COMBINED GESTURE-SPEECH ANALYSIS AND SPEECH DRIVEN GESTURE SYNTHESIS

M.E. Sargin[1],O. Aran[3],A. Karpov[4],F. Ofli[1],Y. Yasinnik[2],S. Wilson[5],E. Erzin[1],Y. Yemez[1] and A.M. Tekalp[1]

[1] Koç University, Turkey ({msargin,fofli,eerzin,yyemez,mtekalp}@ku.edu.tr),
[2]MIT, USA, (yelena@mit.edu), [3]Bogazici University, Turkey (aran@cmpe.boun.edu.tr),
[4]SPIIRAS, Russia (karpov@iias.spb.su), [5]University College Dublin, Ireland (stephen.m.wilson@ucd.ie)

ABSTRACT

Multimodal speech and speaker modeling and recognition are widely accepted as vital aspects of state of the art human-machine interaction systems. While correlations between speech and lip motion as well as speech and facial expressions are widely studied, relatively little work has been done to investigate the correlations between speech and gesture.
Detection and modeling of head, hand and arm gestures of a speaker have been studied extensively and these gestures were shown to carry linguistic information. A typical example is the head gesture while saying "yes/no". In this study, correlation between gestures and speech is investigated. In speech signal analysis, keyword spotting and prosodic accent event detection has been performed. In gesture analysis, hand positions and parameters of global head motion are used as features. The detection of gestures is based on discrete predesignated symbol sets, which are manually labeled during the training phase. The gesture-speech correlation is modelled by examining the co-occurring speech and gesture patterns. This correlation can be used to fuse gesture and speech modalities for edutainment applications (i.e. video games, 3-D animations) where natural gestures of talking avatars is animated from speech. A speech driven gesture animation example has been implemented for demonstration.

1. INTRODUCTION

The role of vision in human speech perception and processing is multi-faceted. The complementary nature of the information provided by the combinations of visual speech gestures used in phoneme production (such as lip and tongue movements) has been well researched and shown to be instinctively combined by listeners with acoustic and phonological information to correctly identify what is being said. In fact, speech perception is highly dependent on the visual gestures like lip movements. McGurk showed in [1] that perception of a speech sound is affected by a non matching lip/mouth movement. His experiments showed that when subject utters /b/ but lip movements corresponding to /g/ is seen, /d/ is perceived.

Non facial modalities like arm and head gestures are not tightly coupled with speech as the case of facial modalities but they are linked to present the same semantic idea units. The correlation between these two modalities and the analysis methodology is of interest in a number of fields including psychology and linguistics [2, 3].

The origin of the correlation between gestural and acoustical modalities are based on two hypothesis named as excitatory and inhibitory. The excitatory hypothesis states that vocal and gestural events are co-activated by a parallel processing system. In this case, human thoughts are processed by the cerebellum, then the motor neurons associated with vocal and muscular activation are stimulated simultaneously. The latter one called inhibitory hypothesis, in which vocal and gestural events are using the resources of single processing system. In this case, events of each modality co-occur with the counter modalities pauses. Detailed information about these hypothesis can be found in [4].

When audio and visual modalities are highly correlated, one modality's events can be used to predict the complementary modality's events. The more the modalities are correlated, the more reliable will be the predictions. This boils down the problem of prediction to selection of the most correlated features or events which was also studied in our previous work [5]. Estimation of correlated gestural events given speech, can be used to provide natural gesture patterns for the task of artificial gesture synthesis. Artificial gesture synthesis given speech is used in edutainment applications, where humans expect interactive conversations that animated person's speech is aided and complemented by other sensory modalities, including expression, gaze, gesture, grasp, signing, emotion, and factors beyond the textual equivalent of speech [6].

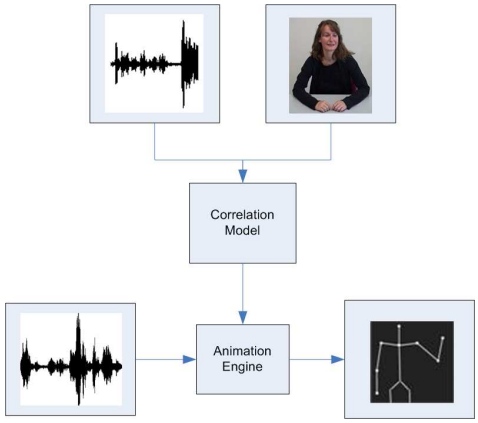


Fig. 1. Proposed System Overview

2. MOTIVATIONS AND INITIAL OBSERVATIONS

A primary motivation of the work presented here was to identify natural classes of gestures that conveyed real linguistic meaning, that is, to identify gestures or groups of gestural patterns that could be

This research was partly funded by SIMILAR, the European Network of Excellence on Multimodal Interfaces, during the eNTERFACE05 Workshop in Mons, Belgium.

clearly correlated with information conveyed in the speech signal. Once identified, these classes would be used to synthesize "natural" gesture patterns using an animated stick figure, given an input speech signal. The work detailed below is intended to be a preliminary investigation and so is restricted to analyzing gestures in a limited but gesture-rich database. An audio-visual database was prepared, comprising 25 minutes of video data. A single native speaker of Canadian English was recorded, providing directions to a number of known destinations in response to assisted questions.

An initial informal analysis was carried out, in order to ascertain potential lexical candidates that had recurring patterns of significant gestures. This involved close viewing of the video data by two investigators with experience of gesture identification and speech annotation. Initial observation highlighted three candidates, "left", "right", and "straight", for further study. The three lexical items were chosen as they showed a high co-occurrence with periods of significant manual gestural activity. Furthermore, they had a high distribution throughout the database indicating a potentially rich source of data for analysis. It was informally noted that 28 instances of the candidate "left", appeared to be accompanied by some sort of gesture. Similarly 31 occurrences of "right" had accompanying gestures, while "straight" had associated gestures 32 times throughout the database. Other candidate words included "across", "no", and "down", but these were dismissed as having too few gesture-marked occurrences (8, 8, and 6 respectively).

3. SPEECH EVENT DETECTION

In this section we investigate automatic spotting of semantic and prosodic events. MFCC coefficients are used in the extraction of semantic events. Pitch, formant frequencies and intensity values are considered as features for prosodic event spotting.

3.1. Feature Extraction

Semantic features are represented with 13 MFCC, 13 delta coefficients and 13 acceleration coefficients. MFCC coefficients are calculated over 25 ms windows for each 10 ms frame, where the resulting speech feature rate is 100 fps.

The nature of prosodic speech events are well described with the temporal variations of intensity, pitch and formant frequencies. Therefore in this study, these three features, pitch (p), intensity (i) and the first three formant frequencies (f), are considered as the potential prosodic features.

The pitch contour is extracted from the speech signal using the autocorrelation method as described in [7]. The squared sound intensities are weighted with 32 ms Kaiser-20 window, and the speech signal intensity is calculated as the sum of these weighted samples. The 32 ms window is shifted by 10 ms for each frame such that, the intensity values have a 100 Hz frame rate. An LP filter is calculated over 50 ms Hamming window for each 10 ms frame. The first three formant frequencies are extracted by tracking the peaks of the LP magnitude spectra.

3.2. Recognition of Semantic Events

Semantic events are considered as keywords uttered in speech. Frequently used words in the speech database are picked as the keywords, which are *left, right* and *straight*. In this section we present an HMM based automatic keyword spotter.

Keyword spotting task is performed using the methodology described in [6]. Manually labelled portion (80%) of the entire database is used for training and the remaining part is used for testing. Each keyword in the training database has at least 30 repetitions. Five HMM models are used for three keywords (*left, right* and *straight*) and two non-keywords (*silence* and *garbage*). The *silence* model is defined as segments corresponding to background noise. The *garbage* model corresponds to any non-keyword utterances. Continuous observation densities are modelled using varying number of Gaussian mixtures and the optimum number of Gaussian mixtures are selected considering the keyword spotting accuracy and false alarm rate.

In the experiments, we obtained 94.3% (33 out of 35) true detection and 1.6% (10 out of 600) false alarm rate for keyword spotting.

Feature Set	$RRate$	$1 - FAR$
$[\Delta f + p + i]$	0.7810	0.6668
$[f + p + i]$	0.7140	0.6724
$[p + i]$	0.7479	0.6966

Table 1. Accent detection performance

3.3. Recognition of Prosodic Events

Prosodic events that are correlated with speech signal are defined as pitch accents. Three different sets of features are used in proposed accent detector scheme: $[p+i]$, $[f+p+i]$ and $[\Delta f+p+i]$. Here, the $+$ operator represents concatenation of features.

In order to establish an initial working hypothesis, an experienced ToBI labeller marked training portion of speech for pitch accents and phrase boundaries. Within the training set, 122 pitch accents are identified. Manually labelled speech sequence is partitioned as *accent*, *non-pitch* and *non-accent*. The *accent* and *non-accent* labels correspond to syllables that are accented and non-accented, respectively. The *non-pitch* label is used for the syllables that pitch can not be extracted. Three left-to-right HMM structures with 6 states and 5 mixtures are used to model these three events.

The system is trained using the features corresponding to 80% portion of manually labelled pitch accents. Remaining 20% portion is used for testing. The position of testing portion is shifted 4 times with 4 new trainings to cover all labelled data in the testing. Table 1 presents the accent recognition rate $RRate$ and false alarm rate FAR. The use of $[\Delta f + p + i]$ feature set yields optimum performance. The $1 - FAR$ is maximized with the $[p + i]$ features, however, considering the trade off between false alarms and the recognized accents, the $[\Delta f + p + i]$ feature set still yields better performance than the other two.

4. GESTURAL EVENT DETECTION

In this section, HMM-based hand and head gesture recognition system is presented. The usage of HMMs for gesture recognition is motivated by the similarities between gesture and speech. Yang *et. al.,* summarizes these similarities in [8]. HMMs have been applied to the speech recognition problem to partition every word into a finite number of speech elements called phonemes. Similarly, the usage of HMMs for gesture recognition allows us to take the advantage of partitioning each gesture into tactemes where hidden states are associated with them. Therefore, the number of states for each HMM associated with a specific gesture should be selected according to the number of tactemes corresponding to that gesture.

4.1. Feature Extraction

In this study, head gesture features are chosen as the 8 global quadratic head motion parameters calculated over the face region. The extraction of head gesture features are described in detail in [6].

A hand gesture is represented with a single numeric feature which is the center of mass position of each hand. The center of mass is tracked over video using a Kalman filter where the states correspond to position and velocity [6].

4.2. Hand Gesture Recognition

Based on the initial observation of directional words and gestures that were salient in the video, three hand gestures were selected. Right and Left Gestures: The right or left hand turns to make a 90° with the arm, pointing to the right for right gesture, or to the left for left gesture. Straight Gesture: The subject starts with her hands in parallel, palms facing each other, fingers directed up, and moves the hands away from the body by extending her elbows. The finishing position is with hands parallel, palms facing each other, fingers pointing away from the subject's body.

An isolated hand gesture recognition scheme using continuous density HMMs with 5 states is employed. The performances obtained on the test video for left, right and straight gestures are 83%, 71% and 70% respectively.

4.3. Head Gesture Recognition

Head gestures, when examined, seemed to be correlated with prominences in speech. Since the evidence for correlation between sharp head movements and prosodic events in speech has previously been presented in gesture literature [9], we have decided to narrow down our investigation of head gestures to nods and head tilts. During nod gesture, the head comes down with chin closer to the body and sharply comes back up. During tilt gesture, the head rotates right or left 45° from its natural vertical position.

Given a set of training examples, three left to right continuous density HMMs are trained to model head gestures related with *nod*, *tilt* and *non-gesture*. These HMMs are then used to spot these gestures in testing sequence. The Viterbi algorithm is applied to determine the most probable gesture labels.

By changing the number of states used in HMMs, different performance metrics are obtained. The optimal number of states for head gesture recognition is achieved when $RRate$ and 1-FAR metrics are equal to each other. The optimal number of states for HMMs is 4 where $RRate$ metric is 80%.

5. CORRELATION ANALYSIS

After labels have been provided for speech and gesture events, correlation analysis has been conducted in order to provide justification for the two hypotheses. Directional hand gestures are closely correlated with the identified lexical candidate tokens, such as "left", "right" and "straight". Sharp head movements, such as nods and tilts, are closely correlated with speech prominences marked as pitch accents. In this section we will describe the correlation analysis procedure and results for both of these two hypotheses.

5.1. Directional Hand Gestures

Within the training portion labelled for directional hand gestures and speech keywords, 23 gestures were manually identified. Of the 23 gestures, 18 were matches with the candidate words "left", "right", and "straight", meaning that there was some degree of temporal overlap between the gestures and corresponding keywords. Of the remaining 5 gestures, 3 were wrongly identified as being related and 2 were designated as "confused", meaning that the speaker has correctly used the gesture to indicate going left, right or straight, but the phase of the gesture has overlapped with another word, usually being used in a different context. For example, the phrase: "Take a left and go *straight* down that street" had two accompanying left hand gestures. The first overlapped with the keyword "left" and was deemed a match, the second with the keyword "straight" and was marked as "confused".

5.2. Head Gestures

The training portion labelled for prosody and sharp head movements was found to contain 122 pitch accents and 81 head gestures 66 nods and 15 tilts. Of the 122 pitch accents, 79 or 64.75% overlapped with a head gesture, either a nod or a tilt. It is worth noting that from the 43 pitch accents that did not overlap with a head gesture, 23 or 53.5% were phrase initial accents, which are known to be problematic in prosody labelling. Often phrase-initial stressed syllables are misidentified as pitch accents due to the fact that both pitch accents and phrase-initial syllables are accompanied by "tense" voice quality [10].

If we disregard the 23 phrase initial syllables that were labelled as accents, only 20 of the 100 pitch accents identified did not overlap with a sharp head movement, that is 80% of remaining accents co-occurred with a head gesture.

The 79 accents that overlapped with a nod or a tilt were also examined for temporal correlation with the relevant head gesture. Time-stamp labels of the accented syllable were compared to the start and end time-stamps of the overlapping gesture using the statistical test of Pearson's correlation and the correlation test produced a correlation coefficient, $r = 0.994$, which implies an almost perfect correlation.

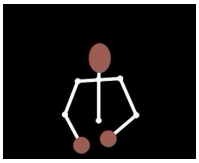
6. ANIMATION

Given a speech sequence, keyword spotter and accent detector are used to extract time-stamps of auditory events. These time-stamps and speech sequence are provided to animation engine to animate the virtual body. In this work, we realized two animation schemes:

Stick Model consists of line segments that corresponds to forearm and upper arm where starting and ending points of these line segments are determined as hand, shoulder and elbow positions. Together with these line segments, head is included with a line segment between head position and the center of the line segment between left and right shoulder. Animation engine for Stick Model uses 2D coordinates of the corresponding points.

3D Body Model consists of 2 arms and head without the body. Animation engine for this model uses a dictionary of gestural events and frames are constructed manually for each event in the dictionary. Animation engine uses each event independently for the animation of head, left arm and right arm. Sample stick and 3D body models are illustrated in Figure 2

In order to animate the body model, the center of mass positions of head and both hands is required by the animation engine. For each acoustical event, related gesture synthesized by considering the duration of acoustical event and the previously recognized gestures.



(a) Stick Model

(b) 3D Body Model

Fig. 2. Body Models

6.1. Hand Motion Model

During the left gesture, the motion of the right hand is limited when compared to the motion of the left hand. Similarly during the the right gesture, the motion of the left hand is limited when compared to the motion of the right hand. However for the straight gesture, both hands have large trajectories. The hand models for each hand gesture are constructed by HMMs. For the left gesture, we train an HMM by using only the left hand trajectory; for the right gesture, we train an HMM by using only the right hand trajectory and for the straight gesture we train two HMMs: one for the left hand and one for the right hand.

To construct an observation sequence from the HMM models, we use the model parameters: state transaction probabilities, parameters of Gaussian distribution for each state and prior probabilities of the states. Using this information, we construct an observation sequence by just providing a sequence length. The methodology used for constructing the observation sequence, given a sequence length and model parameters can be found on [6].

By using this methodology, we produce hand trajectories for each gesture where, for the *left* gesture, only left hand moves; for the *right* gesture, only right hand moves; and for the *straight* gesture both hands move.

Using the 20% portion of the database, we first run the keyword spotting algorithm for finding the time-stamps for words *left*, *right* and *straight*. We then produce the related hand gestures which are animated during the same period with the keyword.

6.2. Head Motion Model

Head motion model is generated according to the duration of accents. Let the duration of the accent be t_a seconds. For $t_a/2$ seconds head center of mass is shifted in $+y$ direction with 25 pixels/second. For the remaining $t_a/2$ seconds head center of mass is shifted back to its resting positions. The practical aspect of this methodology is that, the accents with short period are visually eliminated and the accents with long period are visually amplified.

7. CONCLUDING REMARKS AND FUTURE WORK

In this work, a gesture synthesizer based an audio-visual correlation is presented. Audio-visual correlation analysis is conducted using acoustic and visual events. Acoustic events are divided into semantic and prosodic categories. Visual events are selected as hand and head gestures. The types of events are defined by investigating a portion of the database. The repetitive patterns for acoustic events are mainly keywords (*left*, *right* and *straight*) and accents. The repetitive patterns for head gestures are *nod* and *tilt*. *Left* movement of left hand, *right* movement of right hand and *down* movement of both hands are defined as hand gestures.

Investigating the co-occurring patterns, we concluded that keywords and corresponding hand movements are strongly correlated. Moreover, *nod* movement of head is found out to be highly correlated with accents. Motivated from this fact, using the test portion of the database, first, keywords and accents are detected. Then the virtual body is animated using corresponding visual event at those detected acoustic events. Animation of the virtual body using both stick and 3D model can be found on [11].

As future work, we plan to build up new audio-visual database with new scenarios other than "Direction Giving". The number of keywords and gesture patterns will also be increased using these new scenarios for synthesis of more natural gestures.

8. REFERENCES

[1] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, 1976.

[2] S. Kettebekov, M. Yeasin, and R. Sharma, "Prosody based co-analysis for continuous recognition of coverbal gestures," *Proc. ICMI'02*, 2002.

[3] F. Quek, D. McNeill, R. Ansari, X-F. Ma, R. Bryll, S. Duncan, and K.E. McCullough, "Gesture cues for conversational interaction in monocular video," *Proc. Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems'99*, 1999.

[4] L. Valbonesi, R. Ansari, D. McNeill, F. Quek, S. Duncan, K. E. McCullough, and R. Bryll, "Multimodal signal analysis of prosody and hand motion: Temporal correlation of speech and gestures," *Proc. of the European Signal Processing Conference 2002 (EUSIPCO'02)*, vol. 1, pp. 75–78, 2002.

[5] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, "Lip feature extraction based on audio-visual correlation," *Proc. of the European Signal Processing Conference 2005 (EUSIPCO'05)*, 2005.

[6] M. E. Sargin, F. Ofli, Y. Yasinnik, O. Aran, A. Karpov, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, "Combined gesture-speech analysis and synthesis," *Proc. of the eNTERFACE'05 The SIMILAR Workshop on Multimodal Interfaces*, August 2005.

[7] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. of the Inst. of Phonetic Sciences*, vol. 17, pp. 97–110, 1993.

[8] Jie Yang and Yangsheng Xu, "Hidden markov model for gesture recognition," Tech. Rep. CMU-RI-TR-94-10, Robotics Institute, Carnegie Mellon University, May 1994.

[9] S. Duncan, F. Parrill, and D. Loehr, "Discourse factors in gesture and speech prosody," *Conf. of the International Society for Gesture Studies (ISGS)*, 2005.

[10] M.A. Epstein, "Voice quality and prosody in english," *Proceedings of the XVth International Congress of Phonetic Sciences*, 2003.

[11] M.E. Sargin, "Output of the animation engine," 2005, Available in `http://home.ku.edu.tr/~msargin/icme06/`.