a single microphone and the fine-scale appearance and location of the lips during speech. One possible scenario for this model is that of a human computer interaction: a person's audio and visual speech is captured by a camera and microphone mounted on the computer, along with other interfering signals in the room: machine noise, another speaker, and so on.

In the rest of the paper we present the model structure along with the inference and learning rules, and describe some experiments using it to detect and enhance speech in the presence of noise, and while tracking the lips in video. Finally we suggest possible extensions to the model.

2. AUDIO MODEL

The generative model for audio shown in 1(a) is as follows. A windowed short segment or *frame* of the observed microphone signal is represented in the frequency domain as $w_k \in \mathbb{C}$ where k indexes the frequency band. This observed quantity is described as

the clean speech signal u_k amplified by scalar h and corrupted by Gaussian noise having *precision* (inverse variance) ϕ_k. The speech signal is in turn modeled as a zero mean Gaussian mixture model with state variable s and state-dependent precision σ_{sk}, which corresponds to the inverse power of the frequency band k for state s. Thus the audio model is

$$\begin{aligned} p(u \mid s) &= \prod_k \mathcal{N}(u_k \mid 0, \sigma_{sk}) \\ p(s) &= \pi_s \\ p(w \mid u) &= \prod_k \mathcal{N}(w_k \mid h u_k, \phi_k) \ . \end{aligned} \tag{1}$$

where for the complex sub-band components u_k a Gaussian distribution is defined as $\mathcal{N}(u \mid \rho, \sigma) = \frac{\sigma_k}{\pi} e^{-\frac{\sigma_k}{|}u_k - \rho_k|^2}$ with mean ρ_k and precision σ_k. This is a joint distribution over the real and imaginary parts of u_k, hence the power of two disparity from the usual Gaussian.

We model the audio using a zero mean Gaussian, rather than the traditional cepstral coefficients used in speech recognition. One advantage of this approach is that we can easily extend the model to use phase from inferred microphone delay as in [2]. To use cepstral coefficients derived from the log power spectrum and accommodate inferences about phase is a challenging problem. In addition, the inference of the clean speech in noise is greatly simplified, both mathematically and computationally. The use of nonlinear features such as cepstral components requires either iterative optimization procedures ([6]) or approximations ([7]) to perform noise compensation. Furthermore, whereas cepstral components may work well for speech recognition, high-resolution spectral components may work well for speech enhancement in noisy conditions, because it can take advantage of fine structure in either the signal or the interference [8].

3. VIDEO MODEL

The video model describes an observed frame of pixels from the camera, y as a noisy version of a hidden template v shifted in two dimensions by discrete location parameter l. v in turn is described as a weighted sum of linear basis functions, $A_j \in \mathbb{R}^{N \times 1}$ which make up the columns of A with weights given by hidden variables r. Such a model constitutes a factor analysis model that helps explain the covariance among the pixels in the template v within a linear subspace spanned by the columns of A. This arrangement uses far fewer parameters than the full covariance matrix of v while capturing the most important variances and provides a low-dimensional space of causes, r. In figure 2 r is projected into the subspace of v spanned by the columns of A. It is the further structure within this subspace that we hope to describe using audio.

The video model is parameterized as

$$\begin{aligned} p(l) &= constant \\ p(v \mid r) &= \prod_i \mathcal{N}(v_i \mid \sum_j A_{ij} r_j + \mu_i, \nu_i) \\ p(y \mid v, l) &= \prod_i \mathcal{N}(y_i \mid v_\xi(x_i - x_l), \lambda) \ . \end{aligned} \tag{2}$$

where ν_i is the conditional precision of each pixel, and μ_i captures part of the mean that doesn't depend on the factors. The mapping between two-dimensional coordinates and vector indices is handled by the expression $v_\xi(x_i - x_l)$ in which $x_i \in R^{2\times 1}$ is the position of the ith pixel, $x_l \in R^{2\times 1}$ is the position represented by discrete variable l, and $\xi(x)$ is the index of v corresponding to two-dimensional position x.

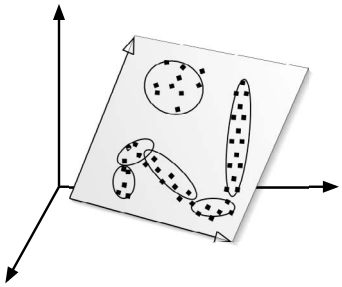

Fig. 2. Video Model as Embedded Subspace Model

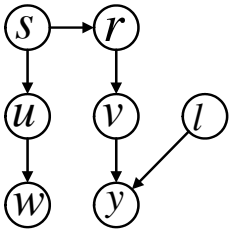

Fig. 3. Audio-Visual Model

4. AUDIO-VISUAL MODEL

Each model by itself is fairly simple, but by exploiting cross-modal fusion we can obtain a system that is more than just the sum of its parts. We fuse the two models together by allowing the mean and precisions of the hidden video factors r to depend on the states s as illustrated in Figure 3:

$$p(r \mid s) = \prod_j \mathcal{N}(r_j \mid \eta_{sj}, \psi_{sj}) \ . \tag{3}$$

The discrete variable s now controls the location and directions of covariance of a video representation that is embedded in a linear subspace of the pixels. Thus we can now represent a nonlinear manifold embedded in a linear subspace, as illustrated in Figure 2.

5. INFERENCE

A variational expectation maximization (EM) algorithm that decouples l from v can be derived to simplify the computation. The posterior $p(u, s, r, v \mid y, w)$ has the factorized form

$$p(u, s, r, v \mid y, w) = q(u \mid s) q(s) q(r \mid s) q(v \mid r, l) q(l) \ . \tag{4}$$

For u we get

$$\begin{aligned} q(u \mid s) &= \prod_k \mathcal{N}(u_k \mid \bar{\rho}_{sk}, \bar{\sigma}_{sk}) \\ \bar{\rho}_{sk} &= \frac{1}{\bar{\sigma}_{sk}} h \phi_k w_k \\ \bar{\sigma}_{sk} &= h^2 \phi_k + \sigma_{sk} \ . \end{aligned} \tag{5}$$

For v we get

$$\begin{aligned} q(v \mid r) &= \prod_i \mathcal{N}(v_i \mid \sum_j \bar{A}_{ij} r_j + \bar{\mu}_i, \bar{\nu}_i) \\ \bar{\nu}_i &= \lambda E_l \alpha_{i+l} + \nu_i \\ \bar{\mu}_i &= \frac{1}{\bar{\nu}_i}(\nu_i \mu_i + \lambda E_l \alpha_{i+l} y_{i+l}) \\ \bar{A}_{ij} &= \frac{\nu_i}{\bar{\nu}_i} A_{ij} \ . \end{aligned} \tag{6}$$

For r we get

$$\begin{aligned} q(r \mid s) &= \mathcal{N}(r \mid \bar{\eta}_s, \bar{\psi}_s) \\ \bar{\eta}_s &= \bar{\psi}_s^{-1} \left[\psi_s \eta_s + A^T \mathrm{diag}(\nu)(\bar{\mu} - \mu) \right] \\ \bar{\psi}_s &= A^T \mathrm{diag}(\nu - \frac{\nu^2}{\bar{\nu}}) A + \psi_s \end{aligned} \tag{7}$$

where $\mathrm{diag}(\nu)$ is a diagonal matrix with the elements of ν along the diagonal

For s we get

$$q(s) = \bar{\pi}_s = \frac{\bar{\pi}'_s}{\sum_s \bar{\pi}'_s} \tag{8}$$

where

$$\begin{aligned} \log \bar{\pi}'_s &= \log \pi_s \\ &+ \sum_k \left(\log \frac{\sigma_{sk}}{\bar{\sigma}_{sk}} - \phi_k \mid w_k - h\bar{\rho}_k \mid^2 - \sigma_{sk} \mid \bar{\rho}_{sk} \mid^2 \right) \\ &+ \log \mid \psi_s \bar{\psi}_s^{-1} \mid - \frac{1}{2} \sum_j \psi_{sj} (\bar{\eta}_{sj} - \eta_{sj})^2 \\ &- \frac{1}{2} \sum_i \nu_i \left[\sum_j (\bar{A}_{ij} - A_{ij}) \bar{\eta}_{sj} + \bar{\mu}_i - \mu_i \right]^2 \\ &- \frac{\lambda}{2} \sum_i \left[E_l \alpha_{i+l} (y_{i+l} - \sum_j \bar{A}_{ij} \bar{\eta}_{sj} - \bar{\mu}_i)^2 + (\bar{A} \bar{\psi}_s^{-1} \bar{A}^T)_{ii} \right] \end{aligned} \tag{9}$$

For l we get

$$\begin{aligned} q(l) &\propto e^{f(l)} p(l) \\ f(l) &= -\frac{\lambda}{2} \sum_i \alpha_{i+l} \left(y_{i+l} - \sum_{sj} \bar{A}_{ij} \bar{\pi}_s \bar{\eta}_{sj} - \bar{\mu}_i \right)^2 \end{aligned} \tag{10}$$

All of the expectations with respect to the hidden location random variable l can be shown to be equivalent to a convolution, and can be efficiently carried out using a fast Fourier transform. To enhance the audio we infer expected value of the audio using the posteriors of u and s calculated above: $\mathrm{E}(u|w,v) = \sum_s \bar{\pi}_s \bar{\rho}_s$. We then invert the Fourier transform and overlap and add using a lapping synthesis window matched to the analysis window.

6. LEARNING

In the M-step we compute the model parameters. The update rules use sufficient statistics which involve two types of averages. We denote by E average w.r.t. the posterior q at a given frame n, and we denote by $\langle \cdot \rangle$ average over frames n. The subscript n will be omitted.

For σ we get

$$\frac{1}{\sigma_{sk}} = \langle |\rho_{sk}^-|^2 + \frac{1}{\sigma_{sk}^-} \rangle \tag{11}$$

For h, ϕ we get

$$\begin{aligned} h &= \frac{\mathrm{Re} \sum_k \phi_k \langle w_k E u_k^\star \rangle}{\sum_k \phi_k \langle E \mid u_k \mid^2 \rangle} \\ \frac{1}{\phi_k} &= \langle \mid w_k \mid^2 \rangle - 2h \mathrm{Re} \langle w_k E u_k^\star \rangle + \langle E \mid u_k \mid^2 \rangle \end{aligned} \tag{12}$$

where

$$\begin{aligned} E u_k &= \sum_s \bar{\pi}_s \bar{\rho}_{sk} \\ E \mid u_k \mid^2 &= \sum_s \bar{\pi}_s \left(\mid \bar{\rho}_{sk} \mid^2 + \frac{1}{\bar{\sigma}_{sk}} \right) \end{aligned} \tag{13}$$

For A, μ, ν we get

$$\begin{aligned} A &= [\langle E v r^T \rangle - \langle E v \rangle \langle E r^T \rangle][\langle E r r^T \rangle - \langle E r \rangle \langle E r^T \rangle]^{-1} \\ \mu &= \langle E v - A E r \rangle \\ \nu^{-1} &= \mathrm{diag}^{-1} \langle E v v^T - A E r v^T - \mu E v^T \rangle \end{aligned} \tag{14}$$

where diag^{-1} in the last equation extracts the diagonal of the matrix as a vector. For the averages we have

$$\begin{aligned} E r &= \sum_s \bar{\pi}_s \bar{\eta}_s \\ E r r^T &= \sum_s \bar{\pi}_s \left(\bar{\eta}_s \bar{\eta}_s^T + \bar{\psi}_s^{-1} \right) \\ E v &= \sum_s \bar{\pi}_s \left(\bar{A} \bar{\eta}_s + \bar{\mu} \right) \\ E v r^T &= \sum_s \bar{\pi}_s \left[\left(\bar{A} \bar{\eta}_s + \bar{\mu} \right) \bar{\eta}_s^T + \bar{A} \bar{\psi}_s^{-1} \right] \end{aligned}$$

$$E v v^T = \sum_s \bar{\pi}_s \left[\left(\bar{A} \bar{\eta}_s + \bar{\mu} \right) \left(\bar{A} \bar{\eta}_s + \bar{\mu} \right)^T + \bar{A} \bar{\psi}_s^{-1} \bar{A}^T + \bar{\nu}^{-1} \right] \tag{15}$$

Finally, for η, ψ we get

$$\begin{aligned} \eta_{sj} &= \langle \bar{\eta}_{sj} \rangle \\ \frac{1}{\psi_{sj}} &= \langle (\bar{\eta}_{sj} - \eta_{sj})^2 + \left(\bar{\psi}_s^{-1} \right)_{jj} \rangle \end{aligned} \tag{16}$$

7. EXPERIMENTS

We conducted experiments to demonstrate the viability of the technique for the tasks of speech enhancement and speech detection. The data consisted of video from the Carnegie Mellon University Audio-Visual Speech Processing Database (by Fu Jie Huang http://amp.ece.cmu.edu/projects/AudioVisualSpeechProcessing).

We trained a speaker-dependent model having 32 states and 16 subspace basis functions on a 30-second sequence of the face of subject "Jon" cropped around the lip area, with accompanying

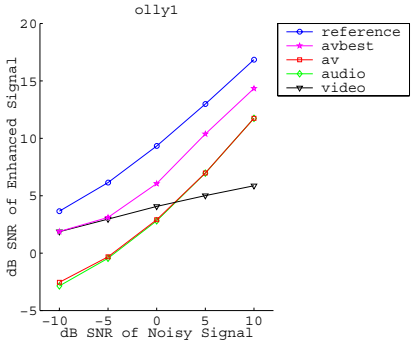

Fig. 4. Audio-Visual Enhancement Results: For inference, the *video* condition used video only, the *audio* condition used the noisy audio only, the *av* condition weighed the audio and video equally, the *avbest* condition selected the best weights in terms of SNR, and the *reference* condition used clean audio to infer the state.

clean audio speech, then trained the noise model on 10 seconds of an interfering audio signal which in this case happened to be another speaker. The model was then tested on a set of data not used during training, consisting of three different 30-second sequences of the same speaker, mixed with different segments of interfering audio signal.

In order to maximize performance it was necessary to vary the contribution of the audio and video components to the state posterior. At test time we vary the log likelihood of audio and video using a single parameter α to control the relative weights. This scheme ensures that when at one extreme we have a valid audio only model, at the other we have a valid video only model, and in between we have the unaltered audio-visual model. We tested inference under five different settings of $alpha$ as described in Figure 7.

Signal-to-noise ratio (SNR) was calculated for the enhanced audio signal relative to the clean signal in the time domain (i.e., $SNR = -10\log_{10}\frac{1}{n}\sum_n(x[n]-y[n])^2$ where x is the clean time domain signal, y is the estimated signal) . Results for each condition are shown in Figure 7.

One plausible explanation for strong video contribution at low SNRs is that with an interfering speaker it is difficult for the audio side of the model to detect when the target speaker is speaking, which is something that is may be easier to determine from video. To test the speech detection performance we turned the enhancement system into a speech detector by thresholding the power of the enhanced signal in each frame and comparing the resulting classification to that obtained by thresholding the clean signal in the same way. Speech detection performance was about 85% at zero dB SNR, with the best setting of α, (the setting used for enhancement in the previous experiment). Performance with the video-only model was comparable at approximately 83%.

In another experiment we used video from the same set, in which the lips are artificially translated in random directions. Tracking was able to almost completely compensate for lip motion, with enhancement to within one dB of that with untranslated video.

8. EXTENSIONS

The systematic nature of the graphical model framework allows us to integrate our generative audio-visual model with other submodules that we have investigated. In particular, the simplistic noise model we have used can be replaced with a mixture model The addition of another microphone as in [2] would further improve both noise robustness and tracking.

The model could also be extended with dynamics, making it a form of hidden Markov model. This would also open up the possibility of exploring time-asynchrony between audio and video streams which may help in interpreting anticipatory motion of the lips. We also intend to explore other applications of the current model, such as unsupervised speaker localization.

9. CONCLUSIONS

We have derived and implemented the inference and learning rules for a novel audio-visual model. The model is capable of tracking video as it translates and changes shape within a low-dimensional linear subspace of pixels. We have shown that the model can be applied to audio-visual speech enhancement, and that useful relationship between audio and video can be learned from small amounts of data. Thus it may be possible to adapt such a system to the prevailing noise and lighting as well as individual differences among speakers in a given situation. Although results are preliminary, we feel this is a promising step toward a completely unsupervised system that can usefully combine the two modalities in a principled way.

10. REFERENCES

[1] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A.Mashari, and J. Zhou, "Audio-visual speech recognition, final workshop 2000 report," Tech. Rep., The Johns Hopkins University, Baltimore, MD, October 2000.

[2] M.J. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. ECCV*, 2002.

[3] J. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds," in *In Advances in Neural Information Processing Systems 12. S. A. Solla, T. K. Leen and K. R. Muller (eds.) 813-819. MIT Press.*, 2000.

[4] Brendan Frey and Nebojsa Jojic, "Learning mixture models of images and inferring spatial transformations using the em algorithm," in *Computer Vision and Pattern Recognition (CVPR)*, 1999.

[5] H. Attias, A. Acero, J.C. Platt, and L. Deng, "Speech denoising and dereverberation using probabilistic models.," 2002.

[6] B.J. Frey, T. Kristjansson, L. Deng, and A. Acero, "Learning dynamic noise models from noisy speech for robust speech recognition," *Advances in Neural Information Processing (NIPS)*, 2001.

[7] M.J.F. Gales and S. Young, "An improved approach to the hidden markov model decomposition of speech and noise," *In Proc. of ICASSP*, pp. 233– 236, 1992.

[8] Trausti Kristiansson and John Hershey, "High resolution signal reconstruction," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2003, in press.