Prosody Based Co-analysis of Deictic Gestures and Speech in Weather Narration Broadcast

Sanshzar Kettebekov, Mohammed Yeasin, Nils Krahnstoever, Rajeev Sharma

Department of Computer Science and Engineering
Pennsylvania State University
220 Pond Laboratory
University Park, PA 16802,USA
[kettebek; yeasin; krahnsto; rsharma]@cse.psu.edu

Abstract

Although speech and gesture recognition has been studied extensively all the successful attempts of combining them in the unified framework were semantically motivated, e.g., keyword co-occurrence. Such formulations inherited the complexity of natural language processing. This paper presents a statistical approach that uses physiological phenomenon of gesture and speech production process for improving accuracy of automatic segmentation of continuous deictic gestures. The prosodic features from the speech signal were co-analyzed with the visual signal to create a statistical model of co-occurrence with particular kinematical phases of gestures. Results indicated that the above co-analysis improves continuous gesture recognition. The efficacy of the proposed approach was demonstrated on a large database collected from the weather channel broadcast. This formulation opens new avenues for bottom-up frameworks of multimodal integration.

1. Introduction

In combination, gesture and speech constitute the most important modalities in human-to-human communication. People use large variety of gestures either to convey what cannot always be expressed using speech only or to add expressiveness to the communication. Motivated by this, there has been a considerable interest in incorporating both gestures and speech as the means for Human-Computer Interaction (HCI).

To date, speech and gesture recognition have been studied extensively but most of the attempts at combining them in an interface were in the form of a predefined signs and controlled syntax such as "*put <point> that <point> there*", e.g., (Bolt, 1980). Part of the reason for the slow progress in multimodal HCI is the lack of available sensing technology that would allow non-invasive acquisition of natural behavior. However, the availability of abundant processing power has contributed to making computer vision based continuous gesture recognition in real time to allow the inclusion of natural gesticulation in a multimodal interface (Kettebekov and Sharma, 2001, Pavlovic et al., 1997, Sharma et al., 2000).

State of the art in *continuous gesture recognition* is far from meeting the requirements of a multimodal HCI due to poor recognition rates. Co-analysis of visual gesture and speech signals provide an attractive prospect of improving continuous gesture recognition. However, lack of fundamental understanding of speech/gesture production mechanism restricted implementation of the multimodal integration at the semantic level, e.g. (Kettebekov and Sharma, 2001, Oviatt, 1996, Sharma et al., 2000). Previously, we showed somewhat significant improvement in co-verbal gesture recognition when those were co-analyzed with keywords (Sharma et al., 2000). However, the implications of using a top-down approach has augmented challenges with those of natural language and gesture interpretation and made automatic processing challenging.

The goal of the present work is to investigate co-occurrence of speech and gesture as applied to continuous gesture recognition from a bottom-up perspective. Instead of keywords, we employ a set of prosodic features from speech that correlate with deictic gestures. We address the general problem in multimodal HCI research, e.g., availability of valid data, by using narration sequences from the weather channel TV broadcast. The paper is organized as follows. First, a brief overview of the types of gestures that occur in the analysis domain is presented. The synchronization hierarchy of gestures and speech is also reviewed. In section 3 we discuss a computational framework for continuous gesture acquisition using a segmental approach. Section 4 presents a statistical method for correlating visual and speech signals. There, acoustically prominent segments are detected and aligned with segmented gesture phases. Finally, results are discussed within the framework for continuous gesture recognition.

2. Co-verbal Gesticulation for HCI

McNeill (1992) distinguishes four major types of gestures by their relationship to the speech. *Deictic* gestures are used to direct a listener's attention to a physical reference in course of a conversation. These gestures, mostly limited to the pointing, were found to be co-verbal, cf. (McNeill, 1992). From our previous studies, in the computerized map domain (***i*MAP,** see Figure 1) (Kettebekov and Sharma, 2000), over 93% of deictic gestures were observed to co-occur with spoken nouns, pronouns, and spatial adverbials.

Iconic and *metaphoric* gestures are associated with abstract ideas, mostly peculiar to subjective notions of an individual. *Beats* serve as gestural marks of speech pace. In the weather channel broadcast the last three categories roughly constitute 20% of all the gestures exhibited by the narrators. We limit our current study to the *deictic* gestures for a couple of reasons. First, they are they are more suitable for manipulation of a large display, which becomes more common for HCI applications. Second, this

type of gestures exhibits relatively close coupling with speech.

2.1. Gesture and Speech Production

The issue of how gestures and speech relate in time is critical for understanding the system that includes gesture and speech as part of a multimodal expression. McNeill (1992) distinguishes three levels of speech and gesture synchronization: semantic, phonological, and pragmatic. The pragmatic level synchrony is common for metaphoric and iconic gestures and therefore is beyond the scope of the present work.

Semantic synchrony rule states that speech and gestures cover the same idea unit supplying complementary information when they occur synchronously. The current state of HCI research provides partial evidence to this proposition. Previous co-occurrence analysis of weather narration (Sharma et al., 2000) revealed that approximately 85% of the time when any meaningful gestures are made, it is accompanied by a spoken keyword mostly temporally aligned during and after the gesture. Similar findings were shown in the pen-voice studies (Oviatt et al., 1997). The implication of the semantic level synchronization rule was successfully applied at the keyword level co-occurrence in the previous weather narration study (Sharma et al., 2000).

At the phonological level, Kendon (1990) found that different levels of movement hierarchy are functionally distinct in that they synchronize with different levels of prosodic structuring of the discourse in speech. For example, the peaking effort in a gesture was found to precede or end at the phonological peak syllable (Kendon, 1980). These findings imply a necessity for viewing a continuous hand movement as a sequence of kinematically different segments of gestures. This approach is reflected in the next section. Issue of using the phonological peak syllables is associated with the complexity of the nature of the tonal correlates, e.g., pitch of the voice. Pitch accent, which can be specified as low or high, is thought to reflect a phonological structure in addition to the tonal discourse, cf. (Beckman et al., 1992). We address this issue by proposing a set of correlate point features in the pitch contour that can be associated with the points on the velocity and acceleration contours of the moving hand (section 4).

3. Gesture Acquisition

Building human computer interfaces that can use gestures involves challenges that range from low-level signal processing to high-level interpretation. A wide variety of methods had been introduced to create gesture driven interfaces. With the advances in technology there has been a growing interest in using vision-based methods (Pavlovic et al., 1997). The advantage of these is in their non-invasive nature. The idea of a natural interface comes from striving to make HCI as close as communicating in ways we are accustomed to. Vision-based implementation therefore can be very useful for a natural interface.

One could expect that the meaning encoded in multimodal communication is somehow distributed across speech and gesture modalities. A number of recent implementations used predefined gesture syntax, e.g., (Oviatt, 1996). A user is confined to the predefined gestures for spatial browsing and information querying. As a result, a rigid syntax is artificially imposed. Therefore the intent of making interaction natural is defeated. However, with imprecise recognition of non-predefined gestures, it may be harder to argue for replacing more precise HCI devices, e.g., electronic pen with fixed predefined functions.

The key problem in building such interface, e.g., using statistical techniques, is the lack of existing *natural* multimodal data. Studies from human-to-human communication do not automatically transfer over to HCI due to artificially imposed paradigms. This controversy leads to a "chicken-and-egg" problem.

While the use of the weather narration domain as a bootstrapping analysis offers virtually unlimited bimodal data it can be assumed as a reasonable simplification of an HCI domain. In the series of the previous studies we employed the weather narration broadcast analysis (Sharma et al., 2000) to bootstrap ***i*MAP** framework (Figure 1) (Kettebekov and Sharma, 2001). It showed that the gesticulative acts used in both domain have similar kinematical structure as well as gesture and keyword co-occurrence patterns. However, the key aspect for choosing the weather domain for the current study is in a possibility of applying simple processing techniques for extraction of prosodic information from uninterrupted narration.

Figure 1. ***i*MAP** testbed in the context of a computerized map. The cursor is shown within the circle.

Over 60 minutes of the selected weather narration data was used in the analysis. The video sequences contained uninterrupted monologue of 1-2 minutes in length. The subject pool was presented by 5 men and 3 women.

3.1. Kinematics of Continuous Gestures

A continuous hand gesture consists of a series of qualitatively different kinematical phases such as movement to a position, hold, and transitional movement. We adopt Kendon's framework (Kendon, 1990) by organizing these into a hierarchical structure. He proposed a notion of gestural unit (*phrase*) that starts at the moment when a limb is lifted away from the body and ends when the limb moves back to the resting position. The *stroke* is distinguished by a peaking effort and it is thought to constitute the meaning of a gesture (Kendon, 1990). After extensive analysis of gestures in weather narration and *i*MAP (Kettebekov and Sharma, 2001, Sharma et al., 2000) we consider following strokes: *contour*, *point*, and *circle*.

Kita (1997) suggested that a *post-stroke hold* was a way to temporally extend a single movement stroke so that the *stroke* and *post-stroke hold* together will

synchronize with the co-expressive portion of the speech. It is thought that a *pre-stroke hold* is a period in which gesture waits for speech to establish cohesion so that the stroke co-occurs with the co-expressive portion of the speech. Therefore, in addition to our previous definitions we also include *hold* as a functional primitive.

3.2. Continuous Gesture Segmentation

Sixty minutes of weather domain gesture data for training and testing was collected from broadcast video using a semi-automatic gesture analysis tool (GAT) (see Figure 2). The tool provides a convenient user interface for rapid and consistent collection of positional data and a easily configurable set of pattern classification tools. GAT is integrated with PRAAT software for phonetics research (Boersma and Weenink, 2002) for speech processing and visualization.

Figure 2. Gesture analysis tool (GAT) interface

The task of positional data ground truthing involves initialization the head and hand tracking algorithms (described in 2.3.1) at the beginning of each video sequence and in the events of self-occlusions of the hands.

3.2.1. Motion Tracking

The algorithm for visual tracking of the head and hands is based on motion and skin-color cues that are fused in a probabilistic framework. For each frame and each tracked body part, a number of candidate body part locations are generated within a window defined by the location of the body part in the previous frame and the current estimate of the predicted motion. The true trajectories of the body parts are defined as the most probable paths through time connecting candidate body part locations. The Viterbi algorithm is used to efficiently determine this path over time. This approach effectively models the hand and head regions as skin-colored moving blobs (Figure 3).

3.2.2. Kinematical Analysis

To model the gestures, both spatial and temporal characteristics of the hand gestures (phonemes) were considered. The time series patterns of gesture phases can be viewed as a combination of ballistic and guided motion of the hand reflected on the skewedness of the velocity profile. In the current study, a gesture phoneme is defined as a stochastic process of 2D positional and time differential parameters of the hand and head over a suitably defined time interval.

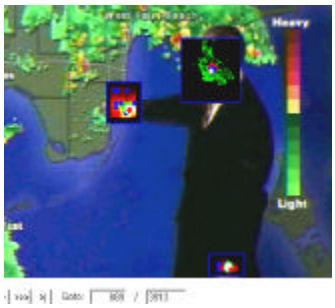

Figure 3. Semi-automatic ground truthing process employing a tracking algorithm;

A Hidden Markov Model (HMM) framework was employed for continuous gesture recognition, as described in (Sharma et al., 2000). The total of 446 phoneme examples extracted from the segmented training video footage were used for HMM training. The results of the continuous gesture recognition showed that only 74.2 % of 1876 were classified correctly. Further analysis indicated that phoneme pairs of preparation-pointing and contour-retraction constitute most of the substitution errors. This type of error, which can be attributed to the similarity of the velocity profiles, was accounted for the total of 33% of all the errors. The deletion[1] errors were mostly due a relatively small displacement of the hand during a pointing gesture. Those constituted approximately 58% of all the errors.

Although purpose of this work was not to introduce a robust algorithm with a high recognition rate there is an inherent limitation with the current acquisition method. I.e., 2D projected motion data can potentially introduce spurious variabilities that can have a detrimental effect on the recognition rate. The gesture model is based on the observed end-effector motion of the hands and the motion of the head projected into the camera plane and is only and indirect measurement of the true body

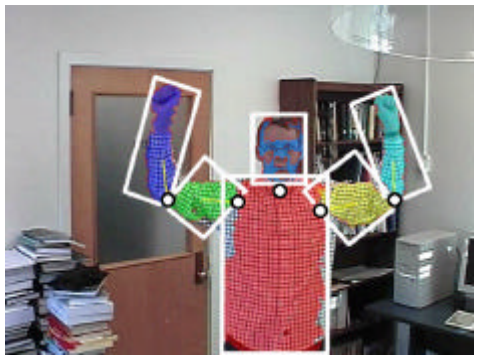

Figure 4. Model based tracking for future extraction of direct kinematical gesture parameters.

[1] Deletion type of errors occur when a gesture phoneme is recognized as part of another adjacent gesture.

kinematics. This observation model can hence introduces distortions and additional spurious variabilities that complicate the differentiation between gestures. Current work in progress, cf. (Krahnstoever et al., 2002), has the goal of visually extracting the true 3D kinematical parameters such as body pose and angles of the shoulder and arm joints (see Figure 4).

4. Prosody Based Co-analysis

Both psycholinguistic, e.g., (McNeill, 1992), and HCI, e.g., ***i*MAP** (Kettebekov and Sharma, 2000), studies suggest that deictic gestures do not exhibit one-to-one mapping of form to meaning. Previously, we showed that the semantic categories of strokes (derived through the set of keywords), not the gesture phonemes, correlate with the temporal alignment of keywords, cf. (Kettebekov and Sharma, 2000). This work distinguishes two types of gestures: referring to a static point on the map and to a moving object (i.e., moving precipitation front). Due to the homogeneity of the context and trained narrators in the weather domain we can statistically assume (mismatch <2%) that pointing gesture is the most likely to refer to the static and contour stroke to the moving objects. Therefore, for simplicity we will use *contour* and *point* definitions.

The purpose of the current analysis is to establish a framework by identifying correlate features in visual and acoustic signals. First we will separate acoustically prominent segments. A segment is defined as a voiced interval on the pitch contour that phonologically can vary from a single phone/foot[2] to intonational phrase units, see (Beckman, 1996) for details. Then we will analyze alignment of the prominent segment with the gesture phonemes. This framework was implemented in GAT.

4.1. Detecting Prosodically Prominent Segments

Pitch accent association in English underlines the discourse-related notion of focus of information. Fundamental frequency (F_0) is the correlate of pitch defined as the time between two successive glottis closures (Hess, 1983). We employed PRAAT software to extract F_0 contour, as described in (Boersma, 1993).

Prominent segments were defined as segments which were relatively accentuated (or perceived as such) from the rest of the monologue. We considered combination of the pitch accent and the pause before each voiced segment to detect abnormalities in spoken discourse. Maximum and minimum of F_0 contour represent features for high pitch and low pitch accents. Maximum gradient of the pitch slope was also considered. A statistical model of prosodic discourse for each narration sequence was created (Figure 5), see (Kettebekov et al., 2002) for details.

To find an appropriate level of threshold to detect prominent segments we employed a bootstrapping technique involving a perceptual study. A control sample set for every narrator was labeled by 3 naïve coders for auditory prominence. The coders had access only to the wave form of speech signal. The task was to identify at least one acoustically prominent sound within the window of 3 seconds. The moving window approach was considered to account for abnormally elongated pauses in the spoken discourse. Allowing 2% of misses, the threshold was experimentally set for each narrator (Figure 5). If a segment appeared to pass the threshold value it was considered for co-occurrence analysis with the associated gesture.

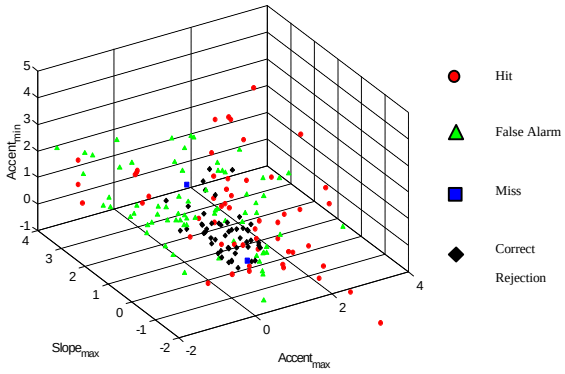

Figure 5. A sample distribution of auditory prominence for a female narrator with the decision boundary from the perceptual study.

4.2. Co-occurrence Models

A statistical model of the temporal alignment of active hand velocity and a set of features of the prominent pitch segments was created for every gesture phoneme class (Figure 6). The features on the pitch profile included max, min, beginning, and max of derivative of F_0, see (Kettebekov et al., 2002) for details. Present formulation

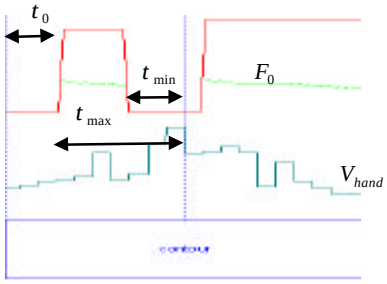

Figure 6. A set of features used for co-occurrence modeling of the hand velocity (V_{hand}) and a pitch (F_0) segment. Red contour represents prominence level of corresponding segments

accounts for the two levels of possible prosodic co-occurrence: discourse and phonological. The onset between a gesture and the beginning of a prominent segment is to model discourse cohesion (pauses). The onset of the peaks in the F_0 and peaks in the velocity profile of the hand addresses phonological level synchronization. All of 446 phonemes that have been used for training gesture phonemes were utilized for training of the co-occurrence models. Analysis of the resulted models indicated that there was no significant difference between *retraction* and *preparation* phases. Peaks of *contour* strokes tend closely to coincide with the peaks of the pitch

[2] Foot is a phonological unit that has a "heavy" syllable followed by a "light" syllable(s).

segments. *Pointing* appeared to be quite silent, however, most of the segments were aligned with the beginning of the *post-stroke hold* interval.

Figure 7 summarizes findings of the co-analysis framework. At the first level we separate co-verbally meaningful gestures (*strokes*) from *auxiliary* phonemes that included *preparation* and *retraction* phases. Also, we exclude strokes that are re-articulate previous gestures such as a stroke can be followed by the identical stroke where the second movement does not have associated speech segment. At the second level co-verbal strokes can be further classified according to their deixis, cf. (Kettebekov and Sharma, 2001). As it was noted before, in the context of the weather narration we can statistically consider those to be represented by *point* and *contour* phonemes without further definitions. *Preparation* and *retraction* phases were eventually collapsed into the same category and were not differentiated.

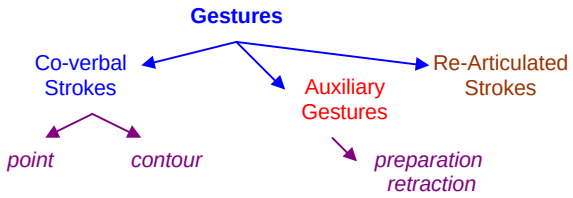

Figure 7. Prosodic co-analysis framework

The co-analysis models for co-verbal strokes were merged with the beginning of the post stroke-*hold* phases for classification purposes. Such redefinition of the co-verbal strokes for the purpose of co-analysis was motivated by the results associated with the *pointing* strokes and it was included into the computational framework.

4.3. Continuous Gesture Recognition with Co-occurrence Models

We employed Bayesian formulation to fuse the gesture framework and the co-occurrence models at the decision level, see (Kettebekov et al., 2002). The resulted segmentation showed significant improvement in the overall performance with the correct recognition of 81.8% (versus 72.4%). Subsequently, there was a significant reduction of deletion (8.6% versus 16.1%) and substitution errors (5.8% versus 9.2%). The deletion type of errors were minimized due to the inclusion of small point gestures, which are quite salient when correlated with prominent acoustic features. Figure 8 shows example of elimination of a deletion error after applying co-analysis. White trace on the figure illustrates visually negligible hand movement trajectory. Improvement of substitution errors can be attributed to the differentiation between the auxiliary gesture phases and the strokes in the co-occurrence analysis.

5. Conclusions

We presented an alternative approach for combining gesture and speech signals from the bottom-up perspective. Unlike commonly controlled gesture

a) b)

Figure 8. Example of deletion error using: a) visual-only signal resulted in *hold* gesture; b) with co-occurrence model *point* was recognized as a part of preceding *hold* (case a.);

recognition domains, we address this problem in the weather broadcast domain, which can be characterized by relatively unrestricted narration. Such formulation is more favorable for automated recognition of continuous deictic gestures then the semantic based (keyword co-occurrence). The current results demonstrate the concept of improving recognition of co-verbal gestures when combined with the prosodic features in speech. This is a first attempt which requires further improvement. The issues of portability to an HCI setting, e.g., ***i*MAP** framework, are currently under investigation.

Applicability of the current formulation for the other types of gestures is probably possible if the segmental approach is considered for the gesture acquisition. In a domain with more spontaneous behavior, e.g., in a dialogue (e.g., ***i*MAP**) (versus monologue as presented in the present work) the methodology of prosodically prominent feature extraction is more complex. It would require acquisition of an improved kinematical model (see section 3.2.2.) that considers additional visual cues such as turn of head (direction of the gaze), and etc.

6. Acknowledgements

The financial support of this work in part by the National Science Foundation CAREER Grant IIS-97-33644 and NSF IIS-0081935 is gratefully acknowledged. We thank Ryan Poore for his help with the data processing and implementation.

7. References

Beckman, M. E., Dejong, K., Jun, S. A., and Lee, S. H. 1992. The Interaction of Coarticulation and Prosody in Sound Change [JAN-JUN]. *Language and Speech* 35:45-58.

Beckman, M. E. 1996. The parsing of prosody [FEB-APR]. *Language and Cognitive Processes* 11:17-67.

Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Paper presented at *Institute of Phonetic Sciences of the University of Amsterdam*.

Boersma, P., and Weenink, D. 2002. PRAAT. Amsterdam, NL: Institute of Phonetic Sciences. University of Amsterdam, NL.

Bolt, R.A. 1980. Put-that-there: Voice and gesture at the graphic interface. *In SIGGRAPH-Computer Graphics*.

Hess, W. 1983. Pitch Determination of Speech Signals. In *Springer Series of Information Sciences*. Berlin: Springer-Verlag.

Kendon, A. 1980. Gesticulation and speech: Two aspects of the process of the utterance. In *The relation between verbal and non-verbal communication*, ed. M.R. Key, 207-227. Hague: Mouton.

Kendon, A. 1990. *Conducting Interaction*: Cambridge: Cambridge University Press.

Kettebekov, S., and Sharma, R. 2000. Understanding gestures in multimodal human computer interaction. *International Journal on Artificial Intelligence Tools* 9:205-224.

Kettebekov, S., and Sharma, R. 2001. Toward Natural Gesture/Speech Control of a Large Display. In *Engineering for Human Computer Interaction*, eds. M.R. Little and L. Nigay, 133-146. Berlin Heidelberg New York: Springer Verlag.

Kettebekov, S., Yeasin, M., and Sharma, R. 2002. Prosody based co-analysis for continuous recognition of co-verbal gestures, submitted to ICME'02.

Kita, S., Gijn, I.V., and Hulst, H.V. 1997. Movement phases in signs and co-speech gestures, and their transcription by human coders. Paper presented at *Intl. Gesture Workshop*.

Krahnstoever, N., Yeasin, M., and Sharma, R. 2002. Automatic Acquisition and Initialization of Articulated Models. Paper presented at *To appear in Machine Vision and Applications.*

McNeill, D. 1992. *Hand and Mind*: The University of Chicago Press, Chicago IL.

Oviatt, S. 1996. Multimodal interfaces for dynamic interactive maps. Paper presented at *Conference on Human Factors in Computing Systems (CHI'96).*

Oviatt, S., Angeli, A. De, and Kuhn, K. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. Paper presented at *Conference on Human Factors in Computing Systems (CHI'97).*

Pavlovic, V. I., Sharma, R., and Huang, T. S. 1997. Visual interpretation of hand gestures for human- computer interaction: A review. *IEEE Trans on Pattern Analysis and Machine Intelligence* 19:677-695.

Sharma, R., Cai, J., Chakravarthy, S., Poddar, I., and Sethi, Y. 2000. Exploiting Speech/Gesture Co-occurrence for Improving Continuous Gesture Recognition in Weather Narration. Paper presented at *International Conference on Face and Gesture Recognition (FG'2000)*, Grenoble, France.