Detecting local audio-visual synchrony in monologues utilizing vocal pitch and facial landmark trajectories

Steven Cadavid[1]
s.cadavid1@umiami.edu

Mohamed Abdel-Mottaleb[1]
mottaleb@miami.edu

Daniel S. Messinger[2]
dmessinger@miami.edu

Mohammad H. Mahoor[3]
mmahoor@du.edu

Lorraine E. Bahrick[4]
bahrick@fiu.edu

[1] University of Miami
Department of Electrical and Computer Engineering

[2] University of Miami
Department of Electrical and Computer Engineering

[3] University of Denver
Department of Electrical and Computer Engineering

[4] Florida International University
Department of Psychology

Speech is a verbal means of communication that is intrinsically bimodal: the audio signal is produced by complex mouth and corporal articulations that form the basic vocal tone into specific, decodable sounds. Both the audible and visible contents of speech carry pertinent information about what is being conveyed.

The motivation behind this work is to derive a synchrony measure between the visual contents of a monologue and its corresponding audio signal. While most of the work in the literature focus on a macro-level analysis of synchrony [1, 3], such as speaker localization and identity verification, we are interested in detecting anatomical features of a speaker that demonstrate synchrony between the sounds of speech (onset, offset) and the visible movements of the face and its features.

We are applying this work to a set of monologue video stimuli that are played to both typically-developing infants and infants who are at risk for autism between the ages of 6 and 10 months. Each monologue is represented by a synchronous (sync) version (e.g. the audio and visual signals are synchronized) and an asynchronous (async) version (e.g. the audio signal is time-shifted with respect to the visual signal).

In this paper, we present an audio-visual synchrony algorithm that employs a Gaussian mutual information method to evaluate the synchrony between vocal pitch and facial landmark trajectories. Pitch is an important feature for detecting the emotional state of a speaker. It provides discernment on the irony, sarcasm, emphasis, contrast and focus of an utterance, which may not be encoded by grammar. The Active Shape Model (ASM) has been widely used for reliably tracking facial landmarks across a sequence of video frames.

In information theory, the mutual information, $M(X,Y)$, between two Gaussian random variables, X and Y, is a quantity that measures the mutual dependence of the two variables. In the case that the random variables are discrete, it is defined as $M(X,Y) = 0.5 \cdot \log\left(|\Sigma X| \cdot |\Sigma Y| / |\Sigma X, Y|\right)$ where Σ denotes the covariance matrix and $|\cdot|$ is the determinant. We use this measure of Gaussian mutual information to compute the temporal contingency between the visual and audio features.

The Active Shape Model, introduced by Cootes et al. [2], is a statistical approach for shape modeling and feature extraction. It represents a target structure by a parameterized statistical shape model obtained from training. The location of n facial landmarks are annotated on a set of training images by a human expert. This set of landmarks is represented by a vector $X = \{(x_i, y_i)\}_{i=1}^{n}$ where x_i and y_i are the coordinates of the i^{th} landmark. Then, by analyzing the variations in shape over the training set, a model is built which can represent these variations.

In this work, the shape model is comprised of 83 landmarks that correspond to salient features on the human face. For each landmark, The absolute horizontal and vertical displacement between adjacent frames are used as the two visual features for computing synchrony.

We employ a frequency domain approach described in [4] for determining the pitch of each temporal set of audio samples, $A(t_k)$. Firstly, the short-term spectrum function, $A(f)$, is obtained by applying the Fourier transform to $A(t_k)$. Suppose that the fundamental frequency is denoted by f_0, then the *sum of the harmonic amplitudes* is defined as $SH = \sum_{n=1}^{N} A(nf_0)$, where N is the number of harmonics to be considered. If only the subharmonic frequencies are considered, equalling one half of f_0, then the *sum of subharmonic amplitudes* is given as $SS = \sum_{n=1}^{N} A\left((n-1/2) f_0\right)$. The subharmonic-to-harmonic ratio (SHR), given by $SHR = SS/SH$, is the amplitude ratio between subharmonics and harmonics. We utilize the absolute difference between the pitch estimates of adjacent audio bins as the audio feature.

The human visual system is capable of distinguishing rigid and non-rigid motion of an articulator during speech. In an attempt to emulate this process, we separate rigid and non-rigid motion and compute the synchrony attributed to each. Non-rigid and rigid motion are separated using pose normalization. To obtain the non-rigid motion, the fitted landmarks of each video frame are registered to the landmarks of the reference frame (first frame) using Procrustes alignment.

The acquired raw synchrony estimates generally contain false-positive synchrony values due to over-sensitivity. We propose a postprocessing method that filters out spurious synchrony by accounting for the onset, offset, and mean synchrony energy of audio-visual events. The synchrony filtering is performed by classifying each audio event (e.g. word, phrase) of both the sync and async versions of the video clip as being either synchronized or asynchronized with its coinciding visual events (e.g. the magnitude displacement of a facial landmark). We perform synchrony filtering separately for each facial landmark. That is, the classification of an audio event is determined on a per facial landmark basis. If an audio event is classified as being synchronized with the coinciding visual events of a given facial landmark, then the synchrony values of the facial landmark are retained, otherwise they are discarded.

We conducted a series of experiments to evaluate the performance of the system described above. The experiments are conducted on 20 pairs of monologue video clips where each pair consists of a sync and async version of the video clip with identical visual content. In each video clip, the speaker is articulating a set of phrases using child-directed speech. For each pair of video clips, the difference in the amount of synchrony detected between the sync and async versions of the video clip is computed. The amount of synchrony detected in the sync video clip is compared against that of the async video clip, where the offset of the audio signal in the async video clip is varied.

The results indicate that the proposed method is capable of detecting a greater amount of synchrony in the sync video clip than in its async counterpart across 97.2% of the experimental trials. The results also illustrate that the amount of synchrony detected for the rigid motion generally surmounts that of the non-rigid motion and the combined non-rigid + rigid motion . Although the result of this motion analysis is somewhat surprising, it is understandable because phrase/words that are communicated using child-directed speech are often accompanied by an exaggerated level of looming (rigid) motion.

[1] H. Bredin, A. Miguel, I.H. Witten, and G. Chollet. Detecting replay attacks in audiovisual identity verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 621–624, May 2006.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models—their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[3] J.W. Fisher III and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Transactions on Multimedia*, 6(3):406–413, June 2004.

[4] X. Sun. A pitch determination algorithm based on subharmonic-to-harmonic ratio. In *the 6th International Conference of Spoken Language Processing*, pages 676–679, 2000.