International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

A Study of the Time Delay between Head Gestures and Pitch Patterns

Nallakkagari Phani Kumar

Abstract: ***We investigated the pragmatic effects of gesture-speech lag by asking participants to narrate four different stories in both English and in their mother tongue, in four conditions: sync, video or audio lag (±15 ms), audio only conditions: sync, video or audio lag (±15 ms), audio only. All three video groups rated the task as less difficult compared to the audio only group and performed better. The scores were slightly lower when sound preceded gestures (video lag), but not when gestures preceded sound (audio lag). Participants thus compensated for delays of 15 milli seconds in either direction, apparently without making a conscious effort. This greatly exceeds the previously reported time window for automatic multimodal integration.***

Keywords: Speech, Audio Lag, Video, Delay, Head Gestures, Integration, Stories

1. Introduction to Head Gestures

Head gestures reveal the way we see things and how we feel about them. As the sensory center of our body, the head turns towards the things we like, and away from the things we want to avoid.

Truth is, it's quite instinctive to "get it" when it comes to the meaning of head gestures. We know that a nod means a 'yes' and shaking the head means a 'no' (most of us anyway). We also learned to recognize many other more subtle movements subconsciously, meaning we get a certain feedback from them but often if we were asked why is that so - we couldn't say.

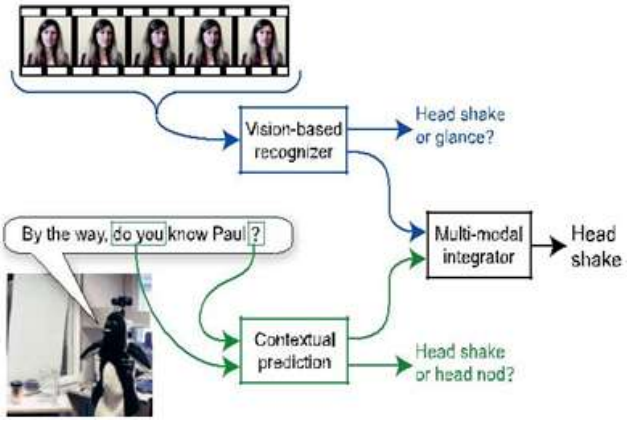

Figure 1: Contextual Recognition of Head Gestures

Figure 2: Sensor Arrangements to the Subject

In the above figure we can observe that totally we are using 10 sensors for the data acquisition. Among those sensors 4 were placed on the forehead (FH1, FH2, FH3, FH4) with the help of a band and 2 were placed on the nose (N1, N2) and the rest were placed on the hands (H1,H2,H3,H4). Camera calibration was done with the help of OPTITRACK's software. This software specifies the position of the markers in the subject's body (see Figure 1). A close-talk microphone is used for the audio recording purpose. We are obtaining 3 angles from each sensor (x, y & z) and mainly we are concentrating on the head gestures so we are using the angles from the forehead and nose and the sensors we have on hands are used for the recognition of start of speech. So, totally we are obtaining 18 angles from the head (6 x

3=18) and by using these angles we are calculating the rotational and translational angles.

2. Relationship between Pitch and Head Gestures

Pitch is defined as the perceived tone frequency of a sound in comparison with the perceptively best match with a pure sound.

During natural face-to-face conversations, a wide range of visual information from the movements of the face, head, and hands is available to conversational partners. In the work reported here, we studied the impact on speech perception of watching one component of this rich visual stimulus—a talker's head movements. It is well known that the intelligibility of degraded auditory speech is enhanced when listeners view a talker's lip movements. Watching these lip movements can also influence the perception of perfectly audible speech or be the sole basis of speech perception

People naturally move their heads when they speak, and our study shows that this rhythmic head motion conveys linguistic information. Three-dimensional head and face motion and the acoustics of a talker producing sentences were recorded and analyzed. The head movement correlated strongly with the pitch (fundamental frequency) and amplitude of the talker's voice. In a perception study, Subjects viewed realistic talking-head animations based on these movement recordings in a speech-in-noise task. The animations allowed the head motion to be manipulated without changing other characteristics of the visual or acoustic speech. Subjects correctly identified more syllables when natural head motion was present in the animation than when it was eliminated or distorted. These results suggest that nonverbal gestures such as head movements play a more direct role in the perception of speech than previously known.

Figure 3: Animated face with different speech sounds

Above figure is the "Two views of the animated face with different speech sounds being produced and with the head at different positions and orientations".

The advantage of using animation is that head motion can be systematically varied independently of the acoustics and face motion in order to determine the influence of head motion on speech perception.

3. Motivation for Study

When I was going through annotating the words from stories I had observed that for some certain words there is a definite head motion throughout the stories but there is delay in head motion with respect to the audio (pitch). Example, I had annotated the word "why_should_i" in that I observed there is a confirmed head motion but with some delay at some places.

Motivation 1

Manual gestures facilitate speech production, evidenced by the fact that they persist when blind people speak among themselves or when the listener is not visible. Furthermore, gestures may improve listening comprehension, especially when speech is ambiguous or when there is a lot of background noise. But how exactly are gestures temporally related to speech? How important is this temporal relation to successful communication? An influential view is that speech and gesture share a common origin and are best seen as two forms of the same communicative process. Their temporal relationship is determined by the semantic and pragmatic synchrony rules: if speech and gestures co-occur, they must either present the same semantic information or perform the same pragmatic function. It is well established that gestures are generally initiated simultaneously with – or slightly before – the onset of their lexical affiliates. But a new question immediately arises: Are they synchronized because this is necessary for successful comprehension or simply because speech and gesture stem from the same "idea unit"? One way to answer this question is to see how a disruption of the natural synchronization affects comprehension. Since speech and gesture exploit different modalities, this is a case of multisensory integration, which is affected by the synchronicity of the two channels. Of course, the time-window of tolerance for asynchrony varies depending on the nature of stimuli. Several studies have found effects of gesture asynchrony on event-related potentials elicited around 400 ms after the onset of a word (N400) indicative of integration difficulty. Habets et al. found a greater N400 to mismatched versus matched gesture-speech sequences only when speech lagged by 0 and 160, but not by 360 ms. The authors conclude that gesture and speech are integrated automatically when they fall within 160 ms of each other, so that a gesture which does not semantically match speech leads to effortful processing. Obermeier and Gunter found an N400 effect for gestures related to either dominant or subordinate meanings of an ambiguous word from approximately -200 ms (speech lag) to +120 ms (gesture lag). Other studies have found a greater perceived emphasis on words when they are synchronized with gestures.

Motivation 2

This article investigates the rhythmic relationship between gesture and speech. Four subjects were filmed in natural conversations with friends. From the resulting videos, several thousand time-stamped annotations pertaining to rhythm were manually recorded in a digital annotation tool, and exported for statistical analysis. They revealed a rich rhythmic relationship between the hands, head, and voice. Each articulator produced *pikes* (a general term for short, distinctive expressions, regardless of the modality) in complex synchrony with other articulators. Even eye blinks were synchronized, with eyelids held closed until reopening on the rhythmic beat, akin to a pre-stroke hold before a gestural stroke. Average tempos similar to previously reported natural human tempos — e.g. Fraise's (1982) 600 ms figure — were found in hands, head, and speech, although hands tended to move most quickly and speech most slowly. All three also shared a common tempo of around a third of a second, perhaps to synchronize inter-articulator meeting points. These findings lend empirical weight to earlier observations of a rhythmic relationship between gesture and speech, providing support for the theory of a common cognitive origin of the two modalities.

4. Mutual Information for Quantifying the Relation between Head Gestures and Pitch Patterns

Relative Entropy

- Relative entropy is a measure of the distance between two distributions.
- It is a measure of the inefficiency of assuming that the distribution is q when the distribution is p.

Mutual Information

- Entropy H(X) is the uncertainty (``self-information") of a single random variable
- Conditional entropy H (X|Y) is the entropy of one random variable conditional upon knowledge of another.
- The average amount of decrease of the randomness of X by observing Y is the average information that Y gives us about X.

Definition: The mutual information I(X; Y) between the random variables X and Y is given by

$$I(X;Y) = H(X) - H(X|Y)$$
$$= \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}$$
$$= E_p(X,Y)\log_2 \frac{p(X,Y)}{p(X)p(Y)}$$

Correlation Coefficient

The correlation coefficient, r, is a summary measure that describes the extent of the statistical relationship between two interval or ratio level variables. The correlation coefficient is scaled so that it is always between -1 and +1. When r is close to 0 this means that there is little relationship between the variables and the farther away from 0 r is, in either the positive or negative direction, the greater the relationship between the two variables. The two variables are often given the symbols X and Y. In order to illustrate how the two variables are related, the values of X and Y are pictured by drawing the scatter diagram, graphing combinations of the two variables. The scatter diagram is given first, and then the method of determining Pearson's r is presented. In presenting the following examples, relatively small sample sizes are given. Later, data from larger samples are given.

Correlation is given by

$$corr_{x,y} = \sum_{n=-\infty}^{\infty} x(n)y(n)$$
$$= \sum_{n=0}^{N-1} x(n)y(n)$$
$$corr - norm_{x,y} = \frac{\sum_{n=0}^{N-1} x(n)y(n)}{\sqrt{\sum_{n=0}^{N-1} x^2(n)\sum_{n=0}^{N-1} y^2(n)}}$$

The verbal and nonverbal channels of human communication are internally and intricately connected. As a result, gestures and speech present high levels of correlation and coordination. This relationship is greatly affected by the linguistic and emotional content of the message. The present paper investigates the influence of articulation and emotions on the interrelation between facial gestures and speech. The analyses are based on an audio-visual database recorded from subjects with markers attached to their face, who were asked to tell stories; a multilinear regression framework is used to estimate facial features from acoustic speech parameters. The levels of coupling between the communication channels are quantified by using correlation between the recorded and estimated facial features. The results show that facial and pitch features are not strongly interrelated, showing levels of correlation in the range of 0.4 to 0.6 when the mapping is computed at sentence-level using spectral envelope speech features. The results reveal that the lower face region provides the highest activeness and correlation levels. Furthermore, the correlation levels present significant intermotional differences, which suggest that emotional content affect the relationship between facial gestures and speech. Principal component analysis (PCA) shows that the audiovisual mapping parameters are grouped in a smaller subspace, which suggests that there is an emotion-dependent structure that is preserved from across sentences. The results suggest that this internal structure seems to be easy to model when prosodic-features are used to estimate the audio visual mapping. The results also reveal that the correlation levels within a sentence vary according to broad phonetic properties presented in the sentence. Consonants, especially unvoiced and fricative sounds, present the lowest correlation levels. Likewise, the results show that facial gestures are linked at different resolutions. While the facial area is locally connected with the speech, other facial gestures such as eyebrow motion are linked only at the sentence-level. The results presented here have important implications for applications such as facial animation and multimodal emotion recognition.

5. Experimental Setup

Here in this part we are performing the Mutual Information between pitch and angles (19, 20 & 21). At first we have deleted all the zero values from the pitch file and noted down the start and stop frame indices of a continuous numbered indices segment of pitch values and again we are eliminating the segment length whose length is below 30 in order to get the information from these we created a delay

between -50 to +50 milliseconds (±50, ±40, ±30, ±25, ±20, ±15, ±10, ±7, ±5, ±3, 0) by taking pitch as the reference and passing the angle values through this delay we have created 21 samples. Here we are mainly concentrating on the delays between 0 and ±10 milliseconds expecting some delay in head motion w.r.t. pitch After that we have concatenated all stories of the particular subject belongs to the particular delay and these concatenated angles are allowed to form a cluster of size 64 and the pitch file is to be allowed to 64 size cluster and then these 64 size clusters of pitch and angles are passed through the Mutual information program to get the MI's.

After that we have taken the mean of each delay samples and plotted. From those we observed that definitely delay is present, maximum mean values are shifting towards positive delay for most of the subjects.

MI – Mutual Information

Coefficient Correlation outputs

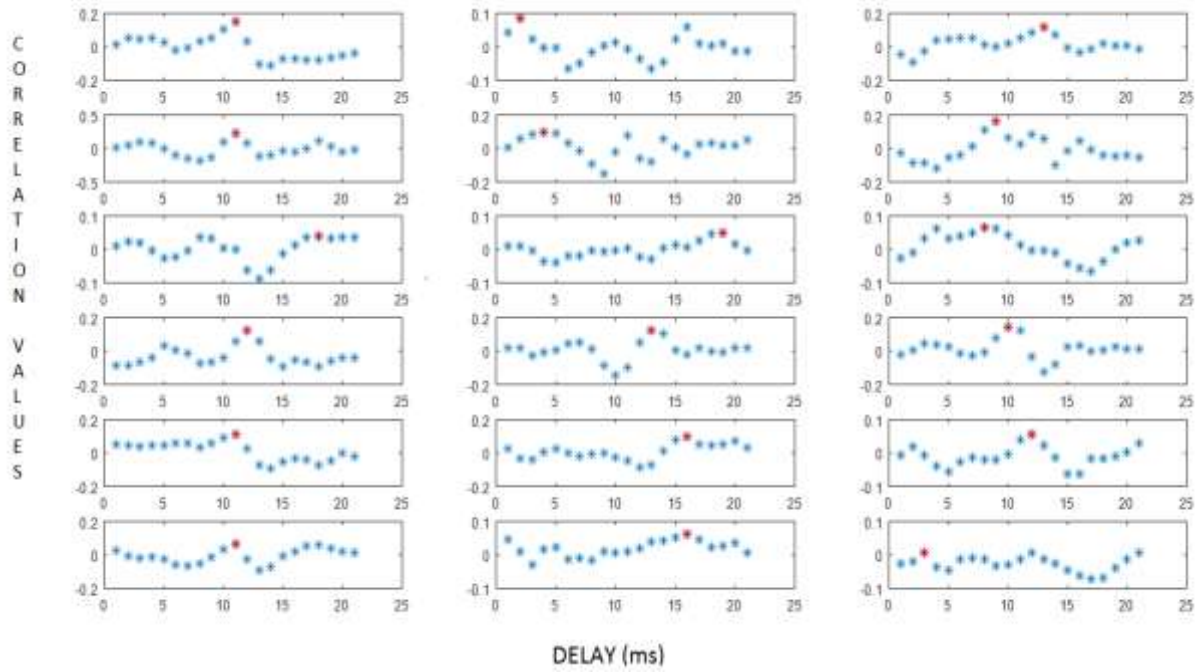

Figure 4: Plots of Delay (1:6) parameters vs Correlation Values (X, Y , Z AXIS)

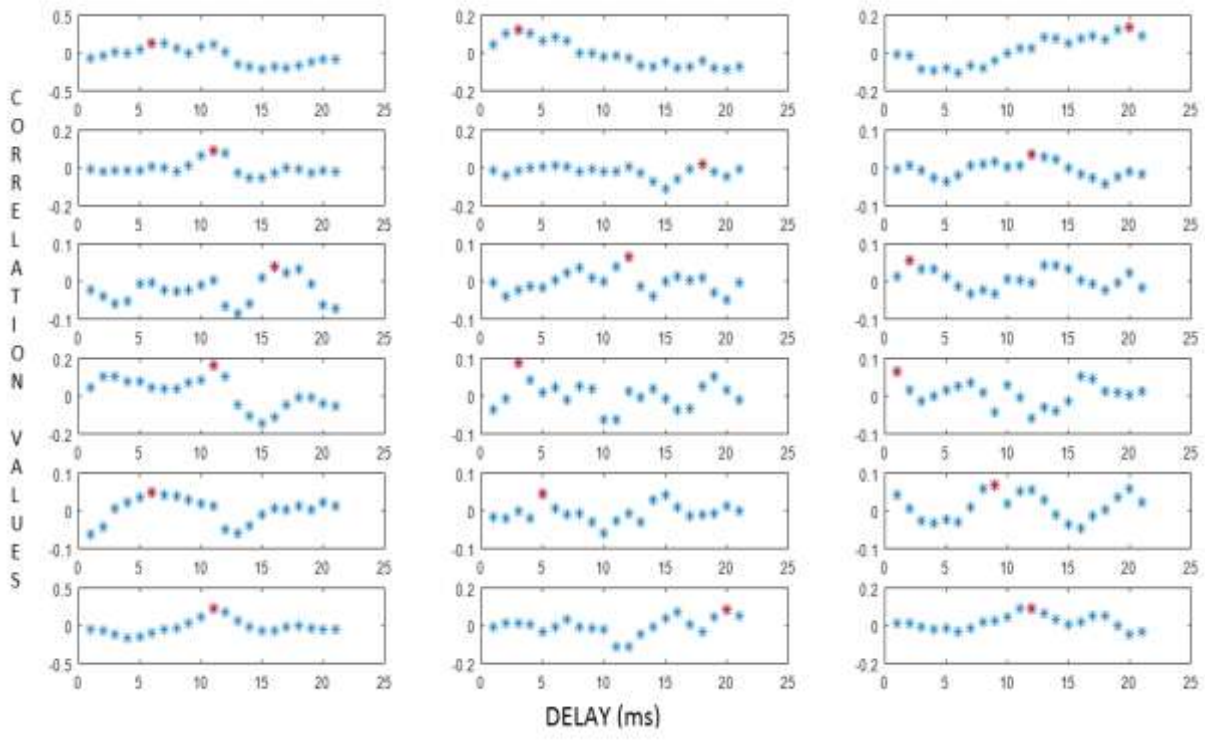

Figure 5: Plots of Delay(7:12) parameters vs Correlation Values (X, Y , Z AXIS)

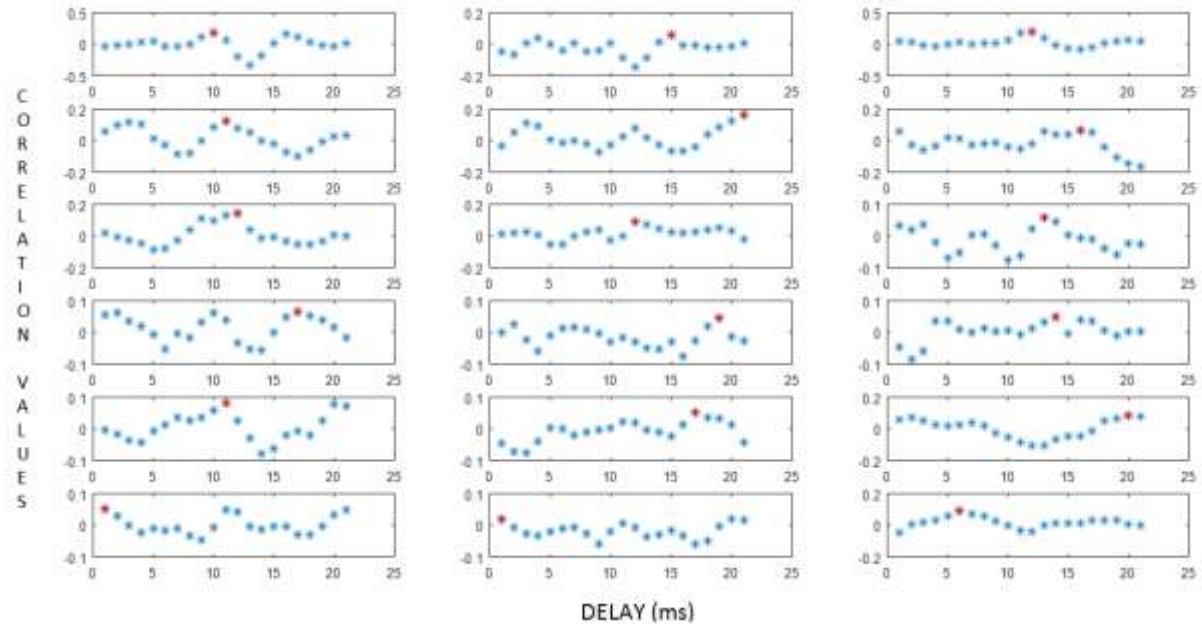

Figure 6: Plots of Delay(13:18) vs Correlation Values (X, Y , Z AXIS)

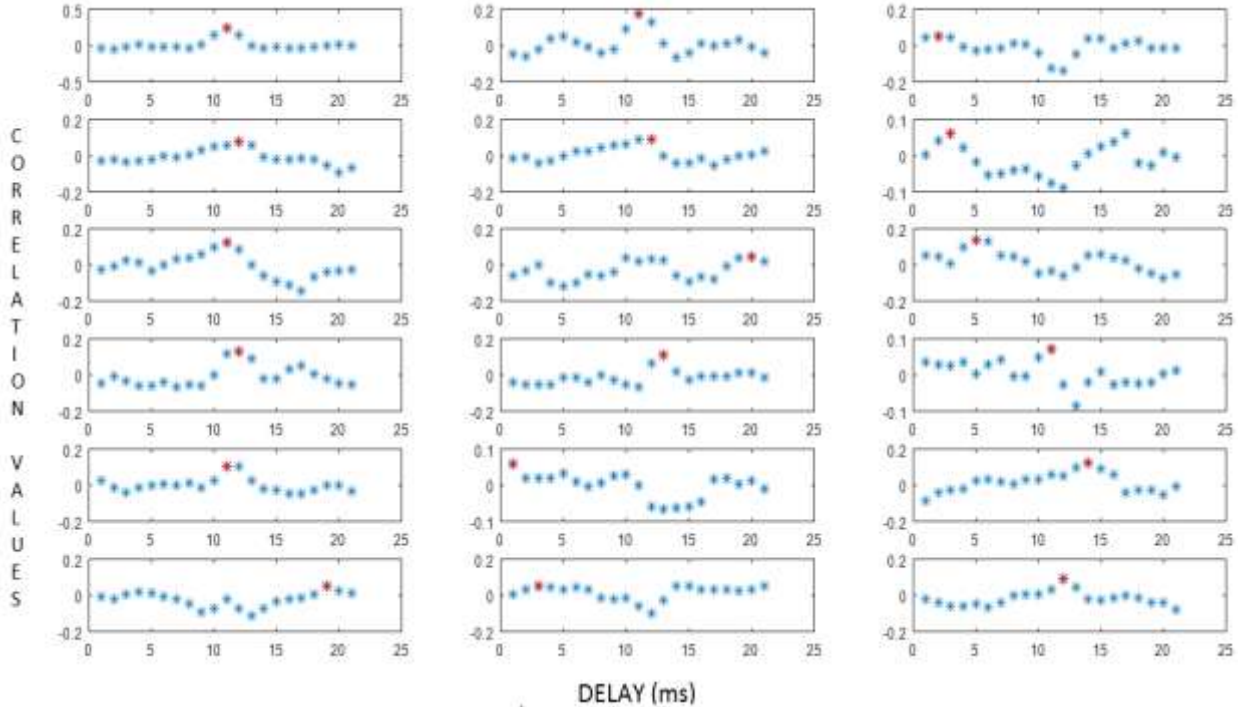

Figure 7: Plots of Delay(X,Y,Z and 22:24) parameter vs Correlation Values(X, Y , Z AXIS)

MI FOR ENGLISH & NATIVE LANGUAGE OUTPUTS

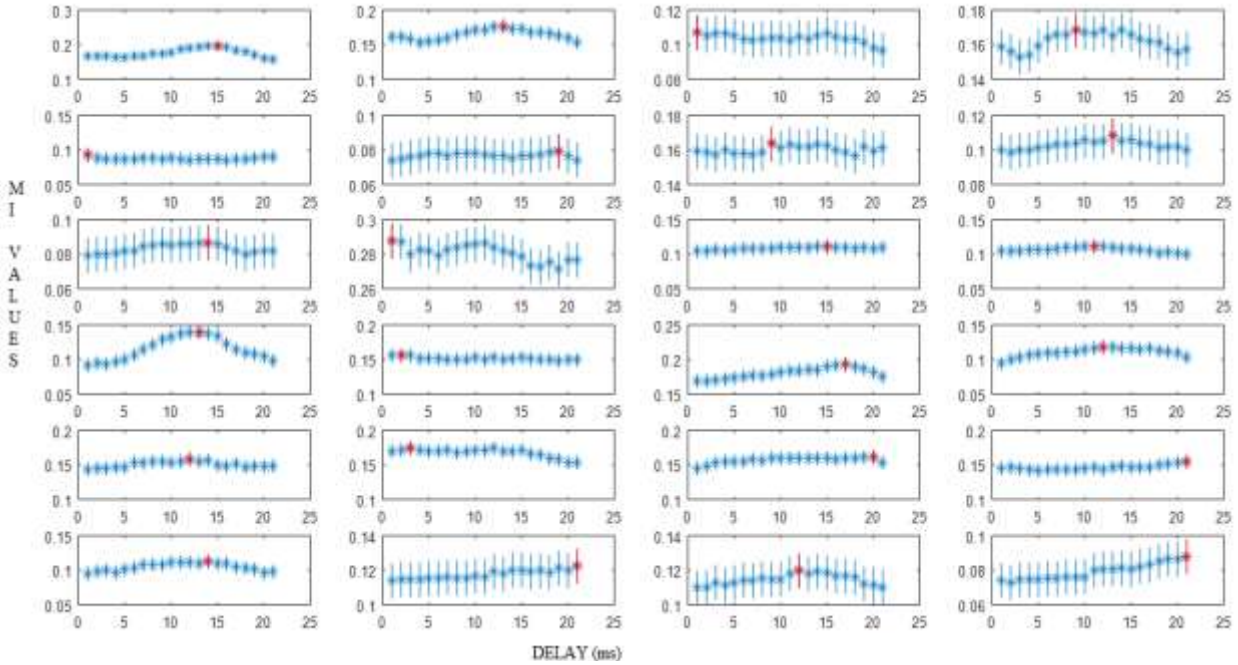

Figure 8: Plots of Delay vs English Language MI Values

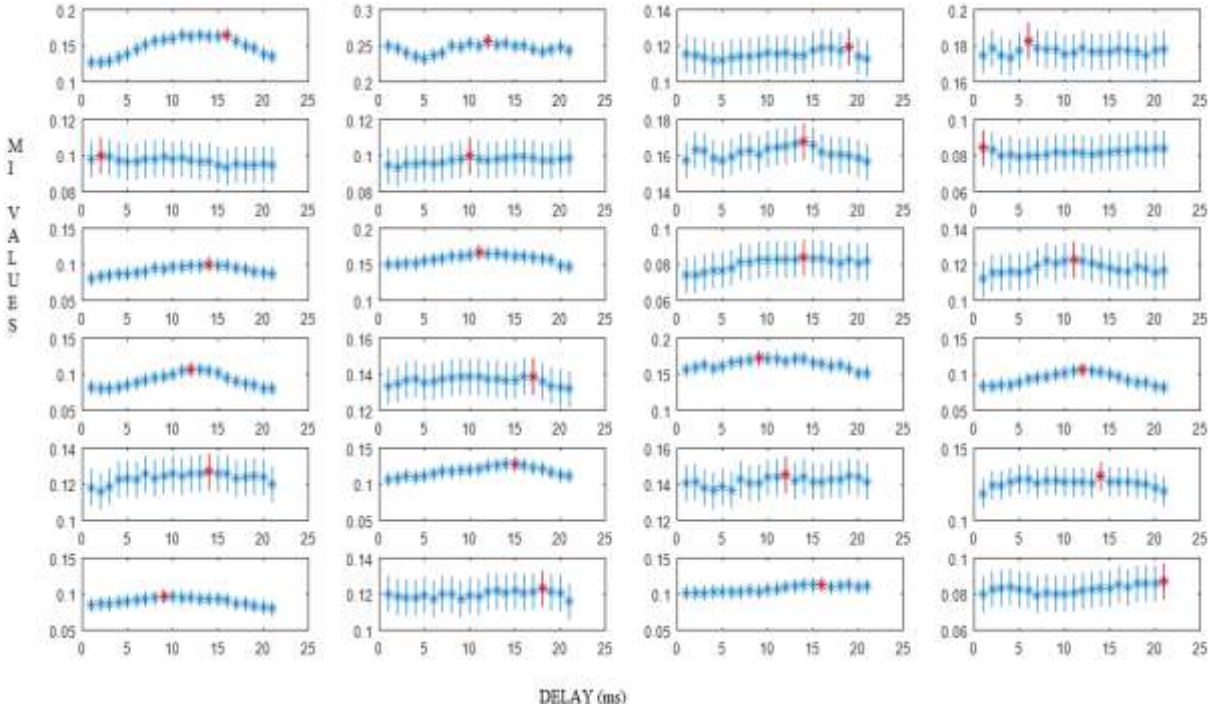

Figure 9: Plots of Delay vs Native Language MI Values