MULTIMODAL SIGNAL ANALYSIS OF PROSODY AND HAND MOTION: TEMPORAL CORRELATION OF SPEECH AND GESTURES

L. Valbonesi[†], R. Ansari[†], D. McNeill[‡], F. Quek[#], S. Duncan[‡], K. E. McCullough[‡], R. Bryll[#]
† University of Illinois at Chicago, Department of Electrical and Computer Engineering, Chicago, IL, USA
‡ University of Chicago, Department of Psychology, Chicago, IL, USA
#Wright State University, Department of Computer Science and Engineering, Dayton, OH, USA

ABSTRACT

This paper is concerned with the processing and analysis of signals pertaining to two communicative modalities, speech and gestures, and investigating the nature of their temporal relationship. Two hypotheses have been proposed about the activation of the gestural system during speech production: the *inhibitory hypothesis* and the *excitatory hypothesis*. The validation of either one of these hypotheses necessitates the processing of large amounts of data by experts. The work described here is an effort to develop valuable signal processing tools to facilitate the multi-modal analysis, at least partially, in an automated manner. In this work algorithms are developed to determine points of emphasis in each of the modalities of speech and gesture data using prosody and hand motion traces obtained from two experiments. The results agree with the excitatory hypothesis: if we compare the temporal locations of the speech focal points with the temporal locations of the gesture focal points, they co-occur in more than 90% of the locations.

1. INTRODUCTION

Speech and *gestures* are two different yet interdependent ways of expressing human thoughts; they clearly belong to different modalities of expression but they are linked on several levels and work together to present the same semantic idea units. The nature of the temporal relationship between the two modalities is of interest in a number of fields including psychology and linguistics. The need for computational tools for an automated analysis of gesture and speech data has led to a collaborative effort among researchers in the fields of signal processing, computer vision, and psychology, and some results of this effort are reported here.

Under the assumption of multiple interactions between the gestural and the verbal systems, one may ask if the nature of these interactions is *excitatory* or *inhibitory*. The excitatory hypothesis is a co-activation hypothesis according to which vocal and manual movements are triggered simultaneously from the computational level shared by the two systems; the activation of the gestural system during speech production is assumed inevitable, and gestures and speech are considered as sharing the same origin and then separating into two different output channels. Alternatively, the inhibitory hypothesis states that gestures and speech can be conceived as similar but rival activities, so competition would occur in the display of substitutable forms. A key difference between these two hypotheses, the excitatory and the inhibitory, is in the analysis of temporal relationship between gestures and speech: co-activation models assume close temporal associations in the production of speech and gestures, whereas competition models suggest a close relation between gestures and hesitation pauses.

The purpose of this paper is to analyze speech accompanying speaker's gestures, i.e. the spontaneous movement of his/her hands, and to investigate the type of relationship that exists between speech and gestures. We focus our attention on the temporal analysis of the data: whether the relation is excitatory wherein gestures bear a high correlation with voice events, or the relation is inhibitory wherein spontaneous movements exhibit high correlation with pauses and hesitations.

In order to establish the nature of gesture-speech relation, suitable experiments are carried out and recorded and data are collected, processed and analyzed. Results will be shown to establish that the time correlation between speech and gestures is of the first type, i.e. the analysis supports the theory of co-activation according to which speech and gestures are co-expressive.

2. DATA AND PRE-PROCESSING

In order to create data for analysis, two experiments are set up and recorded, and from the recordings the audio and video information are extracted.

In the first experiment two subjects are recruited to serve as speaker-interlocutor pair. A female speaker is asked to describe a strategy to surround and evacuate intelligent wombats from a village. This dialogue is of a long duration in which the two subjects take turns at speaking. In order to "transform" this set of data into a monologue, the audio segments corresponding to the interlocutor are eliminated manually. Figure 1 shows a frame extracted from the recording of the second experiment.

Figure 1 - Frame of Sue data

In this second experiment, a female speaker describes her house to an interlocutor. This experiment is mainly a monologue, as the interlocutor rarely interrupts the speaker and consequently this data set is easier to analyze, as it does not require separating speakers voice traces.
Among the features that characterize the voice data, we choose to examine prosodic information, specifically *amplitude* and *fundamental frequency F0*, which are extracted from the recorded data using Entropic x-waves software package.
For characterizing hand motion, the coordinates of the hand positions are extracted from the video, and from these the velocity components are computed.
Once speech and hand motion data have been pre-processed, they are first analyzed separately, significant features are extracted from them, and finally the results are combined together and compared in order to study the speech-gesture interaction.

3. SPEECH FOCAL POINT IDENTIFICATION

The method of analysis involves the determination of small segments, or "prosodic phrases" in the speech data, so that the investigation of significant events can be obtained in the analysis of each of the segments separately. The same procedure is applied to the study of the hand motion traces, which are segmented and analyzed section by section.
Using suitable adaptive thresholds, the speech data are partitioned into segments or speech units, which roughly corresponds to phrases in the discourse.
Once the data have been segmented, we look for significant features within each segment. The purpose of the analysis is the identification of the points in the speech where the greatest amount of information is concentrated. These points will be referred to as *focal points*.
In order to identify these points, it is necessary to extract some events that characterize the speech and locate all the possible points of emphasis. Using a combination of the extracted information a list of *rules* was identified for detecting the locations of focal points. Details of the procedure can be found in [9].
Five parameters are identified to analyze the audio information and they are extracted both from F0 and amplitude data.
The first candidate as key event is the *F0 peak*: within each segment the highest value of the fundamental frequency F0 is identified. However, the location of the F0 peak by itself is not sufficient to determine all the focal points, and it is necessary to integrate it with other key events. Following the studies and the results of a research group of the University of Bonn, an additional parameter is selected: the *minimum of the frequency gradient* within each segment. In fact, the German research group found out that in pre-focal position there is no down-stepping, while after a focal accent down-stepping is significant and characteristic. Consequently the location of the steepest fall in the F0 course is also included as a cue to a focal point. In our method the gradient of the fundamental frequency is computed and within each segment the largest negative peak is identified.
However, since the computation of the gradient could accentuate any residual errors that may not have been removed in F0 pre-processing, the largest negative gradient may be a "false alarm". Therefore not only is the first minimum taken into consideration, i.e. the highest negative value, but also the next largest local minimum. In fact, in some segments the negative peak of the frequency gradient has a value very close to the one of other points in the same segment. By considering two locations of high negative gradient in each segment as cues to candidate key events avoids losing information and false detections.
F0 information does not completely characterize the speech prosody, and amplitude information needs to be integrated with F0. The last cues extracted from the data are hence based on to the amplitude of the speaker's voice. Peaks in amplitude of a speaker's voice are used as cues to points of emphasis. Consequently, we identify the maximum value of the amplitude within each segment as our fourth cue to a focal point. Since the audio data were recorded in a noisy environment, an error that may not have been removed in pre-processing could mislead our analysis again. Moreover, in one segment it is possible that two local maxima may have very close high amplitude values and it is difficult to decide which of the two corresponds to a focal point. This problem is solved in the same way as was done when extracting of parameters from F0 gradient: not only is the first local maximum selected, but also the second local maximum within each segment.
In this way five cues are identified, three from the frequency data and two from the amplitude data. The next step is the combination of the information of these cues in order to establish *detection rules*.
At this point we consider the distinction between data sets as a *training set* and a *test set*. The first is used to determine the "rules" that lead to the determination of the focal accents, while the test set is used to check the validity of the rules.
Expert psychologists were asked to hand-mark the focal points in speech and their location was compared with the positions of the identified parameters: through this comparison it is possible to define a list of simultaneous occurrences of event cues that are necessary in order to detect a focal point.
The percentage of correct detection for the training set, i.e. the set of data used to identify the "rules" of the algorithm, is around 93%, while for the test set, i.e. the set used to check validity of the algorithm, the success rate is 77%. This success rate is an improvement over the performance of the algorithm proposed by researchers at the University of Bonn in which only the F0 gradient is used as cue and where the percentage of correct detection goes down to 69%. The improvement is attributed to the introduction of additional F0 and amplitude cues and to the adaptive segmentation method.

4. IDENTIFICATION OF GESTURE FOCAL POINTS AND CORRELATION WITH SPEECH

The first step in the temporal correlation analysis between speech and gestures is to examine the detected speech focal points and select those that co-occur with an appropriately identified event in the movement of the hands.
In order to do this, the x- and y- components of the velocity are examined for both hands and the velocity trace of each component is segmented to isolate sections of actual movement from those in which the hand rests.
Among the focal points detected in speech segments, we choose only those that occur in one of the motion segments, because these are the only ones that can be correlated to features in the hand movement.

Once speech focal points have been selected, we need to determine the location of the most significant features in the hand traces. In order to accomplish this, the local maxima and minima of the hand traces motion in each segment are located. In the extraction of these parameters, thresholds are used to avoid having detected locations too close to each other.
Now that the location of the speech focal points and the key events in hand motion traces have been identified, it is possible to compare the two sets of data in order to show how speech and gestures are correlated.
Two focal points, one in each of the two modalities of speech and gestures, are considered coincident if one focal point falls within a small window centered in the position of the other focal point. The window defines coincidence of events because the possibility that they occur exactly at the same instant is very low. While a precise instant for the speech focus can be identified, in the case of hand movement the "peak" of the gesture has wider time support and the identified position is one of many possible.
With this analysis we are able to identify the focal points in speech and gestures that are strongly correlated in time. However we are also interested in a weak correlation, i.e. the case in which one maxima or minima location falls within a wider window centered in a focal accent's position. So we need to take into account windows of different sizes.
This analysis was carried on for all the data of the hand traces, for both the right and the left hand, with correlation windows of three sizes, the second equal to twice the first and the third equal to three times the first, and the identified correlations are summarized in the following table

	RIGHT HAND		LEFT HAND		# foci = 22	
	r-	θ-	r-	θ-	tot	%
W1	2	2	1	3	7	31.81%
W2	5	4	4	6	14	63.63%
W3	7	8	7	6	16	72.72%

Table I - Results of strong and weak correlation

From this table one observes that the percentage of correlation between hand movement parameters and speech focal points increases when the width of the window increases, but this increase is obviously not proportional to the window size. Increasing the size of the window further beyond moderate size will not improve the results significantly.

5. COMPARISON WITH HAND-MARKED FOCAL POINTS

During the whole analysis for determining temporal correlation, the speech focal points that were used were only those that were automatically detected. The hand-marked focal points were not used. It should however be pointed out that some of the machine-detected locations do not coincide with the hand-marked ones. The following table summarizes the results on strong and week correlation, similar to that shown in Table I, but this time obtained using hand-marked foci instead that the detected ones.

	RIGHT HAND		LEFT HAND		# foci = 22	
	r-	θ-	r-	θ-	tot	%
W1	0	2	2	3	5	22.72%
W2	4	4	7	5	15	68.18%
W3	6	7	9	5	17	77.27%

Table II - Results of strong and weak correlation (hand-marked foci)

Comparing the results in the two tables, it is observed that the overall correlation rate is almost the same in the machine detected and hand-marked cases, confirming the fact that machine detection of focal points is satisfactory though the success rate is not 100%. This is due to the fact that 80 - 90% of the focal accent positions are correctly identified, while the remaining positions, even if they do not coincide with hand-marked ones, are very close to them and consequently a correlation can still be found. The windows centered at the detected focal points and those centered at the hand-marked focal points may partially overlap and the maximum or minimum of the hand position traces could fall within the overlapping section.
The percentage of success is almost the same using the two sets of focal points also because most of the incorrectly detected foci fall outside one of the segments containing hand motions. This suggests that it is easier to detect correctly a speech focal accent if it is accompanied by a gesture.
For the proposed analysis either sets of focal accents' positions can be used.

6. STROKE OF GESTURES

The percentage of the speech focal points that are correlated in time with a significant movement of the hands is not very high. The second part of the correlation analysis is to investigate the nature of the focal points that are not included in this classification. In order to do this we need to analyze further the nature of gestures.
So far attention has been given to maxima and minima of hand positions since, when the speaker is talking and she moves her hands at the same time. The moments at which the position of the speaker's hands reaches a maximum or a minimum are often the moments at which she points at something or shows something with her hands, and consequently they are focal points in the hand movement analysis.
An aspect that is difficult to capture in motion traces is the exact moment at which the meaning of the gesture is conveyed. This moment may be the "hold" of the gesture, i.e. the moment at which the hand is largely still while pointing at or showing something.
For this reason we have to distinguish between "hand movement" and "gesture"; where the former implies only a non-zero velocity in the hand traces, while the latter is associated with a specific meaning and with an idea conveyed with hands in order to complement the information carried by the speech. Some "gestures" can coincide with a movement, in cases where the information that is added with the hands to the speech is not

a static description but something dynamic and the hands need to move to convey the idea, but in general a movement is not a gesture and not every gesture implies movement.
In the analysis presented so far, the focal positions of the speech are correlated in time only with the hand movement. We are now interested in investigating if the remaining focal points that do not satisfy this type of correlation can be classified differently and be correlated to a gesture that is not necessarily accompanied with a movement.
In order to do this, we invoke the help of expert psychologists, who were able to mark not only the starting and the ending point of the gestures, but also the starting and the ending point of the *stroke*, which is the essence of the gesture, and may often coincide with the *hold* of the hand and not to its movement.
Table III summarizes the results obtained both for the machine-detected focal points and the hand-marked ones.
"Condition 1" refers to the case where the focal points have already been time-correlated with the hand movements, where the window of intermediate size is used as this choice resulted in the best compromise for detection.
"Condition 2" refers to the case where focal points occur within the duration of a gesture's stroke.

	Detected foci		Hand-marked foci	
	Number	%	Number	%
Cond1	14	63.63	15	68.18
Cond2	15	68.18	15	68.18
Cond1or2	20	90.91	21	95.45
None	2	9.09	1	4.55

Table III - Time correlation total results

From this table it is readily observed that again the results obtained using the machine-detected hand motion foci and the hand-marked gesture foci are very similar and that the focal positions that do not satisfy either one of the two conditions are just a small percentage of the total number. Moreover, in the case of the machine-detected foci, one of the two focal positions that belong to the last row (no condition satisfied) is the incorrect focus, and consequently disappears in the case of hand-marked foci.
Analyzing this table we can draw a very important conclusion: leaving aside a small percentage, all focal positions, both the detected and the hand-marked ones, satisfy at least one of the two conditions. This means that when we analyze the time correlation between speech focal accents and hand movement, some of the foci co-occur with significant points in the hand movement, and those that do not satisfy this condition are still correlated in time with hand features, only a different kind of feature. Earlier we investigated the movement of the hands. Now we consider the meaning of the gesture, and we find that the remaining focal points co-occur with the stroke of a gesture.

7. CONCLUSION

The purpose of this paper was to investigate the temporal relationship between prosodic focal accents and significant cues in hands traces. The analysis consists of determining the temporal correlation between two sets of focal point data, one obtained from the speech data and the other one obtained from a combination of information about hand motion and expert-marked gesture strokes.
It was found that gestures and speech are highly correlated in time. Two types of correlation are observed, a correlation with hand movements and a correlation with expert-marked gestures. With the proposed algorithm we can identify the positions of the local maxima and minima of the hand traces and correlate them with the focal positions; all the foci not included in this correlation identify with very high probability (more than 90%) the presence of the stroke of a gesture.
Therefore we can conclude that the results support the excitatory hypothesis of the interaction between speech and gestures production.

8. REFERENCES

[1] Ansari, R., Day, Y., Lou, J., McNeill, D., and Quek, F.: Representation of prosodic structure in speech using nonlinear methods. University of Illinois at Chicago and University of Chicago
[2] Elsner, A.: Prediction and perception of focal accents. Institut fur Kommunikationsforschung und Phonetik, University of Bonn, Germany, 1995
[3] Feyereisen, P., and de Lannoy, J. D.: Gestures and Speech: Psychological investigations. Cambridge University Press, Cambridge, 1991
[4] McNeill, D.: Hand and mind: what gestures reveal about thought. University of Chicago Press, Chicago, 1992
[5] Nobe, S.: Representational gestures, cognitive rhythms, and acoustic aspects of speech: a network/threshold model of gesture production. Ph.D.'s thesis, University of Chicago, Chicago, 1996
[6] Oppenheim, A. V., and Schafer, R. W.: Discrete-time signal processing. 1999
[7] Petzold, A.: Strategies for focal accent detection in spontaneous speech. Institut fur Kommunikationsforschung und Phonetik, University of Bonn, Germany, 1995.
[8] Quek, F., Bryll, R., McNeill, D., Kirbas, C., Arslan H., McCullough, K.E., Furuyama, N., and Ansari, R.: Gesture, Speech, and Gaze Cues for Discourse Segmentation.
[9] Valbonesi, L., Multimodal signal analysis of prosody and hand motion: temporal correlation of speech and gestures, M.S. thesis, University of Illinois at Chicago, 2001