IMPROVING ACOUSTIC SPEAKER VERIFICATION WITH VISUAL BODY-LANGUAGE FEATURES

Christoph Bregler, George Williams, Sally Rosenthal, Ian McDowall
Courant Institute of Mathematical Sciences, New York University
{chris,george,sally,ian}@movement.nyu.edu

Abstract

We show how an SVM based acoustic speaker verification system can be significantly improved in incorporating new visual features that capture the speaker's "Body Language." We apply this system to many hours of Internet videos and TV broadcasts of politicians and other public figures. Our data ranges from current and former US election candidates to the Queen of England, the President of France, and the Pope, while giving speeches.

Index Terms— Speaker recognition, Machine vision, Motion analysis, Multimedia systems

1. Introduction

Among the reasons for recent advances in speaker recognition are discriminative classification based on SVMs [10] and novel feature extraction methods, based on both short-term spectral features [5,18] as well as prosodic and other linguistic information [16]. Another important speech modality that so far has only been studied in the context of lip-reading is the visual signal [1,6]. Besides lip motions, the rest of the body—the eyes, head, arms, torso, and their various movements--sends important signals. In this project, we study how to process these additional signals, the sum of which we call the "body signature." We hypothesize that every person has a unique body signature, which we are able to detect and use in a speaker verification setting. We present a new video-based feature extraction technique and several experiments with an SVM based speaker verification technique [5]. Compared to acoustic speech, the body signature is much more ambiguous. Despite this more challenging task, we show 20% Equal Error Rate (EER) using visual features only, and up to 4.1% EER using both acoustic and visual features. In all experiments we showed an improvement over pure acoustic verification performance.

In Section 2 we outline our new visual feature extraction technique that converts the video signal into a sequence of vectors similar to acoustic feature front-ends. Section 3 describes how we integrate these new modalities into an acoustic based speaker verification system, and section 4 details our experiments with the audio-visual database of political speeches.

2. Robust Visual Motion Features

Tracking visual features on people in videos is very difficult. It is easy to find and track the face because it has clearly defined features, but hands and clothes in standard videos are very noisy. Self-occlusion, drastic appearance change, low resolution (i.e. the hand is sometimes just a few pixels in size), and background clutter make the task of tracking very challenging. The most impressive people tracking recently has been demonstrated by [13]. It recognizes body parts in each frame by probabilistic fitting kinematic color and shape models to the entire body. Many other related techniques have been proposed, but an extensive literature review is beyond the scope of this paper. Please see the survey article by [9]. Explicitly tracking body parts would be of great advantage for our problem, but given the low-resolution web footage, it might be impossible to explicitly track the hands this way. Our technique builds on the observation that it is easy to track just a few reliable features for a few frames (instead of tracking body parts over the entire video). Given those short-term features at arbitrary "un-known" locations, we apply an implicit feature representation that is inspired by techniques that compute global orientation statistics of local features. Examples include [8,3,12,19,7].

We are interested in a robust feature detector that does not use explicit tracking or body part localization (because these techniques will fail frequently, especially on low-res TV and web footage). We are interested in a feature extraction process that is always able to report a feature vector, no matter how complex the input video is.

2.1 MOS: Motion Orientation Signatures

The first step in our extraction schema is the visual 2D flow computation at such reliable feature locations. We detect reliable features with the "Good Features" technique by [15] and then compute the flow vector

with a standard pyramidal Lucas & Kanade estimation [4]. Given these subpixel resolution flow estimates, we compute a weighted angle histogram: The 2D flow directions are discretized into N angle bins (we had good experience with N=18). Each angle bin then contains the sum of the flow magnitudes in this direction. i.e., large motions have a larger impact than small motions. We clip flow magnitudes larger than a certain maximum value before adding it to the angle bin. This makes the angle histogram more robust to outliers. We then normalize all bin values in dividing them by the number of total features. This factors out fluctuations caused by a different number of features found in different video frames. The bin values are then blurred across angle bins and across time with a Gaussian kernel (sigma=1 for angles, and sigma=2 for time). This avoids aliasing effects in the angle discretization and across time. (Many web-videos only have 15 fps; some videos are with 24 fps and up-sampled to 30 fps.) After the spatio-temporal blurring, we further normalize the histogram values to 0-1 over a temporal window (currently t=10). This factors out video resolution, camera zoom and body size (double resolution creates double flow magnitudes), but could also factor out important features. Some people's motion signature is based on subtle motions, while others' large movements are much more part of their signature. For this reason, we keep the normalization constant as one extra feature. This is related to the energy value in established acoustic front-ends.

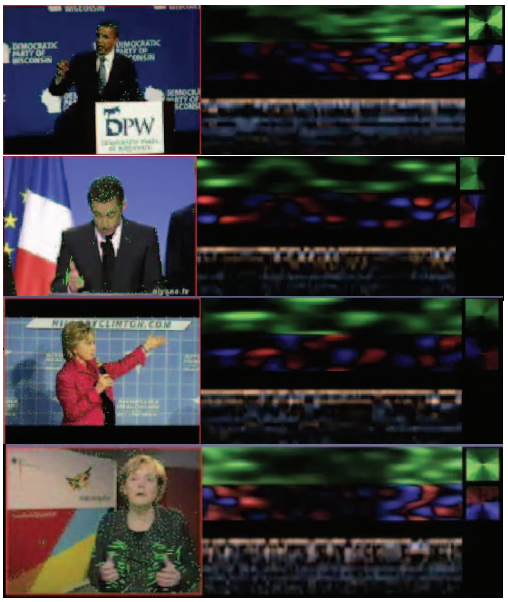

Figure 1: Several politicians doing different hand waving motions. The top rows (green) show the angle bin values over time. The middle rows (red is positive, blue is negative) show the delta-features over time. The bottom rows show the acoustic features.

As with acoustic speech features, we also compute "delta-features," the temporal derivative of each orientation bin value. Since the bin values are statistics of the visual velocity (flow), the delta-features cover acceleration and deceleration. For example if a subject claps her hands very fast, it produces large values in the bin values that cover 90° and 270° (left and right motion), but also large values in the corresponding delta-features. If a person just circles her hand with constant velocity, the bin values have large values across all angles, but the delta-features have low values. Figure 1 shows some example motion orientation signatures.

One very important aspect of this feature representation is that it is invariant to the location of the person. Given that the flow vectors are computed only at reliable locations, and we clip large flow vectors, the histograms are also very robust to noise.

Example MOS feature videos can be found at: http://movement.nyu.edu/ICASSP09

2.2 Shot Detector

If the footage is coming from TV or the Web it might be edited footage with scene cuts. Our recognition system should only work on one shot (scene) at a time, not the entire video. At shot boundaries, we see in the motion histograms drastic changes and could use that for segmenting scenes. Instead, we had better experience in additionally computing histograms over the color-values in each frame. If the difference between color-histograms is above a threshold (we use the histogram intersection metric), we then split the video [20]. With shots that are longer then 5 minutes (i.e. a speech), our shot-detector cuts the video into 5 minute shots. Sometimes we get very short shots of just a few seconds. Every shot that is below 5 seconds will be discarded. Shot-detection is an active research field, and we expect to incorporate a more advanced shot-detector in the future.

2.3 Limitations

As it happens with multiple acoustic speakers in one audio channel, when additional speakers are seen in the video, additional features will end up in the MOS signature. That is, if people are in the background clapping their hands while a political candidate gives a speech, the motion of the hand clapping produces "visual noise" similar to additive acoustic noise. In a future version we plan to incorporate a face-detector to constrain the visual features to body locations only (in order to avoid "visual noise").

3. GMM-Super-Features and SVM models

There are many possible architectures that produce state-of-the-art results for the task of speaker verification. We chose a technique proposed by [5] that converts an arbitrary long speech segment into a fixed-length feature vector and applies a SVM to

perform a classification. In a first step, a Gaussian Mixture Model (GMM) is used to estimate a "Universal Background Model" (UBM-GMM) from a large un-labeled speech corpus. We estimated such a UBM from the acoustic data and the visual MOS features. The video data included 1556 shots of automatically (randomly) downloaded YouTube videos containing unlabelled campaign speeches, music videos, TV commercials, audience reaction shots, and many other examples. We used half of these shots for training the UBM, and reserved the other half for testing. Although acoustic UBMs usually have GMMs with 64 or more mixtures, we achieved best results with 32 mixtures. Given such a UBM, so-called "Super-Features" can be calculated. A MAP adaption of the UBM [14] is performed for each acoustic speech segment or each video shot separately. The difference between the MAP adapted means and the UBM means is the so called GMM-Super-Feature vector.

3.1 Audio-Visual Integration

One challenge is how to integrate both modalities (as in related audio-visual lip-reading tasks [1,2,6]). This can be done at different abstraction levels. With our architectural choice, there are at least 2 different possible integration levels: 1) at the feature level, i.e. the GMMs are computed over the concatenated acoustic and visual vectors, 2) after the super-feature calculation, before they are fed into the SVM (i.e. the GMM-UBM clustering and the MAP adaption is done separately). We achieved superior results with the second integration method. We can imagine that with a significant larger database we might be able to afford more mixture models without over-fitting, and the first integration option might become superior.

Figure 2 shows a diagram of our system architecture.

For the acoustic front-end we used standard Mel Frequency Cepstral Coefficient (MFCC) features (12 Cepstral values, 1 energy value, and delta values).

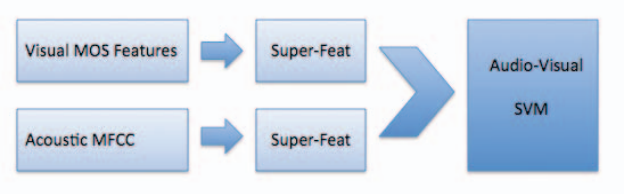

Figure 2: The audio-visual integration.

4. Experiments

Using the second half of the 1556 shots of random YouTube videos and 208 shots of 9 famous public figures (approx 4h data, see Figure 3) that were labeled by categories, we trained several SVM architectures. We ran 90 trials of different split-up between training set and test set for 7 different scenarios: 1) Clean acoustic speech, 2) Acoustic speech with 17dB of background noise (recorded in a pub including other chatter and noises), 3) Acoustic speech with 9.5dB of background noise, 4) Visual data only, 5-7) the 3 different noise-degraded acoustic speech data sets combined with visual speech. In all cases we could reduce the acoustic-only error rate in incorporating visual information: In perfect settings with clean acoustic data (Figure 4), the equal error rate of 4.7% EER (audio only) is reduced to 4.1% EER using audio-visual input (visual only is 20% EER). We got a very dramatic improvement on the 17dB SNR acoustic data in Figure 5, from acoustic EER of 9.4% error down to audio-visual EER of 4.9%, which cuts the error by almost half. In the 9.5dB SNR (heavier acoustic noise) environment in Figure 6, the EER goes from 21.8% (audio only) to 15% (audio visual).

Figure 3: Example Video Clips

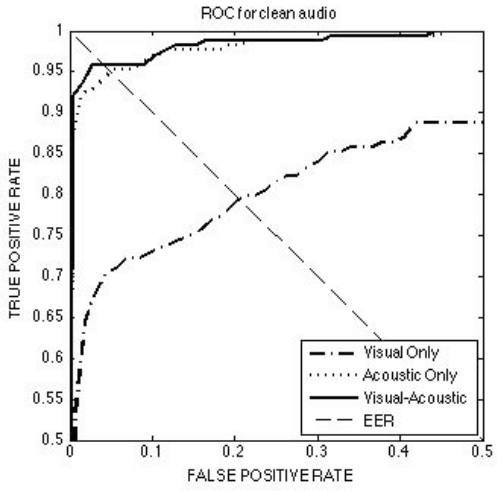

Figure 4: Clean audio: Visual EER: 20%, Acoustic EER: 4.7%, AV EER: 4.1%

5. Discussion and Future Plans

We have shown most significant improvement of the performance of a Speaker Verification System if the acoustic data is noise degraded, and we add visual

information from the speaker. But surprisingly, even on clean acoustic data where current acoustic systems perform with very low error, we could show additional reduction of the error rate with additional visual information. We plan to further improve the visual feature representation in adding other visual descriptors. We also plan to investigate different recognition architectures, including convolutional network architectures (TDNNs), and graph-based architectures (HMM, Bayes-nets), and also plan to apply these new visual features to other tasks, including body-language clustering. We also plan to incorporate into our models recent advances in modeling the variability of features across different samples for the same speaker, like nuisance attribute projection [17] and factor analysis [11]; such techniques should benefit both the acoustic and the visual features.

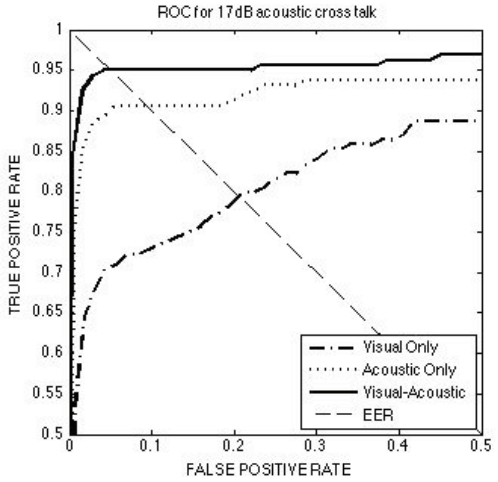

Figure 5: 17dB noise: Visual EER: 20%, Acoustic EER: 9.4%, AV EER: 4.9%

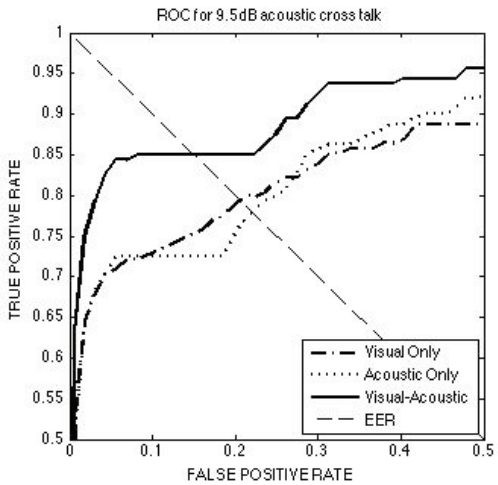

Figure 6: 9.5dB noise: Visual EER: 20%, Acoustic EER: 21.8%, AV EER: 15%

6. Acknowledgements

We would like to thank Peggy Hackney for helpful discussions, Andreas Stolcke for discussions on speaker recognition approaches, and the Office of Naval Research (ONR N000140710414) and National Science Foundation (NSF 0329098, NSF 0325715) for supporting this research.

7. References

[1] C. Bregler, Y. Konig, “Eigenlips” for Robust Speech Recognition, IEEE ICASSP 1994.

[2] C. Bregler, H. Hild, S. Manke, A. Waibel, Improving Connected Letter Recognition by Lipreading, IEEE ICASSP 1993.

[3] C. Bregler and J. Malik, Learning Appearance Based Models: Mixtures of Second Moment Experts, NIPS, 1997.

[4] J.Y. Bouget, Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the Algorithm. Intel Corp, 1999.

[5] W.M. Campbell, D.E. Sturim, D.A. Reynolds, Support Vector Machines Using GMM Supervectors for Speaker Verification, IEEE Signal Processing Letters, Vol. 13, No. 5, May 2006.

[6] C.C Chibelushi, F Deravi, J.S.D. Mason, A Review of speech-based bimodal recognition. IEEE Trans. On Multimedia, Vol 4, 2002.

[7] N. Dalal, B. Triggs, C. Schmid, Human detection using oriented histograms of flow and appearance, European Conference on Computer Vision, 2006.

[8] W.T. Freeman and M. Roth, Orientation histograms for hand gesture recognition. International Workshop on Automatic Face and Gesture Recognition, 1995.

[9] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, D. Ramanan, Foundations and Trends in Computer Graphics and Vision Volume 1 Issue 2/3 (255pp), 2006.

[10] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf and C. Burges and A. Smola (ed.) MIT-Press, 1999.

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor Analysis Simplified", Proc. ICASSP, vol 1, pp. 637-640, 2005.

[12] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, Int. Journal of Computer Vision, Vol 60, Number 2, pages 91-110, 2004.

[13] D. Ramanan, D.A. Forsyth, A. Zisserman, Strike a Pose: Tracking People by Finding Stylized Poses, Proc. CVPR 2005.

[14] D. A. Reynolds, T.F. Quatieri, R.B. Dunn, Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10, 19-41, 2000.

[15] J. Shi, C. Tomasi, Good features to track, CVPR 1994.

[16] E. Shriberg, "Higher-Level Features in Speaker Recognition", in Speaker Classification I, C. Mueller (ed.), pp. 241-259, Springer, 2007.

[17] A. Solomonoff, C. Quillen, and I. Boardman, "Channel Compensation for SVM Speaker Recognition", Proc. Odyssey Speaker Recognition Workshop, pp. 57-62, Toledo, Spain, 2004.

[18] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, A. Venkataraman, MLLR Transforms as Features in Speaker Recognition. Ninth European Conference on Speech Communication and Technology, 2005.

[19] L. Zelnik-Manor and M. Irani, Statistical Analysis of Dynamic Actions, IEEE Trans. On Pattern Analysis and Machine Intelligence, p 1530-1535, 2006.

[20] H. Zhang, A. Kankanhalli, S.W. Smoliar, Automatic partitioning of full-motion video. Readings in Multimedia Co