eNTERFACE'10 Project Proposal

Title: Multimodal Speaker Verification in NonStationary Noise Environments

Principal Investigators: Cenk Demiroglu - Ozyegin University, Turkey
Devrim Unay - Sabanci University, Turkey

Date: 10.12.2009

Project Objective: The aim of this research is to provide a robust multi-modal speaker verification system in non-stationary noise environments. The idea has been investigated before both for speech recognition and speaker verification problems. One commonly used and effective method is to measure the uncertainty in speech and video features and take the uncertainty into account in the decision process. Although uncertainty decoding is shown to improve the performance, estimating the uncertainty in speech features is a difficult problem especially in nonstationary noise environments. Here, we will investigate the possibility of estimating that uncertainty using both video and speech features to develop a noise-robust system.

Background Information: This project aims in addressing the problem of multimodal speaker verification in the presence of nonstationary noise. Speaker verification is an active research area, where main attempts have focused on using only audio information. There are two prevailing approaches to audio-only speaker verification: the Gaussian Mixture Model (GMM)-based approach [1] and the Support Vector Machine (SVM)-based approach [2]. Both approaches can yield competitive results. However, they are both sensitive to variability between different sessions and environmental noise during sessions. Although stationary noise can be effectively dealt with using existing techniques developed for speech recognition systems [3][4], nonstationary noise is still a big challenge on the way to successful use of speaker verification systems in real-life applications.

Speaker verification problem can be regarded as a biometric identity verification problem as well. Among various biometry-based solutions introduced in the literature, face recognition offers a good compromise between reliability and social acceptance, balancing security and privacy [5][6]. Despite considerable research effort, current solutions are still not able to fully solve the problem of face recognition especially in the presence of illumination and pose variations [7].

Due to the fact that performance of audio-only speaker verification systems largely vary in the presence of noise and visual-only systems are susceptible to some variations, recent research has shifted to combining information from these two modalities to build a more robust solution [8]. While initial works on speaker recognition combined information from audio and visual speech (video sequences of mouth or face region) [9], recent attempts have focused on reinforcing these two with face information and achieving final decision by dynamically weighting each source with respect to its reliability [10][11].

Detailed Technical Description:

I. **Technical Description:** The project will focus on the research and development of the following modules/subsystems:

- **Audio Processing:** The first step in the project is the implementation of a baseline speaker verification system. The second step is the modification of the baseline system to handle the uncertainty in the speech features in the speaker verification phase. Typically, uncertainty in speech and video features are computed independently in uncertainty based approaches to noise-robust audiovisual speech processing systems. The third step will be the implementation of an uncertainty detection system that uses video in addition to speech cues to estimate the uncertainty in speech features. The method that will be used in reliability detection is the lip distance information that will be obtained from the video processing block. Although closed lips or short lip distance do not always imply silence, they do indicate the presence of low speech energy which will be used in the uncertainty detection module to faster adapt to rapidly changing noise spectrum. Hence, we expect the final system to be more robust to nonstationary noise conditions compared to speech-only uncertainty detection. In the fourth workpackage, a new component will be developed to output the reliability of the decision which will be computed from the overall reliability of the features.
- **Visual Processing:** Visual Processing of the project will include the following workpackages: 1) pre-processing, such as denoising, illumination compensation, contrast enhancement; 2) detection (and tracking, for video input) of face regions; 3) facial features extraction, such as distance between eye-centers, lip shape from still images or video; and 4) face recognition. Considering the limited time available, we will focus on the state-of-the-art face recognition algorithms that are partially or fully available as open-source implementations. Our visual processing stage will take a video sequence or a single image as input, and output a matching score and a reliability measure about the user being the person he/she claims to be. Matching score will be computed using the feature vectors of the input image/video and those coming from the indexed database. Reliability will be based on some image quality measures defined in MPEG standards (e.g. brightness, contrast), as well as those dependent on the face detector (e.g. spatial resolution and in-plane rotation of face). These measures will be incorporated by the final decision stage.
- **Audio-Visual Information Fusion**: At the final decision stage, we aim to fuse the multimodal information coming from the visual and audio stages by hard-level max and soft-level weighted summation rules, as well as some popular classification schemes (e.g. support vector machines).

II. **Resources needed:**

All team members will need to have access to MATLAB software. The video processing team will also require access to Microsoft Visual Studio. The video processing and information fusion teams will need to have laptops with a camera and microphone. Everybody involved in the project will need to have access to M2VTS audio-visual database.

III. Project Management:

Dr. Demiroglu will be responsible from the development of the audio-related workpackages. Dr. Unay will be responsible from the development of the video-related workpackages. Information fusion team will work under the guidance of both Dr. Demiroglu and Dr. Unay.

Workplan and Implementation Schedule:

WP1 (First week): Specification – Definition of the Setup – Assignment of responsibilities – Data collection

WP2 (Second week): Implementation of the baseline GMM-based voice verification system (both training and testing algorithms). Implementation of the visual speech and face recognition system. Implementation of the fusion systems that will combine classifier outputs using synthetic data.

WP3 (Third week): Testing the GMM-based voice verification system and implementation of the uncertainty decoder to adapt to the noise environment. The uncertainty decoder will also be tested in this phase. Testing the visual speech and face recognition system under varying conditions, such as illumination/pose variations and (signal dependent) nonstationary noise. Implementation of the fusion systems using GMM-based voice verification system output and the outputs of visual speech and face recognition based verification parts.

WP4 (Fourth week): Fusion of the audio-based and video-based modules, and evaluation of the final system.

Profile of the team:

I. Principal Investigator: Dr. Cenk Demiroglu – Ozyegin University, Turkey

Cenk Demiroglu is an assistant professor in the Electrical and Computer Engineering department at Ozyegin University in Turkey. His Research interests include speaker verification, speech recognition, statistical text-to-speech systems, and the machine learning theory. He got his PhD from Georgia Institute of Technology in 2005. He worked as a team leader for three years in a speech recognition engine development project at CustomspeechUSA Inc., USA. After that, he worked as a team leader in the embedded statistical speech synthesis system development project for two years at Sensory Inc . He joined Ozyegin University in May 2009.

II. Principal Investigator: Dr. Devrim Unay – Sabanci University, Turkey

Devrim Unay is a visiting faculty member in the Faculty of Engineering and Natural Sciences at Sabanci University, Turkey. His research interests include medical image analysis, content-based information retrieval, feature extraction and selection, and classification as well as machine vision and quality inspection. Unay has a PhD in Applied Sciences from Faculté Polytechnique de Mons, Belgium, and an MS in Biomedical Engineering and a BS in Electrical

and Electronics Engineering from Bogazici University, Turkey. Previously he worked as a senior scientist and a Marie Curie fellow in the Video Processing and Analysis Group at Philips Research Eindhoven. Recently, he has received a 2-year EU FP7 Marie Curie Reintegration Grant on the topic of medical image analysis and retrieval. Unay has written more than 30 papers and has 2 patents pending.

III. Other researchers needed

2 researchers working on vision-based speaker verification
2 researchers working on audio-based speaker verification
2 researchers working on audio-visual information fusion

References:

[1] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19-41, 2000.

[2] Dehak, N., Dehak, R., Kenny, P. and Dumouchel, P. "Comparison Between Factor Analysis and GMM Support Vector Machines for Speaker Verification", In *Proceedings of IEEE Odyssey 2008 - The Speaker and Language Recognition Workshop (Odyssey 2008)*, Stellenbosch, South Africa, January 21-25, 2008.

[3] J. Droppo, A. Acero, and L. Deng, “Uncertainty decoding with SPLICE for noise robust speech recognition,” in *Proc. ICASSP*, Orlando, Florida, May 2002.

[4] H. Liao and M. J. F. Gales, “Joint uncertainty decoding for noise robust speech recognition,” in *Proc. Interspeech*, 2005.

[5] Tan, X.; Chen, S.; Zhou, Z.; and Zhang, F., “Face recognition from a single image per person: A survey,” *Pattern Recogn.,* Vol.39, no.9, pp.1725-1745, Sep. 2006

[6] Abate, A. F.; Nappi, M.; Riccio, D.; and Sabatino, G., “2D and 3D face recognition: A survey,” *Pattern Recogn. Lett.,* Vol.28, no.14, pp.1885-1906, Oct. 2007

[7] Zhang, X.; and Gao, Y. “Face recognition across pose: A review,” *Pattern Recogn.,* Vol.42, no.11, pp.2876-2896, Nov. 2009

[8] Aleksic, P.S.; Katsaggelos, A.K., "Audio-Visual Biometrics," *Proceedings of the IEEE* , vol.94, no.11, pp.2025-2044, Nov. 2006

[9] Wark, T.; and Sridharan, S.,”Adaptive Fusion of Speech and Lip Information for Robust Speaker Identification,” *Digital Signal Processing*, Vol.11, no.3, pp.169-186, July 2001

[10] Fox, N.A.; Gross, R.; Cohn, J.F.; Reilly, R.B., "Robust Biometric Person Identification Using Automatic Classifier Fusion of Speech, Mouth, and Face Experts," *Multimedia, IEEE Transactions on* , vol.9, no.4, pp.701-714, June 2007

[11] P. George, K. Athanassios, P. Vassilis, P. Maragos, “Adaptive Multimodal Fusion by Uncertainty Compensation with Application to Audiovisual Speech Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, No. 3, Mar 2009.