Using Maximum Entropy (ME) Model to Incorporate Gesture Cues for SU Detection

Lei Chen
School of Electrical and
Computer Engineering
Purdue University
West Lafayette, IN, US 47906
chenl@ecn.purdue.edu

Mary Harper
School of Electrical and
Computer Engineering
Purdue University
West Lafayette, IN, US 47906
harper@ecn.purdue.edu

Zhongqiang Huang
School of Electrical and
Computer Engineering
Purdue University
West Lafayette, IN, US 47906
huang37@ecn.purdue.edu

ABSTRACT

Accurate identification of sentence units (SUs) in spontaneous speech has been found to improve the accuracy of speech recognition, as well as downstream applications such as parsing. In recent multimodal investigations, gestur]al features were utilized, in addition to lexical and prosodic cues from the speech channel, for detecting SUs in conversational interactions using a hidden Markov model (HMM) approach. Although this approach is computationally efficient and provides a convenient way to modularize the knowledge sources, it has two drawbacks for our SU task. First, standard HMM training methods maximize the joint probability of observations and hidden events, as opposed to the posterior probability of a hidden event given observations, a criterion more closely related to SU classification error. A second challenge for integrating gestural features is that their absence sanctions neither SU events nor non-events; it is only the co-timing of gestures with the speech channel that should impact our model. To address these problems, a Maximum Entropy (ME) model is used to combine multimodal cues for SU estimation. Experiments carried out on VACE multi-party meetings confirm that the ME modeling approach provides a solid framework for multimodal integration.

Categories and Subject Descriptors: H.5.1 [Multimedia Information Systems] Audio and Video Input, H.5.5 [Sound and Music Computing] Modeling and Signal Analysis, I.2.7 [Natural Language Processing] Meeting Processing

General Terms: Algorithms, Performance, Experimentation, Languages.

Keywords: multimodal fusion, gesture, prosody, language models, sentence boundary detection, meetings.

1. INTRODUCTION

In human-to-human conversations, ideas tend to unfold in a structured way. For example, speakers organize their utterances into *sentence units* (SUs). An SU tends to express a complete thought or idea. Accurate identification of sentence units (SUs) in spontaneous speech has been found to improve the accuracy of speech recognition [5], as well as downstream applications such as parsing [9]. Both lexical and prosodic cues from the speech channel have been utilized to detect SUs [12, 15]. However, humans tend to use not only speech but also visual cues, such as gesture, gaze, and body posture, to communicate with each other. It is commonly held that visual cues carry meaningful information to support the production and/or the perception of the communication [1, 8]. Some recent psycholinguistics studies suggest that gesture and speech stem from a single underlying mental process [18] and are related both temporally and semantically [17]. Because gesture is imagistic and uses quite a different expressive capacity than speech in a communication, we would expect gesture to provide additional important information that can be exploited when modeling structural events. In this paper, we describe a Maximum Entropy (ME) approach for incorporating gestural cues in a multimodal SU model.

This paper is organized as follows. Section 2 summarizes the previous related research. Section 3 describes the multimodal corpus used in this study. Section 4 describes the ME model. Section 5 describes the setup of the experiment and results. Section 6 draws some conclusions and describes future research directions.

2. PREVIOUS RESEARCH ON SU DETECTION

Gestures are used to provide meta-discursive information about a speaker's discourse. One such example involves a speaker counting on his/her fingers while itemizing several points using speech. McNeill [18] proposes that gesture and speech are two synchronized expressive streams generated and unpacked from a single underlying image-language entity, a *growth point* (GP). A GP is the minimal unit that retains the essential properties of an image-language expression as a whole. Since an SU is the word stream corresponding to a complete idea, it should be generated from a GP. It therefore follows that one would expect that some gestural cues should help signal the start or end of an SU. For example, near the end of an SU, if a speaker had been gesturing, he/she would often drop his/her hands to an area of rest

ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.
Copyright 2006 ACM 1-59593-541-X/06/0011 ...$5.00.

(e.g., the lap) to terminate the gesture.

The synchrony and co-expressivity of gesture and speech has inspired research on the incorporation of gestural cues into a multimodal SU model. Chen et al. [3] combined lexical, prosodic, and gestural cues into a multimodal SU model using a direct modeling approach. Figure 1 depicts the architecture for their multimodal SU detection system. This study was carried out on a small multimodal dialog corpus (KDI), in which two speakers discussed how to evacuate wombats from a town that was displayed as a model. Eisenstein and Davis [6] have constructed and evaluated a similar SU model using a small monolog instruction corpus. Their data set was elicited by having participants describe how a machine works by using speech and gesture while they were facing a diagram of a mechanical device drawn on a whiteboard.

In these two studies, SU detection has been treated as a classification task such that for each inter-word boundary, the classifier decides whether there is an SU end boundary or not. The SU boundary decision is based on three somewhat independent knowledge sources: lexical, prosodic, and gestural cues, as can be seen in Figure 1. A vector of prosodic and gestural features is extracted for each word boundary. Given that E denotes the word boundary class sequence (SU or not), W denotes the corresponding word sequence, and F and G denote the corresponding prosodic and gestural feature vectors, the goal is to estimate $P(W, F, G, E)$ and then choose the boundary classifications that have the highest probability given the observed words and multimodal features:

$$\begin{aligned} arg \max_E P(E|W, F, G) &= arg \max_E P(W, F, G, E) \\ &= arg \max_E P(W, E)P(F, G|W, E) \end{aligned}$$

There are two sets of parameters to estimate in this model. Let W_i denote the ith word, E_i denote the boundary event after W_i, and F_i and G_i denote W_i's prosodic and gestural features, respectively. $P(W, E)$ (where $W, E = W_1, E_1, W_2, E_2, \ldots W_{n-1}, E_{n-1}, W_n, E_n$) is estimated using a hidden event language model, which models the joint distribution of the word and event sequences. This word/event LM is trained from a corpus containing SU markups (but no markup for non-events). It is implemented using the SRILM toolkit [23] with a 4-gram model and Kneser-Ney smoothing.

The second set of HMM parameters are the observation likelihoods estimated from the prosodic and gestural features. Assuming conditional independence of the prosodic and gestural features and the word sequence given the events, we can approximate $P(F, G|W, E)$ as follows:

$$\begin{aligned} P(F, G|W, E) &\approx P(F, G|E) \\ &\approx \prod_{i=1}^{n} P(F_i, G_i|E_i) \end{aligned}$$

Although this assumption is not true (e.g., the phonetic makeup of a word can affect prosodic features), this independence assumption is reasonable in that we never use a word's identity directly when extracting prosodic and gestural features (only its alignment). We use a decision tree model that estimates $P(E_i|F_i, G_i)$ to obtain the observation probabilities as follows:

$$P(F_i, G_i|E_i) = \frac{P(E_i|F_i, G_i)}{P(E_i)} P(F_i, G_i)$$

Since $P(F_i, G_i)$ does not impact maximization, it can be ignored. Note that in practice we have found that better results are obtained by constructing separate decision tree classifiers that output posterior probability estimates $P(E_i|F_i)$ and $P(E_i|G_i)$, respectively, with the posterior probability $P(E_i|F_i, G_i)$ obtained as follows:

$$P(E_i|F_i, G_i) \approx \lambda P(E_i|F_i) + (1 - \lambda)P(E_i|G_i) \qquad (1)$$

To find the optimal event label, the forward-backward algorithm is used for HMMs [20] since this generally minimizes the boundary error rate.

Although similar HMM models were used in these two studies, different multimodal features were used. In [3], a series of prosodic features, which capture the duration, pitch, and energy patterns of speech, were extracted using SRI's prosodic feature extraction tool [7]. In [6], pause duration between words was the only prosodic feature. In [3], gestural features were obtained automatically on a small KDI data set. The VCM algorithm was used to track $3D$ hand positions directly from video. Based on the tracked hand position data, two important characteristics of hand gesture, *hold* and *effort*, were computed [19]. A series of numerical features were then derived, including 1) duration of hold within an inter-word boundary region, 2) basic statistics on effort, e.g., *minimum*, *maximum*, and *average*, within a window preceding (or following) a word boundary, and 3) effort change across an inter-word boundary. Instead of using automatically extracted gestural features, Eisenstein and Davis [6] manually annotated gestures according to gesture units, gesture phrases, and movement phases.

Chen et al. [3] found that the three-way model combination provided a lower overall error rate than the single and pair-wise model combinations. However, the improvement from adding gestural cues was not statistically significant. Eisenstein and Davis [6] obtained a similar result. They also obtained some improvement when adding gestural cues to their speech-only model, although the improvement was not statistically significant. They followed up with a regression analysis to argue that gestural cues may be providing redundant information to speech cues.

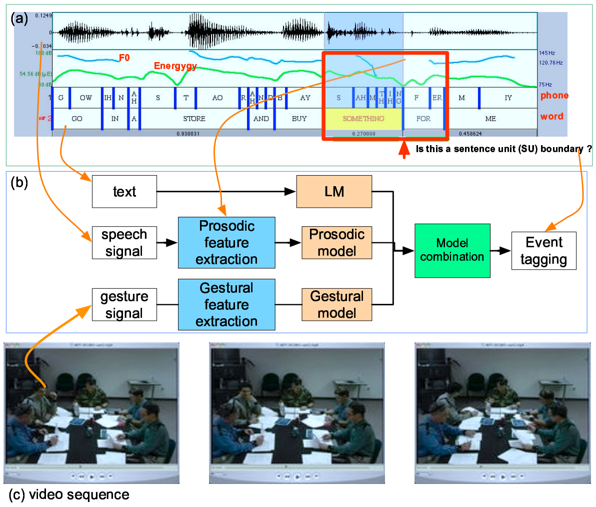

Figure 1: An HMM scheme for integrating lexical, prosodic, and gestural models

3. VACE MEETING DATA

The data sets used by Chen et al. [3] and Eisenstein and Davis [6] for multimodal SU detection were quite small. In this paper, we utilize a larger data set from the VACE multimodal meeting corpus [4] to investigate multimodal cues for SU detection. A comparison between the two previous multimodal SU detection data sets and the data set used in this study is presented in Table 1. The VACE meeting data set is larger and provides significantly more words and SUs than the other sets. Although both of the previous studies obtained improvements from the addition of gestural features, the error reductions were not statistically significant. However, since both studies were based on small data sets, the number of SU boundaries were quite limited. The VACE data set provides us with a greater number of participants, words, and SUs, thereby increasing the size and diversity of the data and the opportunity for gestural features to make a difference.

Data	# Participants	Type	# Words	# SUs
MIT [6]	9	Monolog	2,698	241
KDI [3]	3	Dialog	3,951	594
VACE	14	Meeting	24,566	3170

Table 1: A comparison of the VACE meeting data set to the previous data sets used for multimodal SU detection

The VACE multimodal meetings consist of planning exercises by four to five participants who were civilians, military personnel, or some mixture. For each recorded meeting, multichannel time synchronized audio and video recordings were collected and a series of audio and video processing tasks were carried out [4].

The Air Force Institute of Technology (AFIT) modified a lecture room to collect multimodal, time-synchronized audio, video, and motion data. In the middle of the room, up to 8 participants sit around a rectangular conference table. An overhead rail system permits the data acquisition technician to position 10 Canon GL2 camcorders in any configuration required to capture all participants by at least two of the camcorders. Using S-video transfer, 10 Panasonic AG-DV2500 recorders capture video data from the camcorders. The rail system also supports 9 Vicon MCam2 near-IR cameras that are driven by the Vicon V8i Data Station. The Vicon system records temporal position data. For audio recording, participants wear Countryman ISOMAX Earset wireless microphones to record their individual sound tracks. In addition, table-mounted wired microphones are used to record the audio. All audio signals are routed to a Yamaha MG32/14FX mixing console for gain and quality control. A TASCAM MX-2424 stores the sound tracks from both the wireless and wired microphones. In this study, we used the audio from the individual wireless microphones.

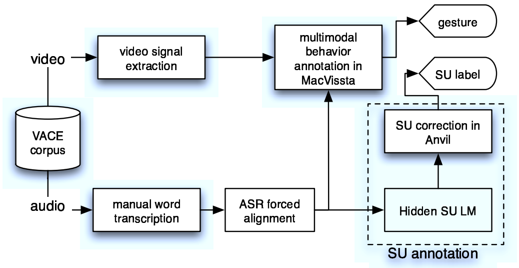

Figure 2: Data flow diagram for word transcription, SU annotation, and gesture annotation for the VACE meeting corpus

Figure 2 depicts the data flow used to transcribe and annotate the VACE meeting data. The meetings were transcribed by human transcribers (at the word level) based on LDC Quick Transcription guidelines, and then the transcripts were time aligned with the audio. Because the cameras were positioned to record the meeting participants from different viewing angles, accurate annotation of each participant's gaze direction and gestures was supported. Gesture annotation was carried out by researchers in the McNeill laboratory at University of Chicago using MacVissta [21], a general-purpose multimodal video display and annotation tool that runs on Mac OS X. The annotators had access to time aligned word transcriptions and all of the videos when producing the gesture annotations. Using the McNeill laboratory's gesture coding manual, the duration and type of five common types of gestures that are related to the content of the concurrent speech were annotated, including *metaphoric*, *iconic*, *emblematic*, *deictic* and *beat* gestures [17]. The annotators excluded fidgeting movements (e.g., tapping fingers while thinking, touching clothes), as well as instrumental movements (e.g., holding a cup, arranging papers on a desk).

Given the time aligned word transcriptions, we used the EARS MDE annotation specification V6.2 [1] to annotate the SUs using an Anvil [11] interface developed for this purpose. We used a hidden event LM trained on the MDE RT04S conversational dialog data set to initially estimate the position of each SU boundary. The annotator then corrected any inaccurate SUs and specified one of the four SU types (statement, question, incomplete, and backchannel) using a pull down menu in the Anvil SU interface. The SU interface displays time aligned word transcriptions with the automatic markups and allows the annotator to listen to audio and view video corresponding to selected portions of the transcripts.

For this experiment, we use three VACE meetings with time-aligned words, annotated SUs, and hand gesture annotations. These three meetings are named according to their recording date: Jan07, Mar18, and Apr25. Jan07 involves five participants who are discussing a plan to evaluate a foreign weapon component. Mar18 involves a collaborative discussion to select awardees for graduate fellowships by five faculty members at AFIT. Apr25 involves four participants discussing possible methods to stop the erosion of the Lincoln Memorial. To determine meeting duration, we measured it from the beginning of the first word to the end of the last word spoken by all meeting participants. Durations of the three meetings are: Jan07 (2,489.6 sec), Mar18 (2,624.83 sec) and Apr25 (2,096.94 sec).

Table 2 provides some basic statistics for each meeting participant's verbal communication, including the number of words spoken, the duration of all words, the number of SUs

[1] http://projects.ldc.upenn.edu/gale/Transcription/

Participant	# words	dur.(sec)	# SUs	SU (%)	AWPS
Jan07_C	1145	266.00	170	14.85	6.73
Jan07_D	2027	392.32	189	9.32	10.72
Jan07_E	3145	639.33	373	11.86	8.43
Jan07_F	2095	434.43	220	10.5	9.52
Jan07_G	1459	278.49	163	11.17	8.95
Mar18_C	2095	546.6	226	10.79	9.27
Mar18_D	1285	305.87	201	15.64	6.39
Mar18_E	1380	314.74	230	16.67	6
Mar18_F	1467	341.55	158	10.77	9.28
Mar18_G	1320	330.12	255	19.32	5.17
Apr25_C	1210	255.97	195	16.12	6.20
Apr25_D	3057	792.22	323	10.57	9.46
Apr25_F	2091	449.38	352	16.83	5.94
Apr25_G	790	162.59	115	14.56	6.87

Table 2: Basic statistics for each participant in the VACE meeting corpus

produced, the ratio of SUs to words, and the average words per SU (AWPS). Table 3 provides some basic statistics on each meeting participant's gesture behavior, including the total duration of his/her gestures, percentage of his/her gestures given the duration of each meeting, and percentage of the gestures given the duration of the words spoken by the participant. The overall percentage of gestures produced by each of the participants over the entire meeting is quite low. Although the percentage is higher when focusing on the time when the participants are speaking, there is still a considerable amount of time in which gestures do not co-occur with speech for each of the participants.

Participant	dur. (sec)	(%) meeting dur.	(%) word dur.
Jan07_C	181.68	7.32	68.3
Jan07_D	121.92	4.91	31.07
Jan07_E	393.43	15.86	61.54
Jan07_F	255.09	10.28	58.72
Jan07_G	223.29	9.00	80.18
Mar18_C	119.76	4.56	21.91
Mar18_D	60.29	2.30	19.71
Mar18_E	91.59	3.49	29.1
Mar18_F	214.25	8.16	62.73
Mar18_G	75.72	2.88	22.94
Apr25_C	128.56	6.13	50.22
Apr25_D	332.47	15.86	41.97
Apr25_F	133.3	6.37	29.66
Apr25_G	107.11	5.1	65.87

Table 3: Statistics of gesture usage for each participant in the VACE meeting corpus

4. MAXENT MODEL

4.1 Motivation

The HMM SU model investigated by Chen et al. [3] and Eisenstein and Davis [6] is a generative modeling approach; it is a stochastic process with hidden variables (SU labels) that produces observable data, including words, prosody, and gestures. The standard HMM training method maximizes the joint probability of the observed and the hidden events; however, the criterion used to evaluate an SU detection system is classification accuracy for each inter-word boundary. Another important issue for an HMM model relates to the fact that gestures do not occur over the entire communication stream, in contrast to the speech features, as can be observed from the gesture statistics presented in Table 3. A challenge for integrating gestural features is that their absence sanctions neither SU events nor non-events; (think of a telephone conversation with no gesture for example). It is only the co-timing of gestures with the speech channel that should impact our model. In regions where gestures are absent, the integration of a gesture model into an HMM may be less effective than it could be.

To address these two issues, we chose to integrate gestural features using a conditional Maximum Entropy (ME) model instead of the generative HMM. ME has been successfully applied to a variety of natural language processing tasks [2], including structural event detection in spontaneous speech [14, 16]. An ME model, which takes the exponential form, estimates the conditional probability of events given the multimodal features, which is a better match for our task.

$$P(E_i|W,F,G) = \frac{1}{Z_\lambda(W,F,G)}\exp(\sum_k \lambda_k g_k(E_i,W,F,G))$$

where $Z_\lambda(W,F,G)$ is the normalization term:

$$Z_\lambda(W,F,G) = \sum_{E_i}\exp(\sum_k \lambda_k g_k(E_i,W,F,G))$$

Note that $g_k(E_i,W,F,G)$ are indicator functions corresponding to features defined over SU events, words, and prosodic and gestural features, where k indicates different features, each of which has an associated λ_k. With the constraint that the expected values of the various multimodal feature functions $E_P[g_k(E_i,W,F,G)]$ match the empirical averages in the training data, the ME model is estimated by finding λ_k values that ensure the maximum entropy (uncertainty) of the distribution. The resulting ME model maximizes the conditional likelihood $\prod_i P(E_i|W,F,G)$ over the training data. In contrast to an HMM, the ME model decides on each inter-word boundary independently.

In our multimodal study, all indicator functions are divided into two groups: functions based only on words and prosodic features and functions also involving gestural features.

$$\begin{aligned} arg\max_E P(E_i|W,F,G) &= arg\max_E \exp(\sum_k \lambda_k g_k(E_i,W,F,G)) \\ \sum_k \lambda_k g_k(E_i,W,F,G) &= \sum_{k1} \lambda_{k1} g_{k1}(E_i,W,F) + \\ &\quad \sum_{k2} \lambda_{k2} g_{k2}(E_i,W,F,G) \end{aligned}$$

where, $g_{k1}(E_i,W,F)$ are indicator functions from the integrated HMM speech model described in Section 4.2, and $g_{k2}(E_i,W,F,G)$ are indicator functions involving gestural features described in Section 4.3. Figure 3 depicts the ME model architecture.

4.2 HMM Speech Model Features

The first set of features is derived using an HMM speech

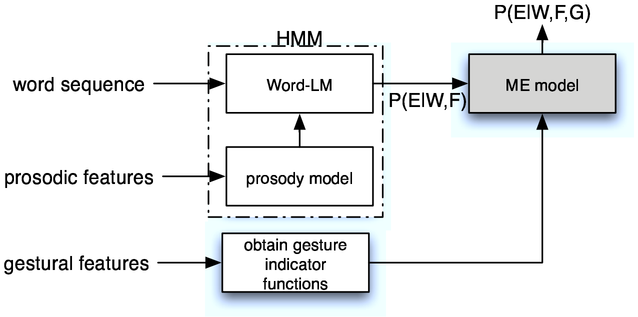

Figure 3: Maximum Entropy (ME) scheme used to combine lexical, prosodic, and gestural cues for SU detection.

model that combines lexical and prosodic features as follows:

$$\begin{aligned} arg\max_{E} P(E|W,F) &= arg\max_{E} P(W,F,E) \\ &= arg\max_{E} P(W,E)P(F|W,E) \\ &\approx arg\max_{E} P(W,E)P(F|E) \\ &\approx arg\max_{E} P(W,E)\prod_{i=1}^{n} P(F_i|E_i) \end{aligned}$$

$P(W,E)$ is estimated by a hidden event language model, using the SRILM toolkit [23] with a 4-gram model and Kneser-Ney smoothing. The prosodic features are modeled by a decision tree classifier that outputs posterior probability estimates $P(E_i|F_i)$, where E_i is the boundary event after W_i, and F_i is the corresponding prosodic feature vector. Hence, we obtain the observation probabilities as follows:

$$P(F_i|E_i) = \frac{P(E_i|F_i)}{P(E_i)}P(F_i)$$

$P(F_i)$ does not impact maximization and so can be ignored.

To model the prosody of sentence boundaries, we extract prosodic features around each word boundary. These features capture duration, pitch, and energy patterns in regions around the word boundaries [22]. The prosodic features are extracted using Purdue Prosodic Feature Extraction (PPFE) tool designed for *Praat* [10]. We briefly describe the prosodic features we used:

- **Duration Features:** Pause durations after each word boundary is extracted based on the alignment of human transcriptions. Phone durations are also computed. We also include the duration of a pause preceding the word that ends an SU. One possible indicator of an SU boundary in speech is preboundary lengthening. To capture such lengthening, we measure vowel and rhyme duration. We extract features such as the duration of the last vowel or stressed vowel in a multisyllabic word, as well as their normalization.

- **F_0 Features:** To obtain F_0 features, we first use Praat's auto-correlation based pitch tracker to obtain the raw pitch contour. The pitch baseline, top-line, and pitch range are computed based on the mean and variance of the log F_0 values. Voiced/unvoiced (V/UV) regions are identified and the pitch curve is stylized over each voiced segment. Using the stylized pitch contour, we compute several different types of F_0 features:

  Range features: These features reflect the pitch range of a single word or windows around an inter-word boundary. These include the minimum, maximum, mean, and last F_0 values for each word boundary. These features are normalized by baseline F_0 values using a linear difference and log difference. It is expected that speakers are more likely to fall near the bottom of their pitch range at a sentence boundary.

  Movement features: These features are obtained from the stylized F_0 contours for the voiced regions of the word preceding and the word following a boundary. Examples of such movement features are the minimum, maximum, and mean F_0 values, and the starting or ending stylized F_0 values, using various normalization methods.

  Slope features: The stylized pitch values generate pitch slope within a word or a predefined window length. We also consider the slope across a boundary to capture local pitch variation. A continuous trajectory is more likely to correlate with non-boundaries; whereas, a broken trajectory tends to indicate a boundary of some type.

- **Energy Features:** Speakers tend to start an utterance loudly and taper off over time. We compute energy features based on the intensity contour produced by Praat. Similar to the F_0 features, a variety of energy-related range features, movement features, and slope features are computed, using various normalization methods.

- **Additional Features:** The gender of each meeting participant is also included as an additional feature.

One set of indicator functions for the audio model are derived from the posterior probabilities, $p(E_i|W_i, F_i)$, which are estimated by a combined language and prosody model using the HMM modeling approach discussed above. For each word boundary i, the posterior is converted to a series of binary features using cumulative bins, for example $p > 0.1$, $p > 0.2$, $\ldots$, $p > 0.9$, and these binary features are used in the ME model. We also add a second set of indicator functions to encode more contextual information given the fact ME makes a decision about each inter-word boundary independently. Hence we include the binned posterior probabilities for the following word boundaries, $p(E_{i+1}|W_{i+1}, F_{i+1})$.

4.3 Gestural Features

In [3], the gestural features were automatically extracted from hold and effort measurements. However, these measurements are not currently available for the VACE meeting data. We do, however, have human-annotated gesture segmentations that can be utilized in our models, but these cannot be used to produce the types of numerical gestural features needed to train a decision tree gestural model like that studied in [3]. Since we have gesture segments, which have very course-granularity, we focus here on exploiting the temporal relationships between gestures, words, and SU events using the ME model.

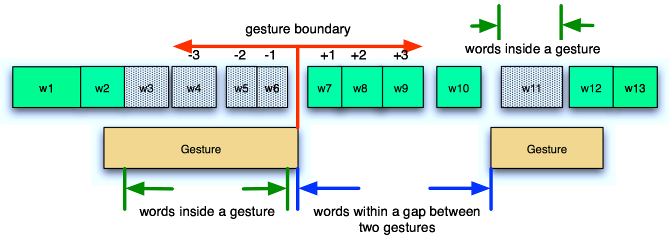

Figure 4: Word types given their temporal relationships with gestures

In order to design effective gestural features for ME modeling, we calculated the frequency of SUs on words with various temporal relationships to the beginning and ending points of the gestures in our data. Figure 4 depicts some regions of importance, in particular:

1. words inside a gesture,
2. words located around a gesture boundary, either at a gesture's start or end; here we consider a window of three words before and after a gesture boundary,
3. words within a gap between two gestures.

To get a sense for the impact of the temporal relationships of gestures on SU detection, we calculated the SU ratio, which is defined as the proportion of words in a set of words associated with a region that act as an SU boundary, to measure the contribution of the regional class to SU estimation. For example, we consider words immediately before a gesture (i.e., word[-1]), after a gesture (i.e., word[+1]), as well as words crossing a gesture boundary (i.e., word[0]).

Over all words in the corpus, the average SU ratio is 12.9%. For gesture-related regions mentioned above, the SU ratio measurements provide some interesting insights. For example, for words spoken within a gesture, the SU ratio is only 9.12%. Words uttered during a gesture are less likely to signal the end of an SU. This is consistent with a prediction based on Growth Point (GP) theory [18]; when a speaker is in the process of generating a gesture, he/she may still be unpacking the GP and so would still be in the process of forming an SU. For words around a gesture boundary, especially in some specific locations, there are higher or lower SU ratios than the average. SU ratios of words in these specific locations are reported in Table 4. As pointed out in Section 2, speakers often terminate their on-going gestures when they stop talking. Therefore, we tend find a higher SU ratio on the words just preceding a gesture end (see word [-1] in the "gesture end" row with a 20.01% SU ratio in Table 4). According to GP theory, a gesture's start or end can signal a GP's initiation or completion, respectively. Therefore, in the gap between the end of the previous gesture and the start of the next gesture, the current GP may finish and a new GP start. Completion of SUs should often co-occur with completion of a GP. Hence, we also consider words within a gap between two adjacent gestures. When we examine the words in the gap with the largest posterior probability from the combined language and prosody model, we find that its SU ratio is as high as 63.55%. Based on these measurements, a series of gesture features for SU estimation were defined:

$$g_{inside}(\cdot) = \begin{cases} 1 & \text{if the word is inside a gesture} \\ 0 & \textit{otherwise} \end{cases}$$

$$g_{start}(\cdot) = \begin{cases} 1 & \text{is the word } ith \text{ word from a gesture start,} \\ & \text{where } i \text{ is from } -3 \text{ to } +3 \\ 0 & \textit{otherwise} \end{cases}$$

$$g_{end}(\cdot) = \begin{cases} 1 & \text{is the word } ith \text{ word from a gesture end,} \\ & \text{where } i \text{ is from } -3 \text{ to } +3 \\ 0 & \textit{otherwise} \end{cases}$$

$$g_{gap}(\cdot) = \begin{cases} 1 & \text{if word is within a gap of two} \\ & \text{gestures, and has the largest } p(E_i|W_i, F_i) \\ 0 & \textit{otherwise} \end{cases}$$

5. SU DETECTION EXPERIMENT

5.1 Setup

In this experiment, we compare the performance of several models to our ME model, which integrates the gesture and speech features described in the previous section. For each model, we discuss how it was trained.

1. LM_{CTS}**:** A hidden event language model, denoted LM_{CTS}, uses lexical information only to estimate SUs. LM_{CTS} is trained using the MDE RT04S conversational dialog data set, which was annotated using the EARS MDE annotation specification V6.2 and contains approximately $480,000$ word tokens and $63,651$ SU labels. Due to its limited size, we did not build a language model using the VACE data set. LM_{CTS} is used to tag the VACE test data based on the word sequence information alone.
2. PM_{CTS}**:** Using PPFE tool described in section 4.2, we extracted prosody features from the RT04S data and trained 350 balanced CART trees to serve as the prosody model. The obtained prosody model tags the word sequences in the VACE test set using prosodic information only.
3. PM_{VACE}**:** Using PPFE tool, we also extracted prosody features using the VACE meeting audio data. For this model, we used 10-fold cross-validation on the three VACE meetings. From preliminary investigations, we found that a single decision tree trained using the entire training set with no downsampling or bagging performed the best. See [13] for more information on sampling and bagging for SU detection.
4. LM_{CTS}+PM_{CTS} and LM_{CTS}+PM_{VACE}: Using the SRILM toolkit [23], we integrated the hidden event language model with each of the prosody models. We used the default settings for two parameters used to adjust the contribution of the language and prosody models ($lmw = 1$ and $mapw = 1$).
5. $\{LM_{CTS}+PM_{CTS}\}_{ME}+GM_{ME}$: Using an ME approach, we constructed a model that integrates the binned posterior probabilities from the HMM SU detection model described in section 4.2 and the gestural

location	word[-3]	word[-2]	word[-1]	word[0]	word[+1]	word[+2]	word[+3]
gesture start	13.85%	15.62%	17.94%	11.03%	19.6%	8.91%	8.3%
gesture end	8.74%	10.10%	20.01%	16.61%	11.2%	11.74%	11.94%

Table 4: The ratio of SUs to words in specific regions around gesture boundaries, i.e., the words before a gesture (word[$-n$]), after a gesture (word[$+n$]), or crossing a gesture (word[0]), where $n = 1, 2, 3$.

indicator functions described in section 4.3. We used the ME toolkit of [24], which uses L-BFGS parameter estimation and Gaussian prior smoothing (with a Gaussian prior of 1). We use the same 10-fold cross-validation train-test splits as were used for the PM_{VACE}.

To evaluate the performance of our models, we use the Error Rate metric defined by NIST for the DARPA EARs metadata evaluation. To calculate the SU Error Rate, the estimated SU string is compared with the gold standard SU string to determine the number of misclassified boundaries per SU. Since SU boundaries may be incorrectly deleted or inserted, we also provide the Insertion Rate and Deletion Rate to examine the patterns of insertions and deletions among the different models. The Insertion Rate is the number of incorrect insertions of an SU in the estimated SU string that does not appear in the gold standard per SU boundary; whereas, the Deletion Rate is the number of incorrect deletions of an SU that appears in the gold standard string per SU boundary. The three metrics appear below:

1. *Error Rate* = (# *Deletion* + # *Insertion*)/ # SUs in the reference
2. *Insertion Rate* = # *Insertion*/ # SUs in the reference
3. *Deletion Rate* = # *Deletion*/ # SUs in the reference

5.2 Results and Discussion

Table 5 compares the performances of several SU models, including the ME model incorporating speech features and additional gestural features. To evaluate the various SU models, we utilize a threshhold, $\theta = 0.5$. This threshold indicates the minimum posterior probability required for a model to produce an SU label (rather than a non-even label). This value is commonly used in the literature when reporting SU detection results, and this threshold provides us with the lowest error rate for our HMM speech models (providing a strong HMM baseline). Note that all results were obtained using 10-fold cross validation runs repeated twenty times with different data splits. The results were then averaged.

Approach	*INS*	*DEL*	*ERR*
LM_{CTS}	17.13	40.54	57.67
PM_{CTS}	11.48	39.59	51.07
PM_{VACE}	9.78	42.71	52.49
$LM_{CTS} + PM_{CTS}$	19.68	18.68	38.36
$LM_{CTS} + PM_{VACE}$	15.17	23.72	38.89
$\{LM_{CTS} + PM_{CTS}\}_{ME} + GM_{ME}$	15.79	21.15	36.94

Table 5: Comparing HMM and ME models on the VACE meeting data set

The language model LM_{CTS} has the largest SU error (57.67%) of all the models in Table 5. When the same language model has been applied to the RT04 eval data set (a genre-matched test set), it obtained a far lower SU error of 45.48% [10]. LM_{CTS} was trained on dialog data, and so it is not surprising that it performed better on the RT04 test set than on the VACE meeting data, which is from a different genre.

The prosody model PM_{CTS} has an SU error of 51.07%. This prosody model, when evaluated on the RT04 eval set, obtained an SU error of 59.08% [10]. The speech in our meetings involves many more speakers per conversation, and so there tends to be more short utterances (e.g., backchannels or short responses to questions) that are separated by long silences. Hence it is not surprising that SU prediction in this setting could be somewhat easier for the prosody model than in a dialog setting. The prosody model trained on the VACE meetings, PM_{VACE}, performs more poorly than PM_{CTS}, possibly due to the limited training data size. This is, in fact, a good explanation since the deletion error is greater for PM_{VACE} than PM_{CTS}.

The combinations of the language model and prosody models using SRILM toolkit [23], $LM_{CTS} + PM_{CTS}$ and $LM_{CTS}+PM_{VACE}$, both obtain error reductions relative to the individual prosody and language models. The decrease in error results from a substantial reduction in SU deletions. Using all of the indicator functions described in 5.1, which incorporate binned posteriors from HMM speech model and gestural features, our ME model, $\{LM_{CTS} + PM_{CTS}\}_{ME} + GM_{ME}$, obtains a significant reduction in error relative to the baseline HMM speech model, $LM_{CTS} + PM_{CTS}$, $p = 0.0175 < 0.05$ using the sign test.

To get a better picture of the factors contributing to our ME model's performance, we have conducted a study wherein some of the feature types are deleted from the model. There are several types of features, including **HMM$_i$:** binned posteriors $p(E_i|W_i, F_i)$ from the speech model; **HMM$_{i,i+1}$:** combined binned posteriors $p(E_i|W_i, F_i)$ and $p(E_{i+1}|W_{i+1}, F_{i+1})$ to add contextual cues to the model; and **Gesture:** the gestural features. We also decompose the HMM posteriors into binned language model and prosody model posteriors (denoted **(LM, PM)**) to evaluate their loose coupling in the ME model. Note that in this experiment, only the language model and prosody model trained on RT04S data are utilized. Results for each of these models appear in Table 6.

Feature Set	*INS*	*DEL*	*ERR*
Gesture	11.84	76.18	88.03
LM,PM	21.23	31.56	52.79
HMM$_i$	16.15	21.63	37.79
HMM$_{i,i+1}$	15.34	22.24	37.57
HMM$_{i,i+1}$+Gesture	15.79	21.15	36.94

Table 6: A comparison of models with different feature combinations using ME

The ME model using only **Gesture** features is quite weak but is better than the model where we always choose the majority class, which has an SU error of 100% (0% INS error and 100% DEL error). The error obtained from the integration of the language model and prosody model features in ME, (**LM,PM**) in Table 6, has a much larger error than that obtained by the HMM model, $LM_{CTS} + PM_{CTS}$ in Table 5. However, using $\mathbf{HMM}_i$ features, we obtain a model that is slightly more accurate than the original HMM speech model, and by adding additional context, $\mathbf{HMM}_{i,i+1}$, we obtain further error reductions. Finally, when we add the **Gesture** features, we further reduce the error to 36.94%. It is interesting to note that the addition of gestural features to $\mathbf{HMM}_i$ resulted in only a negligible reduction in error. It appears that for gesture features to be effectively integrated in ME, we must provide contextual information from the speech model.

6. CONCLUSIONS

By carefully observing gesture usage and its impact on sentence formation, we have created an ME model that reduces error significantly over an HMM speech-only model. This experiment builds on a strong baseline HMM model that incorporates lexical and prosodic features. We also investigated the factors that contributed to our ME model's error reductions. It was very important to incorporate speech features in a way that provides context for our gestural features to exploit.

In future research, we will investigate incorporating lexical and prosodic features directly into the ME model. We also plan to expand our multimodal meeting corpus, as well as identify additional gesture constraints. Finally, we plan to utilize automatically derived gesture features in future investigations. In [3], we were able to utilize some of the dynamics of gesture, which were not utilized here. We believe that gesture dynamics are also quite important for predicting SU boundaries.

7. ACKNOWLEDGMENTS

The authors are grateful to Dr. Yang Liu for her valuable input. We also thank all of our VACE team members for their efforts in producing the VACE meeting corpus: Dr. David McNeill, Dr. Susan Duncan, Jim Goss, Fey Parrill, Haleema Welji, and Amy Franklin from University of Chicago; Dr. Francis Quek, Dr. Yingen Xiong, Travis Rose, Bing Fang, and Dulan Wathugala from Virginia Tech; Dr. Ron Tuttle, Dr. David Bunker, Jim Walker, Kevin Pope, and Jeff Sitler from AFIT. This research has been supported by ARDA under contract number MDA904-03-C-1788. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views ARDA.

8. REFERENCES

[1] M. Argyle and M. Cook. *Gaze and Mutual Gaze.* Cambridge Univ. Press, 1976.

[2] A. Berger, S. Pietra, and V. Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–72, 1996.

[3] L. Chen, Y. Liu, M. Harper, and E. Shriberg. Multimodal model integration for sentence unit detection. In *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, University Park, PA, Oct 2004.

[4] L. Chen, T. Rose, F. Parrill, X. Han, J. Tu, Z. Huang, I. Kimbara, H. Welji, M. Harper, F. Quek, D. McNeill, S. Duncan, R. Tuttle, and T. Huang. VACE multimodal meeting corpus. In *Proc. of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, 2005.

[5] S. Coquoz. Broadcast news segmentation using MDE and STT information to improve speech recognition. Technical report, International Computer Science Institute, 2004.

[6] J. Eisenstein and R. Davis. Gestural cues for sentence segmentation. *MIT AI Memo*, 2005.

[7] L. Ferrer. Prosodic features extraction. Technical report, SRI, 2002.

[8] S. Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11), Nov 1999.

[9] M. Harper, B. Dorr, B. Roark, J. Hale, Z. Shafran, Y. Liu, M. Lease, M. Snover, L. Young, R. Stewart, and A. Krasnyanskaya. Final report: Parsing speech and structural event detection. Technical report, John Hopkins CSLP Summer Workshop, 2005.

[10] Z. Huang, L. Chen, and M. Harper. An open source prosodic feature extraction tool. In *Proc. of the Language Resources and Evaluation Conference (LREC)*, May 2006.

[11] M. Kipp. Anvil: A generic annotation tool for multimodal dialogue. In *Proc. of European Conf. on Speech Processing (EuroSpeech)*, 2001.

[12] Y. Liu. *Structural Event Detection for Rich Transcription of Speech.* PhD thesis, Purdue University, 2004.

[13] Y. Liu, N. V. Chawla, E. Shriberg, A. Stolcke, and M. Harper. Resampling techniques for Sentence Boundary Detection: A Case Study in Machine Learning from Imbalanced Data for Spoken Language Processing. *Computer Speech and Language*, to appear.

[14] Y. Liu, E. Shriberg, A. Stockle, and M. Harper. Comparing HMM, Maximum Entropy, and Conditional Random Fields for disfluency detection. In *Proc. of InterSpeech*, Lisbon, Sept. 2005.

[15] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, H. D., M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. Structural Metadata Research in the EARS Program. In *Proc. of ICASSP*, 2005.

[16] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper. Comparing and combining generative and posterior probability models: Some advances in sentence boundary detection in speech. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[17] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought.* Univ. Chicago Press, 1992.

[18] D. McNeill. Growth points, catchments, and contexts. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, 7(1), 2000.

[19] F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. E. McCullough, and R. Ansari. Multimodal human discourse: Gesture and speech. *ACM Trans. Comput.-Hum. Interact.*, 9(3):171–193, 2002.

[20] L. R. Rabiner and B. H. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

[21] T. Rose, F. Quek, and Y. Shi. Macvissta: A system for multimodal analysis. In *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, 2004.

[22] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000.

[23] A. Stockle. SRILM - An extensible language modeling toolkit. In *Proc. of Int. Conf. on Spoken Language Processing (ICSLP)*, 2002.

[24] L. Zhang. *Maximum Entropy Modeling Toolkit for Python and C++.* http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html.