# 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
Coral Gables, FL 33146

[2] University of Miami
Department of Psychology
Coral Gables, FL 33146

[3] University of Denver
Department of Electrical and Computer Engineering
Denver, CO 80208

[4] Florida International University
Department of Psychology
Miami, FL 33199

**Abstract**

We describe a novel approach for determining the audio-visual synchrony of a monologue video sequence utilizing vocal pitch and facial landmark trajectories as descriptors of the audio and visual modalities, respectively. The visual component is represented by the horizontal and vertical displacement of corresponding facial landmarks between subsequent frames. These facial landmarks are acquired using the statistical modeling technique, known as the Active Shape Model (ASM). The audio component is represented by the fundamental frequency, or pitch, obtained using the subharmonic-to-harmonic ratio (SHR). The synchrony between the audio and visual feature vectors is computed using Gaussian mutual information. The raw synchrony estimates obtained using this method may contain spurious synchrony values due to over-sensitivity. A filtering method is employed for discarding synchrony values that occur during non-associated audio/visual events. 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. Experiments are conducted on a dataset of monologue video clip pairs. Each pair is composed of an asynchronous and synchronous version of the video clip. For the asynchronous video clips, the audio signal is displaced with respect to the visual signal. Experimental results indicate that the proposed approach is successful in detecting facial regions that demonstrate synchrony, and in distinguishing between synchronous and asynchronous sequences.



BMVC 2009 doi:10.5244/C.23.10

# 1 Introduction

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 [3, 4, 6, 7, 8], 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. We are interested in analyzing the correlation between an infants' gaze data and regions of the stimuli that exhibit synchrony. Generating ground truth data for individual facial regions is a difficult task for a human expert because of the highly complex relationship that exists between the audio and visual signals. An example of a challenge is that changes in the audio signal do not necessarily imply changes in the visual signal and vice versa. Our objective is to develop a systematic approach for determining synchronous facial regions in a video frame based on computed synchrony using computer vision techniques.

Audio-visual synchrony algorithms are generally comprised of two principal stages - 1) front-end processing and 2) evaluating synchrony between audio and visual features. Front-end processing and feature extraction aim to reduce the raw input data in order to achieve a good subsequent modeling. The second stage uses the features acquired during the front-end processing stage to measure the correlation between the video and audio signals in order to compute a synchrony measure.

Several features for describing the audio and visual components of a video sequence have been reported in the literature. For the audio component, feature extraction has been performed in both the frequency and time domains. Time domain features, such as the root mean square amplitude and log energy [1, 3], generally describe the intensity of the audio signal by aggregating a sequence of audio samples and computing their average acoustic energy [11]. Frequency domain features, such as Periodograms [8], Spectograms [2], line spectral frequencies [19] and particularly Mel-frequency cepstral coefficients (MFCC) [6, 12, 16], are frequently used because they are the state-of-the-art in parameterization for speech processing [18]. Visual speech features are categorized into the use of raw pixel intensities [3, 11, 12], holistic methods [16], lip-shape methods [7, 9], and dynamic features [4].

Synchrony evaluation methods are generally categorized as methods that assume a linear correlation between the audio-visual feature vectors [11, 13, 17], and those that model the correlation using parametric [16, 18] and non-parametric models [6].

In this paper, we present an audio-visual synchrony algorithm that employs a Gaussian mutual information method [11, 17] 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 insight on whether an utterance is a statement, a question, or a command. Furthermore, pitch 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.

This paper is organized as follows: Section 2 describes the proposed approach including audio and visual feature extraction using ASM and pitch detection, respectively, and a filtering process, based on Hierarchical K-means clustering, used for eliminating spurious synchrony. An experimental evaluation is presented in Section 3, followed by conclusions and future work in Section 4.

# 2 System Approach

In the following sections, we present a system for detecting audio-visual synchrony based on Gaussian mutual information. Synchrony detection is performed on audio and visual feature vectors that are representative of the two modalities. The mutual information algorithm requires that the feature vectors be of equal lengths. Generally, audio sampling rates are significantly higher than those of video. To cope with this difference, the audio samples are partitioned and aggregated into bins such that a one-to-one correspondence between video frames and audio bins is established. The audio samples in a bin are used to estimate the fundamental frequency, or pitch, of the bin. This process is repeated on all bins so that each one has a corresponding pitch estimate. The visual component is represented by 83 facial landmarks, which have been tracked across the video sequence using the ASM technique. The visual feature vector is composed of the horizontal and vertical displacement of corresponding facial landmarks between subsequent video frames.

The synchrony values yielded by the Gaussian mutual information method are then filtered by accounting for the onset and offset of audio and visual events. The audio signal is partitioned into events such as words or phrases. The hierarchical k-means clustering technique is employed to derive a classification of synchronous and asynchronous audio events. The dimensions of the classification space consist of 1) the distance of an audio event's onset (the beginning of an audio event) to its nearest visual event, 2) the distance of an audio event's offset (the end of an audio event) to its nearest visual event, and 3) the mean of the synchrony values detected during the audio event. Figure 1 illustrates the system overview.

## 2.1 Gaussian Mutual Information

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:

$$M(X,Y) = \frac{1}{2}\log_2 \frac{|\Sigma X|\,|\Sigma Y|}{|\Sigma X, Y|} \tag{1}$$

where $\Sigma$ denotes the covariance matrix and $|\cdot|$ is the determinant. Mutual information is both non-negative ($M(X,Y)) \geq 0$) and symmetrical ($M(X,Y) = M(Y,X)$). It can also be shown that the random variables are independent if and only if $M(X,Y) = 0$. We use this measure of Gaussian mutual information to compute the temporal contingency between a video and audio signal. Consider a sequence of $s$ consecutive frames of visual data and audio data co-occurring with those visual frames. Let $V(p_x, p_y, t) \in \Re^2$ be a vector describing the spatial location, $(p_x, p_y)$, of a visual feature at time $t$. Likewise, let $A(t) \in \Re^1$ be a scalar describing the audio signal at time $t$. Now, consider a set of audio-visual vectors $W(V(p_x, p_y, t_l), A(t_l))_{l=t-s+1,\cdots,t}$ sampled at times $t_k - s + 1, \cdots, t_k$ and

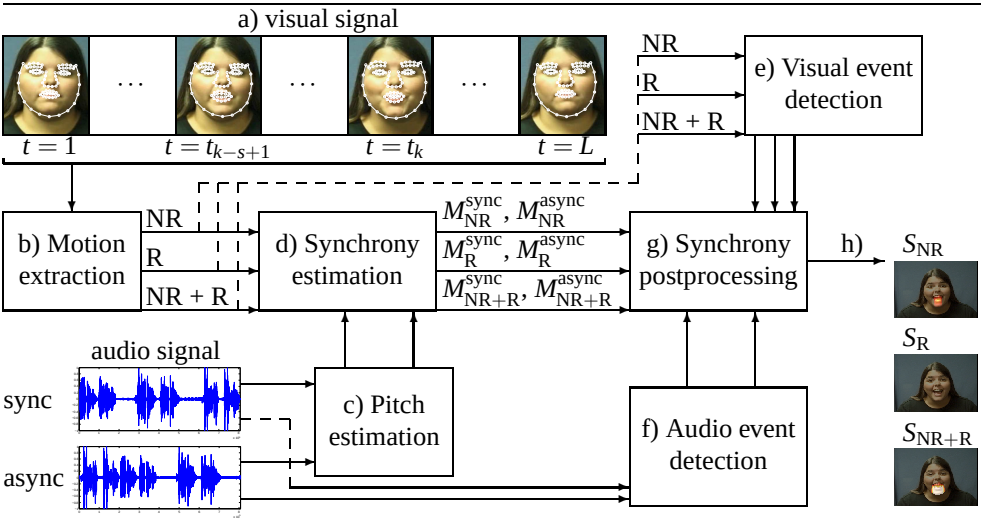


Figure 1: System overview. a) Facial landmarks are tracked across a monologue video sequence. b) The non-rigid (NR), rigid (R), and combined non-rigid + rigid (NR + R) motion is extracted from the mesh models. c) The pitch estimates are derived from both the sync. and async. raw audio inputs. d) Synchrony is estimated separately for the three motions, analyzing a temporal window of $s$ frames at a time. Synchrony for each motion is computed for both the sync. and async. pitch estimates. e) Visual events are detected separately for the three motions. f) Audio events are detected separately for both the sync. and async. raw audio inputs. g) Postprocessing is performed to filter out spurious synchrony values. An output is obtained for all three motions. h) An example output of the synchrony obtained for a video frame between the three motions and the synchronous pitch estimates. Colors superimposed on the images represent detected synchrony; the colors range from dark orange (minimal synchrony) to bright orange (maximal synchrony).

at the visual feature location $(p_x, p_y)$. Both $A(t_k)$ and $V(p_x, p_y, t_k)$ are presumed dependent Gaussian random variables with respective probability distributions $\mathrm{N}_1\left(\mu_{A(t_k)}, \Sigma_{A(t_k)}\right)$ and $\mathrm{N}_2\left(\mu_{V(p_x,p_y,t_k)}, \Sigma_{V(p_x,p_y,t_k)}\right)$, where $\mu$ denotes the sample mean and $\Sigma$ is the covariance. Moreover, each vector $W$ is considered as an independent sample from a joint, multi-variate Gaussian random process with probability distribution $\mathrm{N}_3\left(\mu_{W(p_x,p_y,t_k)}, \Sigma_{W(p_x,p_y,t_k)}\right)$.

An additional term, $\left(1 - \frac{1}{2^{r\alpha}}\right)$, has been multiplied to (1) to decrease the effect of small, sub-audible audio features that are accidently correlated with the visual features [17]. Variable $r$ is set to the maximum value of the audio vector at time $t_k$, and $\alpha = 50$ is a fixed value determined empirically.

## 2.2 Active Shape Model

The Active Shape Model (ASM), introduced by Cootes et al. [5], is a statistical approach for shape modeling and feature extraction. It has been subsequently improved in recent years [15]. It represents a target structure by a parameterized statistical shape model obtained from training. The location of $n$ points, commonly referred to as landmarks, are annotated on a set of training images by a human expert. This set of landmarks is represented by a vector $X = (x_1, y_1, \ldots, x_n, y_n)^T$ 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:

$$X \approx \bar{X} + Pb \tag{2}$$

The vector $\bar{X}$ contains the mean values of the coordinates of the annotated data, $P$ is a matrix of the first $t$ eigenvectors of the covariance matrix of the data, and $b$ is a vector that defines the model parameters. The variance of the $i^{th}$ parameter, $P_i$, across the training set is given by the corresponding eigenvalue $\lambda_i$. By limiting the parameters $b_i$ in the range of $\pm 3\sqrt{\lambda_i}$, we ensure that the generated shape is similar to those in the original training set. To apply the constructed shape model to a given target, a transfer function is required to move from the model coordinate system to the image coordinate system. Typically, this is achieved by a Euclidean transformation defining the translation $(X_t, Y_t)$, rotation $\theta$, and scale $s$. The position of the landmarks, $X$, in the image are then given by:

$$X = T_{X_t, Y_t, \theta, s}(\bar{X} + Pb) \tag{3}$$

For a given new image, the ASM is performed to find where the target object lies on the image. Therefore, we need to find the optimum parameters of the ASM that best fit the model to the target structure. Generally, this optimization problem is solved iteratively. Firstly, the model is initialized by the mean shape. Secondly, a region of the image around each feature point is examined to find the best nearby match (i.e. searching along the profile line for the edge locations). Thirdly, the parameters $X_t$, $Y_t$, $s$, and $\theta$ are updated to best fit the new found landmarks. Lastly, the constraint $|b_i| < 3\sqrt{\lambda_i}$ is applied to the parameters $b_i$. These steps are repeated until there is no significant change in the shape parameters.

In this work, the shape model is comprised of 83 landmarks corresponding to anatomical features on the human face. For each video clip, a model is trained based on the manual annotations of five percent of the video frames. The facial landmarks of the remaining 95% of the video frames are automatically tracked utilizing the trained model. For each tracked landmark, The absolute horizontal and vertical displacement between adjacent frames are used as the two visual features for computing synchrony.

## 2.3 Pitch Detection

Pitch (i.e. fundamental frequency) is an important feature in prosody modeling, and is one of three attributes, along with loudness and quality, for characterizing speech. Pitch determination algorithms (PDA) are designed to estimate the fundamental frequency of a quasiperiodic or virtually periodic signal, such as a digital recording of speech. The fundamental frequency of speech can vary from 40 Hz for low-pitched male voices to 600 Hz for children or high-pitched female voices. PDAs are generally categorized into time domain and frequency domain approaches.

We use a frequency domain approach described in [20] 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:

$$SH = \sum_{n=1}^{N} A(nf_0) \tag{4}$$

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:

$$SS = \sum_{n=1}^{N} A\left(\left(n - \frac{1}{2}\right) f_0\right) \tag{5}$$

The subharmonic-to-harmonic ratio (SHR), given by:

$$SHR = \frac{SS}{SH} \tag{6}$$

is the amplitude ratio between subharmonics and harmonics.

In practice, solving for the SHR directly in (6) is not tractable. Therefore an alternative method, known as the Subharmonic Summation algorithm (SHS), is used [10]. The SHS method sums the spectrum at harmonics of the pitch candidate, and subtracts the spectrum at the middle points between harmonics. Finding the pitch estimate, $f_0$, amounts to solving a maximization problem described in [20]. We utilize the absolute difference between the pitch estimates of adjacent audio bins as the audio feature.

## 2.4 Classifying synchrony based on rigid and non-rigid motion

The human visual system is capable of distinguishing the rigid and non-rigid motion of an articulator during speech. An advantage of the ASM technique is that non-rigid and rigid motion can be separated. Non-rigid motion is due to local deformations such as lip movement and eye constriction. Contrariwise, rigid motion is attributed to head pose variations and looming. In our experiments, we compute synchrony attributed to non-rigid motion, rigid motion, and the combined non-rigid + rigid motion.

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 analysis. Procrustes analysis computes an optimal affine transformation between the two landmark sets by minimizing the alignment error. Each video frame has a corresponding transformation matrix that is used to align the corresponding landmarks to the landmarks contained within the reference frame. The alignment is driven by a set of typically rigid facial landmarks, including the inner and outer eye corners, the bottom of the nose, and the temples. Aligning the landmarks of all frames to the reference frame effectively negates the rigid motion of the landmarks, retaining only the non-rigid motion.

To obtain the rigid motion of the facial landmarks, the inverse of the transformation matrices acquired from the aforementioned non-rigid motion process are each applied to the landmarks of the reference frame.

## 2.5 Unsupervised classification of synchrony using the onset and offset of audio events

The raw synchrony estimates obtained using the aforementioned method 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 (eg. word, phrase) of both the synchronous and asynchronous versions of the video

clip as being either synchronized or asynchronized with its coinciding visual events (eg. 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 may view the problem of classifying an audio event as an unsupervised classification problem, principally because of two reasons; Firstly, the visual events of a facial landmark may, coincidentally, be synchronized with an asynchronous audio signal. Consequently, it would be inappropriate to simply label all audio events contained within a synchronous and asynchronous video clip as synchronous and asynchronous, respectively. Secondly, it is difficult to manually establish a ground truth synchrony classification of audio events on a per facial landmark basis from which to train a supervised classifier. Therefore, we propose an unsupervised classification scheme, based on the Hierarchical K-means clustering method [14], for determining whether an audio event is synchronized or asynchronized with the coinciding visual events of a given facial landmark. It should be noted that all parameters reported in this section have been derived empirically.

The audio signal is first partitioned into events that correspond to either phrases or words surrounded by periods of silence. Each segmented phrase/word is termed an audio event. To partition the audio signal into events, the acoustic energy of the audio signal is first derived by computing its absolute value. Acoustic energy samples that do not surpass an amplitude threshold, $T_a^A = 0.04$, are determined to be silence, while those that do are considered potential candidates for audio events. The candidate samples then undergo connected component analysis to distinguish between consecutive runs of samples. The connected components that have a duration greater than $T_d^A = 300$ms are retained, while all others are discarded.

A similar process is performed on facial landmarks for detecting visual events. However, a fixed displacement threshold is unsuitable for determining visual events because each facial landmark possesses a distinct range of displacements. Landmarks along the lip contour, for instance, generally exhibit larger displacements than, say, eyebrow landmarks. Therefore, a facial landmark demonstrating a displacement greater than the $75^{\text{th}}$ percentile of displacements for the given facial landmark are considered potential candidates for visual events. Candidate samples demonstrating a connected components duration greater than $T_d^V = 150$ms, are declared as visual events.

The synchrony classification of audio events is then performed using the Hierarchical K-means clustering technique. It is well known that K-means is sensitive to initialization. Hence, K-means is run multiple times and the cluster centroids with the minimum intra-class variance are selected as the cluster centroids. Consider a facial landmark, $p_i, i = 1,\ldots,N$, tracked across a video sequence, with a corresponding set of visual events, $E^{p_i} = \left\{e_j^{p_i}\right\}_{j=1}^{M}$, and a set of audio events, $A = \{a_i\}_{i=1}^{Q}$. We construct a three-dimensional feature space utilizing three measures: 1) the Euclidean distance from the onset of an audio event to the nearest visual event ($D_{\text{on}}$), 2) the Euclidean distance from the offset of an audio event to the nearest visual event ($D_{\text{off}}$), and 3) the mean synchrony energy of the audio event ($M$). For a given audio event, $a$, and a given facial landmark, $p$, the three measures are formally expressed as:

$$D_{\text{on}}^{p}(a) = \min_{x \in E^p} |\text{ind}(a_{\text{on}}) - \text{ind}(x)| \tag{7}$$

$$D^p_{\text{off}}(a) = \min_{x \in E^p} |\text{ind}(a_{\text{off}}) - \text{ind}(x)| \tag{8}$$

$$M^p(a) = \frac{1}{\text{ind}(a_{\text{off}}) - \text{ind}(a_{\text{on}}) + 1} \sum_{i=\text{ind}(a_{\text{on}})}^{\text{ind}(a_{\text{off}})} s_i^p \tag{9}$$

where $\text{ind}(a_{\text{on}})$, $\text{ind}(a_{\text{off}})$, and $s_i^p$ denote the index position (in terms of video frames) of the audio event's onset, the index position of the audio event's offset, and the $i^{\text{th}}$ frame's synchrony value of facial landmark $p$, respectively. The samples in the feature space are comprised of the audio events contained within the synchronous and asynchronous version of the video clip. Intuitively, samples that have smaller values for (7) and (8) and a larger value for (9) should be synchronous audio events. Therefore, the cluster centroid that is closest to $C_S = \{0,0,3\}$ is designated as the synchrony centroid (where 3 is typically the largest synchrony value encountered), and the further centroid as the asynchrony centroid. Synchrony values that coincide with the audio event samples contained within the synchrony cluster are classified as synchronous, otherwise they are classified as asynchronous. This classification process is then repeated for all facial landmarks.

# 3 Experimental Results

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 synchronous and asynchronous 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. An example phrase is "come over here by me!". Each phrase is separated by approximately one second of silence accompanied by no facial movement.

For each pair of video clips, the difference in the amount of synchrony detected between the synchronous and asynchronous versions of the video clip is computed. The amount of synchrony detected in the synchronous video clip is compared against that of the asynchronous video clip, where the offset of the audio signal in the asynchronous video clip is varied. In computing the synchrony difference, the audio signal of the asynchronous video clip is time-shifted with respect to the video signal from one second to four seconds, at an interval of one second. When the audio signal is offset, it is circularly shifted so that the end of the audio signal is wrapped around to the beginning. The total duration of each video clip is approximately 17 seconds. The video clips were exported in an uncompressed format with a spatial resolution of $720 \times 486$ pixels. For each video clip, we applied the synchrony algorithm to the non-rigid motion, rigid motion, and combined non-rigid + rigid motion.

It is difficult to assess the performance of an audio-visual synchrony system primarily because it is very hard to obtain a ground truth measure of synchrony. Currently, there are no publicly-available video clip datasets containing ground truth synchrony measures. For this reason, we have developed a quantitative measure for determining whether the system is capable of detecting a greater amount of synchrony in a synchronous monologue than in an asynchronous monologue.

We first construct two histograms of all synchrony values detected within the synchronous and asynchronous versions of the video clip, respectively. Synchrony values typically range between 0 (no synchrony) and 3 (maximum synchrony). The bin centers of the histograms

are within the range of 0 to 3 at an interval of $3/25$, resulting in 26 bins. We compute the distance between the synchrony histograms of the synchronous and asynchronous video clips using the weighted signed distance, given by:

$$D\left(h^S, h^A\right) = \frac{\sum_{i=1}^{n} w_i \left(h_i^S - h_i^A\right)}{\min\left(\sum_{i=1}^{n} w_i h_i^S, \sum_{i=1}^{n} w_i h_i^A\right)} \tag{10}$$

where $h_i^S$ and $h_i^A$ denotes the number of samples contained within the i$^{\text{th}}$ bin of the synchronous and asynchronous histograms, respectively, $n$ is the total number of bins, and $w_i$ signifies the center of the i$^{\text{th}}$ bin. If the histogram distance in (10) results in a positive number, it signifies that a greater amount of synchrony was detected in the synchronous video clip than in the asynchronous video clip.

Table 1 illustrates the computed histogram distances of the monologue video pairs. For conciseness, we report the results for nine video clips. As previously mentioned, four trials are conducted for each pair of monologues, where the offset of the audio signal in the asynchronous version of the video clip is varied. The computed distances are further subdivided into non-rigid motion (NR Motion), rigid motion (R Motion), and the combined non-rigid + rigid motion (NR + R Motion). An average histogram distance is also given at the bottom of each column.

Table 1: Histogram distance using weighted signed distance

| | Histogram Distance | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | audio offset = 1 sec. | | | audio offset = 2 sec. | | | audio offset = 3 sec. | | | audio offset = 4 sec. | | |
| Sub. | NR | R | NR+R | NR | R | NR+R | NR | R | NR+R | NR | R | NR+R |
| 1 | 0.91 | 0.91 | 0.91 | 0.99 | 0.80 | 0.84 | 0.26 | 0.80 | 0.51 | 0.67 | 0.71 | 0.56 |
| 2 | 0.26 | 0.34 | 0.28 | 1.04 | 1.70 | 1.63 | 0.38 | 0.55 | 0.53 | 0.16 | 0.21 | 0.20 |
| 3 | 0.03 | 0.63 | 0.48 | 0.41 | 0.42 | 0.36 | 0.37 | 0.58 | 0.26 | 0.26 | 0.20 | 0.26 |
| 4 | 0.30 | 0.26 | 0.31 | -0.05 | 1.39 | 0.62 | 0.36 | 0.61 | 0.41 | 0.07 | -0.09 | 0.01 |
| 5 | 3.00 | 3.00 | 2.00 | 0.12 | 0.72 | 0.40 | 1.80 | 1.79 | 1.82 | 1.86 | 1.78 | 1.79 |
| 6 | 0.39 | 0.53 | 0.46 | 0.28 | 0.28 | 0.27 | 0.23 | 0.47 | 0.43 | 0.51 | 0.66 | 0.54 |
| 7 | 0.17 | 0.23 | 0.27 | 0.08 | 0.73 | 0.63 | 0.29 | 0.15 | 0.18 | 0.18 | 0.18 | 0.18 |
| 8 | 0.24 | 0.81 | 0.43 | 0.34 | 0.57 | 0.38 | -0.04 | 0.19 | 0.04 | 0.17 | 0.43 | 0.25 |
| 9 | 0.24 | 0.24 | 0.23 | 1.01 | 1.01 | 1.01 | 1.16 | 1.31 | 1.31 | 0.50 | 0.50 | 0.50 |
| **Avg.** | **0.62** | **0.77** | **0.60** | **0.47** | **0.85** | **0.68** | **0.53** | **0.72** | **0.61** | **0.49** | **0.51** | **0.48** |

The results in Table 1 indicate that the proposed method is capable of detecting a greater amount of synchrony in the synchronous video clip than in its asynchronous counterpart across 97.2% of the experimental trials. The results also illustrate that the amount of synchrony detected for the rigid motion (overall synchrony value of 0.71) generally surmounts that of the non-rigid motion (overall synchrony value of 0.53) and the combined non-rigid + rigid motion (overall synchrony value of 0.59). 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.

# 4 Conclusion

In this paper, we presented a method utilizing Gaussian mutual information to determine the audio-visual synchrony between a synchronous and asynchronous version of a monologue video clip, respectively. For the asynchronous video clips, the audio signal is time-shifted with respect to the visual signal by one to four seconds. The audio and visual signals are represented respectively by an estimated pitch trajectory and the displacement between facial landmarks across adjacent frames. Non-rigid, rigid, and the combined non-rigid + rigid motion are extracted from the mesh models, and synchrony is computed separately for each. A postprocessing technique, based on hierarchical k-means clustering, is then employed to discard synchrony values that occur during non-associated audio/visual events. Experimental results demonstrate that the proposed method is capable of detecting a greater amount of synchrony in a synchronous video clip than in an asynchronous video clip. Furthermore, the motion analysis indicates that the rigid motion exhibits the greatest amount of synchrony, followed by the combined non-rigid + rigid motion.

We are applying this work to a set of monologue video stimuli that are played to infants between the ages of 6 and 10 months. We are interested in analyzing the correlation between the infants' gaze data and regions of the stimuli that exhibit synchrony. Our future and ongoing work in this area will include establishing a correlation measure between the captured gaze data and the synchrony values acquired through this method. This work will give us insight into the developmental processes of attention and awareness in both typically-developing infants and infants who are at risk for autism.

# References

[1] J. P. Barker and F. Berthommier. Evidence of correlation between acoustic and visual features of speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 199–202, June 1999.

[2] Z. Barzelay and Y.Y. Schechner. Harmony in motion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383344.

[3] 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.

[4] C.C. Chibelushi, F. Deravi, and J.S.D. Mason. A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37, March 2002.

[5] 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.

[6] R. Cutler and L. Davis. Look who's talking: speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo*, volume 3, pages 1589–1592, 2000.

[7] N. Eveno and L. Besacier. Co–inertia analysis for "liveness" test in audio-visual biometrics. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 257–261, September 2005.

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

[9] R. Goecke and B. Millar. Statistical analysis of the relationship between audio and video speech parameters for australian english. In *Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing*, pages 133–138, Saint–Jorioz, France, September 2003.

[10] D. J. Hermes. Measurement of pitch by subharmonic summation. *Journal of Acoustic Society of America*, 83:257–264, 1988.

[11] J. Hershey and J. Movellan. Using audio-visual synchrony to locate sounds. *Advances in Neural Information Processing Systems 12*, pages 813–819, 1999.

[12] G. Iyengar, H.J. Nock, and C. Neti. Audio-visual synchrony for detection of monologues in video archives. In *International Conference on Multimedia and Expo*, volume 1, pages 329–332, July 2003.

[13] E. Kidron, Y.Y. Schechner, and M. Elad. Pixels that sound. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 88–95, June 2005. doi: 10.1109/CVPR.2005.274.

[14] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[15] M. H. Mahoor and M. Abdel-Mottaleb. Facial features extraction in color images using enhanced active shape model. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, pages 144–148, Washington, DC, USA, 2006. IEEE Computer Society.

[16] H. J. Nock, G. Iyengar, and C. Neti. Assessing face and speech consistency for monologue detection in video. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 303–306, Juan-les-Pins, France, 2002. ACM.

[17] C. G. Prince and G. J. Hollich. Synching models with infants: a perceptual-level model of infant audio-visual synchrony detection. *Cognitive Systems Research*, 6(3):205–228, 2005.

[18] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41, January 2000.

[19] N. Sugamura and F. Itakura. Speech analysis and synthesis methods developed at ecl in ntt – from lpc to lsp. *Speech Commununication*, 5(2):199–215, 1986.

[20] 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.