

# **Audio-Visual Synchrony for Face Location and Segmentation**

**Eric J. Mislivec**

**Faculty Sponsor: Christopher G. Prince**

## **Background**

This project would create a computational system to detect synchrony between portions of audio and visual data, in order to locate and segment faces on the basis of speech-face synchrony. There is a close audio-visual synchrony, or time-based correlation, between speech sounds and facial movements. I propose to use this relationship to locate and segment faces from digital video files. The implemented system would be used in our sensory-augmented word-learning model (Prince, Mislivec & Helder, submitted; Prince, Mislivec, Kosolapov & Lykken, 2002). This sensory-augmented word-learning model (SAWL) augments an artificial neural network (ANN) model of word-learning processes in infants (Hollich, 1999) with a sensory processing layer to explore the significance of sensory embodiment.

Key to SAWL is the ability to provide an adequate sensory representation to the word-learning ANN model. From the onset of their development, infants are presented with vast quantities of sensory information. Senses such as vision and audition continually receive separate input from the external world. Immediate correlations across sensory modalities are important in attentional selection and developing a psychological representation of the world. It has been reported that 7-month-old infants will learn the link between a vowel sound and an object when the movement of the object is synchronized with the production of the vowel sound, but not when the object moves out of synchrony or does not move at all (Gogate & Bahrick, 1998).

Detecting audio-visual synchrony in our model will allow further investigation of the importance of this factor in computational models of learning and development.

An audio-visual synchrony based approach also has certain practical benefits over other methods of sound localization and facial detection. Binaural methods (i.e., two sound source methods), including cross-correlation between channels in a stereo audio signal, can be used to provide an estimate of a sound's location (Nandy & Ben-Arie, 1996), but there are important drawbacks to these methods. For example, these methods have limited precision in spatial location and require stereo signals.

Facial detection methods tend to depend on implicit knowledge of faces (e.g., supervised training methods: Li, Zhu, Zhang & Zhang, 2002; Viola & Jones, 2001), or explicit knowledge of faces (e.g., our current visual processing system implements a template-based face recognition method: Sinha, 1995). However, the issue of facial *detection* is relatively independent of the method at hand. Detecting audio-visual synchrony will allow for speech-face relations to be taken into consideration, without programming our system with knowledge of a 'face.'

## **Research and Implementation**

The algorithm we will use for synchrony detection is to be adapted from an approach implemented by Hershey and Movellan (2000), based on a measure of the degree of mutual information between audio signals and regions within video signals. The centroid of the mutual information data was successfully used by these researchers to estimate a speaker's location. In addition to implementation of the synchrony detection mechanism itself, we will make use of software from the sensory processing portion of the SAWL system. Some of this software will require modification for use in this new system.

The most significant modification to our sensory systems involves providing the synchrony detection mechanism access to the raw input data and processed features for both the audio and visual senses. Input is obtained from digital video files (MPEG format), and currently these audio and visual inputs are processed separately in our system, with integration of this processed information taking place only when being input to the word-learning ANN.

Adjustments to the specific audio and visual features being processed may also need to be investigated, but it is not expected to require extensive modification as Hershey and Movellan (2000) used relatively simple representations of audio and visual energy in their work. Similar features are processed by components of the sensory systems of our current SAWL model.

Applying the spatial localization information to image segmentation will require further effort. Image segmentation techniques, such as seeded region growing, have been used for purposes of facial segmentation (Grinias, Mavrikakis, & Tziritas, 2001). The areas of strong synchrony as detected by Hershey and Movellan (2000) in their speaker localization task roughly corresponded to boundaries of the speaker's face. This is due to the fact that there is normally some movement of a speaker's head in synchrony with their speech, in addition to mouth and jaw movement. Because of this, using synchrony information in conjunction with the original image to grow or restrict a region of maximal synchrony, corresponding to a face, seems plausible. The color segmentation mechanism we are currently using in SAWL (Bruce, Balch & Veloso, 2000) could be applied to this task.

To increase accuracy in face segmentation and location, processing results will be integrated across several frames. Noise or variation in the data for any single frame could compromise our efforts because the detected synchrony would be dependent upon erroneous variations in both audio and visual signals. These variations can arise from environmental and

equipment noise in recording, and also from the process of MPEG format compression. Motion tracking of regions over several frames is one method to compensate for the effects of immediate variations.

In addition to changes to the sensory processing systems, the graphical user interface to the system will need to be modified as well. Configuration of parameters specific to the proposed system will be added, along with graphical displays of the processing.

### **Evaluation and Report**

The ability of this proposed system to localize sound sources in a video sequence and then apply this information to determining the boundaries of faces needs to be assessed. Criteria for measuring performance in these areas are rather difficult to determine. As mentioned, our video processing system in SAWL is currently capable of color image segmentation and template-based face detection, but direct comparison to these methods is inappropriate. Instead, human judgements of face location and boundaries on particular digital video files will be recorded, perhaps assisted by our existing automated methods. This will provide a reference for evaluating and comparing the performance of the proposed system.

### **Project Budget and Schedule**

This is proposed to be a 24-week project, starting July 1, 2003, for which I am requesting a full stipend of \$1,400. This would fund up to 129.15 hours, approximately five paid work hours per week, at the current rate of \$10.84 for students. Additionally, I am requesting a stipend of \$300 to offset registration and travel expenses to present this work at a scientific meeting. A 24-

week time frame will allow adequate time to implement the system, evaluate its performance, and prepare a report of the results.

### **Time Line**

Synchrony detection implementation:	4 weeks
Integration with audio/video systems:	5 weeks
Facial location/segmentation:	8 weeks
System testing/evaluation:	4 weeks
Summary and report of results:	3 weeks
Total:	24 weeks

### **Dr. Prince's Involvement**

Dr. Prince initiated and coordinates work on the SAWL model and related projects. As such his current research interests are directly related to this proposed project. While related, this project is a separate line of work for which Dr. Prince would provide advice and guidance as appropriate.

### **Research Motivation**

My motivations for pursuing this project are multifaceted. Currently I am studying Computer Science at the University of Minnesota Duluth and have been working with Dr. Prince since February 2001. During this time I have been involved with the development of SAWL, and am the developer of the auditory processing portion of the sensory layer. Multimedia applications have long been an interest of mine as well. Implementing this project would provide an opportunity to pursue new directions in work with which I have been involved, and expand my knowledge and experience in areas of interest.

## Works Cited

- Bruce, J., Balch, T., & Veloso, M. (2000). Fast and inexpensive color image segmentation for interactive robots. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'00)*.
- Gogate, B. A., & Bahrick, L. E. (1998). Intersensory redundancy facilitates learning of arbitrary relations between vowel sounds and objects in seven-month-old infants. *Journal of Experimental Child Psychology*, 69, 133-149.
- Grinias, I., Mavrikakis, Y., & Tziritas, G. (2001). Region growing colour image segmentation applied to face detection. *International Workshop on Very Low Bitrate Video Coding*.
- Hershey, J., & Movellan, J. (2000). Audio-vision: Using audio-visual synchrony to locate sounds. In S. A. Solla, T. K. Leen, & K. R. Muller (eds.), *Advances in Neural Information Processing Systems 12* (pp. 813-819). Cambridge, MA: MIT Press.
- Hollich, G. J. (1999). *Mechanisms of Word Learning: A Computational Model*. Unpublished Dissertation, Temple University, Philadelphia, PA.
- Li, S. Z., Zhu, L., Zhang, Z., & Zhang, H. (2002). Learning to detect multi-view faces in real-time. *The 2nd International Conference on Development and Learning (ICDL02)*. IEEE Press.
- Nandy, D., & Ben-Arie, J. (1996). Estimating the azimuth of a sound source from the binaural spectral amplitude. *IEEE Transactions on Speech and Audio Processing*, 4, No. 1, January 1996, 45-55.
- Prince, C. G., Mislivec, E. J., Kosolapov, O. V., & Lykken, T. R. (2002). Towards a theory grounded theory of language. *The 2nd International Conference on Development and Learning (ICDL02)*. IEEE Press.
- Prince, C. G., Mislivec, E. J., & Helder, N. A. (submitted). Towards an ontogenetic design mindset: Advancing developmental algorithms. *International Joint Conference on Artificial Intelligence (IJCAI-03)*.
- Sinha, P. (1995). *Perceiving and Recognizing Three-dimensional Forms*. Unpublished Dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Viola, P., & Jones, M. J. (2001). *Robust Real-Time Object Detection*. Technical report, Cambridge Research Laboratory, Compaq Computer Corporation.