Multimodal Floor Control Shift Detection

Lei Chen
ECE School, Purdue University
West Lafayette, IN 47906
present contact: Educational Testing Service (ETS)
Princeton, NJ 08541
LChen@ets.org

Mary P. Harper
UMIACS, University of Maryland
College Park, MD 20472
Human Language Technology Center of Excellence, Johns Hopkins Univ.
Baltimore, MD 21211
mharper@umd.edu

ABSTRACT

Floor control is a scheme used by people to organize speaking turns in multi-party conversations. Identifying the floor control shifts is important for understanding a conversation's structure and would be helpful for more natural human computer interaction systems. Although people tend to use verbal and nonverbal cues for managing floor control shifts, only audio cues, e.g., lexical and prosodic cues, have been used in most previous investigations on speaking turn prediction. In this paper, we present a statistical model to automatically detect floor control shifts using both verbal and nonverbal cues. Our experimental results show that using a combination of verbal and nonverbal cues provides more accurate detection.

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, Design, Experimentation, Performance

Keywords

Multimodal fusion, nonverbal communication, floor control, prosody, language models

1. INTRODUCTION

Floor control is an underlying scheme used by human beings to manage the order of speaking in conversations. Knowledge about the floor control structure in multi-party conversations can play an important role in various tasks, including: (1) better analysis of conversational content, and (2) design of more natural human computer interaction systems. For example, based on knowledge of floor control structure, utterances spoken by different meeting participants can be coherently connected to support a clearer description of the meeting content; human-like computer agents can also obtain and release control of the floor to interact effectively with human participants.

A large body of previous studies (e.g., [10, 16, 27, 32]) suggests that people tend to use both verbal (e.g., lexical and prosodic cues) and nonverbal cues (e.g., facial expression, eye gaze, hand gesture, and body posture) to efficiently manage floor control in conversations. However, most of the research on automatic detection of floor control has focused on utilizing speech features and has neglected nonverbal cues. In this paper, we will investigate combining verbal with nonverbal cues (i.e., hand gesture and eye gaze) to detect floor control shifts in multi-party meetings.

This paper is organized as follows: Section 2 describes the related previous research. Section 3 describes the multimodal corpus and floor control shift task used in our experiment. Section 4 describes the metrics used for the evaluation. Section 5 describes multimodal features, including lexical, prosodic, and visual features (i.e., gaze and gesture). Section 6 describes our evaluation plan and experimental results from several statistical models for floor control shift detection. Section 7 summarizes key findings on floor control shift detection using multimodal cues.

2. PREVIOUS RESEARCH

Sacks et al. [32] proposed a turn-taking model, in which turn taking can only happen at transition relevance points (TRPs), which are the projected end points of lexically or semantically defined units, called *turn constructional units (TCUs)*. The principle governing TCUs is that participants in interaction try to minimize the gap between and overlap of TCUs to form fluent speech.

To provide natural turn taking in human-to-computer dialog systems, some computational models of turn-taking have combined lexical and prosodic cues. Schlangen [33] used lexical and prosodic information at the ends of utterances to predict turn changes. His experiments on part of the Switchboard corpus and his own small German corpus showed that predictions based on audio cues achieve greater accuracy than predictions using only silences [12]. Levow [23] conducted a similar experiment on Mandarin, a tonal language. Her investigation suggested that intonation plays an important role for signaling turn-taking, even in a tonal language.

Nonverbal behaviors play an important role in coordinating turn-taking and the organization of discourse. Duncan [10] proposed some nonverbal cues for signaling turn taking in face-to-face communication, including gesture, gaze, and facial expression. Gaze is an important cue for managing floor control [2, 19]. Near the end of an utterance, the floor holder may gaze at an interlocutor to prepare to trans-

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

fer control of the floor [1]. The interlocutor who is gazed at has the advantage for taking the floor because he/she infers that the speaker is conceding the floor to him/her. Kendon [20] observed that utterances that terminate without gaze cues more frequently had delayed listener response. Vertegaal et al. [37] observed that the absence of gaze significantly decreases turn-taking efficiency in a multi-party mediated system. Novick et al. [29] reported on a special pattern of eye gaze that is important for floor control, the *mutual gaze break*. A mutual gaze break contains three components: (1) the current speaker looks towards a listener as an utterance comes to an end, (2) the speaker and the listener share a mutual gaze momentarily, and (3) the listener (following the speaker) breaks the mutual gaze and begins to speak. Novick et al. found that this pattern occurs in 42% of turns in their dialog corpus.

Gesture and body posture are also important cues for managing floor control change. When a speaker keeps gesturing, there is usually no turn transition. Duncan [10] observed that floor control attempts fell to virtually zero when the dominant speaker gesticulated at a phonemic clause boundary. Posture shifts tend to occur during an utterance's beginning and ending, as well as at various discourse boundaries [4].

Some investigations of floor control in meetings, especially using audio and visual signals have emerged recently. For example, Padilha and Carletta [30] investigated small-group discussions using a simulation approach and found that using nonverbal cues improved the smoothness of conversation in their simulations. However, compared with research on floor control (turn-taking) in dialogs, the research on floor control in meetings is new and has many open questions. For example, Novick [28] summarized current studies of eye gaze in multi-party interactions and suggested that a gaze model specific to multi-party meetings should be developed.

Floor control is just one aspect of meeting structure. There has been an increasing number of studies on detecting other aspects of meeting structure and participants' roles. For example, *addressee* detection in meetings is related to floor control detection since the addressee of the current floor holder is frequently the next floor holder given a change in control of the floor. Jovanovic et al. [18] used a *Bayes* network model to identify addressees using verbal, non-verbal, and contextual features. Their results showed that by adding nonverbal cues, the accuracy rate for identifying addressees is improved. A comprehensive review of pertinent research on these topics can be found in [15]. To our knowledge, and as pointed out in [15], research on floor control detection in meetings using multimodal cues has been limited.

3. THE DATA AND THE TASK

3.1 VACE Multimodal Meeting Corpus and Annotations

We utilized the VACE multimodal meeting corpus [7] in our experiment. Each VACE meeting has an associated set of multimodal signals and annotations, including time-aligned word transcriptions, sentence unit (SU) annotations, floor control annotations, and gesture and gaze annotations. Each VACE meeting was named according to its recording date. In our experiment, we used three meetings with full annotations: **Jan07**, **Mar18**, and **Apr25**. These three meetings contain 14 speakers[1]. Table 1 provides information about the participants in each of the three meetings; the first column provides a label for each speaker and the second, the total speaking time duration.

Participant	Dur.(sec.)	# Control	$Dur_{Control}$(sec.)
Jan07_C	337.32	37	299.58
Jan07_D	539.13	26	465.54
Jan07_E	820.51	63	763.31
Jan07_F	579.42	37	523.16
Jan07_G	352.92	31	296.31
Mar18_C	679.39	62	648.73
Mar18_D	390.46	54	359.75
Mar18_E	485.21	49	465.03
Mar18_F	486.60	57	481.70
Mar18_G	470.72	53	422.49
Apr25_C	382.56	63	340.7
Apr25_D	1029.96	72	990.61
Apr25_F	532.86	80	450.11
Apr25_G	197.96	34	185.06

Table 1: Statistics of each meeting participant's speaking time, number of Control type floors, and duration of Control type floors

Using a multi-channel audio and video data collection system as described in [7], each meeting participant's audio and video signals (from several different viewing angles) were collected. The speech content was transcribed by human annotators, and the starting and ending time points of words were added by doing a forced alignment. Then, sentence units (SUs), which represent the smallest complete idea in utterances, were annotated following [35]. A total of 3170 SUs were annotated for these three VACE meetings.

Psycholinguistic researchers hand annotated the gesture and gaze of each of the participants in the meetings using the MacVissta [31] tool, which displayed multiple videos along with the time-aligned words and silences. Gesture onset and offset, as well as the semiotic properties of the gesture as a whole, were coded in relation to the accompanying speech. In addition, gaze patterns were coded for each speaker in terms of the object of the gaze (at whom or what the gaze was directed) at each moment. See [6, 7] for more information on the annotation process.

The floor control annotations primarily focus on speaking turn management. When participant A is talking to participant B, who is listening without attempting to break in, then A clearly has "control of the floor". The person controlling the floor bears the burden of moving the discourse along. The floor control event types annotated in the VACE corpus are:

Control: corresponds to the main communication stream in meetings. We annotated who had control and which participants were involved.

Sidebar: corresponds to sub-floors that have split off of a more encompassing floor.

Backchannel: corresponds to utterances like "yeah" that are spoken when another controls the floor.

Challenge: corresponds to an attempt to grab the floor.

Cooperative: corresponds to an utterance that is inserted into the middle of the floor controller's utterance in a way that is much like a backchannel but has propositional content.

[1]Only C in the **Jan07** meeting is a female speaker.

Other: corresponds to other types of vocalizations that do not contribute to any current floor thread, e.g., self talk.

More details about the floor control annotation can be found in [6]. Using this annotation, the use of nonverbal behaviors, e.g., gesture and gaze, have been analyzed for their effect on floor change management [6].

In addition to the speaking time of each participant in each meeting, Table 1 provides some basic statistics on floor control in the three meetings: the number of Control type floors and total duration of the Control type floors for each meeting participant[2]. Clearly, most of the speaking time is assigned to Control type floors in these meetings.

3.2 Floor Control Shift Detection

Floor control shift (FCS) detection can be treated as a classification task. At the end of each SU spoken by the current floor holder, the goal is to determine whether they will keep or give up the control of the floor. Figure 1 depicts this task. In this figure, there are three SUs in the example floor, which is depicted above the SUs. At the end of each SU, a FCS detection model predicts whether the floor will be kept or yielded. In this example, the control of the floor is not yielded until the end of the third SU.

Each SU ending boundary within all Control type floors is classified as one of two classes: *Keep* or *Change*. The *Keep* class label implies that after the SU's ending boundary, control of the floor continues to be held by the current floor holder, i.e., there is no floor control shift. The *Change* label indicates that after the SU's ending boundary, control of the floor is shifted to a new holder. Table 2 lists the floor control shift class distribution on all SU ending boundaries. Note that the column labeled "None" indicates the number of SUs that do not occur in Control type floors. In the three VACE meetings, among 3170 inter-SU boundaries, 1181 inter-SU boundaries do not occur in Control type floors, most of which are backchannel SUs (865). Of the remaining inter-SU boundaries occurring in Control type floors, there were were 1301 *Keep* FCS and 688 *Change* FCS.

SUs	Keep	Change	None
3170	1301	688	1181

Table 2: Statistics of floor control shifts in the VACE meeting corpus

For FCS detection, we focused on Control type floors for the following three reasons: (1) Control type floors contain the most important information that is conveyed and exchanged in conversations; (2) for SUs in non-Control type events (e.g., Backchannel, Challenge, and Cooperation), the speaker is not the floor holder and so floor control shift is undefined; and (3) although the technique developed from Control type floors could be used on sub-floors within Sidebar type floors, these event types are rare in our corpus.

In addition, for FCS detection, we utilized multimodal cues appearing in the region immediately prior to each SU's ending boundary (either the word or the predefined window just previous to the SU boundary). This is because we intend to build an on-line floor control shift detection system. By only using multimodal features occurring previous to SU boundaries, we do not need to wait for features following the SU boundaries in order to predict potential floor control shifts. Although long-range pragmatic information (e.g., the topic structure) is likely to be quite useful for predicting floor control shifts, here we focused on using cues extracted locally for two reasons: (1) utilizing long-range pragmatic information is quite challenging for current language processing techniques, and (2) our data resources in this study were quite limited, making it difficult to characterize long-range pragmatic information.

[2]It should be noted that these time measurements are based on time intervals that contain some silence.

4. METRICS

To evaluate the performance of our models described in Section 6, we designed a metric (ERR) in the spirit of the one defined by NIST for the sentence unit (SU) evaluation in the DARPA EARS program. To calculate the FCS ERR, the estimated FCS sequence is compared with the standard FCS reference string to determine the number of misclassified boundaries per Change FCS. Since FCS boundaries may be incorrectly deleted or inserted, we also provide the insertion rate (INS) and deletion rate (DEL) to examine the patterns of insertions and deletions among the different models. The three metrics are defined as follows:

$$\text{INS} = \frac{\text{number of incorrect insertions of Change boundaries}}{\text{total number of FCS boundaries with Change labels}}$$

$$\text{DEL} = \frac{\text{number of incorrect deletions of Change boundaries}}{\text{total number of FCS boundaries with Change labels}}$$

$$\text{ERR} = \text{INS} + \text{DEL}$$

Note that the NIST-style Error rate can be greater than 100%. The following example shows a system FCS hypothesis aligned with the reference FCS labels:

```
Reference:   SU1     SU2    SU3  /  SU4
System:      SU1  /  SU2    SU3     SU4
                 INS             DEL
```

where SU_i is an SU and '/' indicates an FCS Change event. There are two misclassified boundaries, one insertion error and one deletion error, in the example above. Since there is only one reference FCS Change boundary, the NIST ERR rate for this system output is 200%. If a system hypothesizes a non-event boundary at each inter-SU boundary, then the NIST error rate would be 100% for the boundary detection task, all due to deletion errors.

The classification error rate (CER) is also calculated to enable us to test whether results from different modeling approaches are significantly different using the sign-test [22]. CER is defined below:

$$CER = \frac{\text{number of incorrect FCS boundaries}}{\text{total number of SU ending boundaries}}$$

5. MULTIMODAL FEATURES

Years of conversational analysis (CA) studies provide evidence that a wide variety of cues, e.g., lexical cues, prosodic cues, and visual cues, are used to efficiently manage floor control in human conversations [10, 16, 27, 32]. Because of this, in our floor control shift detection experiment, we investigated the use of lexical features from word transcriptions, prosodic features extracted from audio signals, and visual features computed from the annotations of gesture and gaze behaviors.

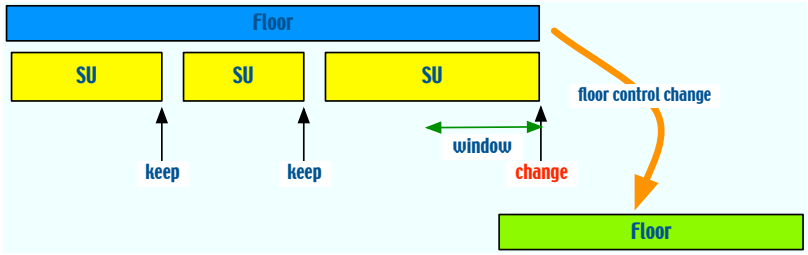

Figure 1: A schematic view of floor control shift detection

5.1 Lexical Features

Lexical cues have been found to be quite useful for signaling floor control shift in previous studies [10, 32]. For example, syntactic completeness is an important indicator for a speaking turn change. Also some phrases, e.g., *any suggestions*, are commonly used by people to yield floor control. Since our floor control shift detection was applied on all SU ending boundaries, the syntactic completeness was largely guaranteed[3]. In addition, we utilized word and part of speech (POS) co-occurrence information, which were successfully used in previous turn-taking detection algorithms [23, 33], in an attempt to model lexical cues provided by word sequences.

Given the word transcriptions and POS tag sequence, which was computed using a POS tagger [36] trained on the transcriptions of spontaneous speech, we extracted the following n-gram features:

- **Word n-gram features:** where the maximum n is four. Given w_i as the word token at position i (the last word of the current SU), the word-ngram features include: $\langle w_{i-3}, w_{i-2}, w_{i-1}, w_i \rangle$, $\langle w_{i-2}, w_{i-1}, w_i \rangle$, $\langle w_{i-1}, w_i \rangle$, $\langle w_i \rangle$.

- **POS n-gram features:** where the maximum n is four. Given P_i as the POS token at position i (the last word of the current SU), POS n-gram features include: $\langle P_{i-3}, P_{i-2}, P_{i-1}, P_i \rangle$, $\langle P_{i-2}, P_{i-1}, P_i \rangle$, $\langle P_{i-1}, P_i \rangle$, $\langle P_i \rangle$.

5.2 Prosodic Features

Previous investigations have suggested that prosody plays a role for managing floor control in conversations. Silent pauses, rises or falls in intonation, variation in speech rate, final lengthening, and other prosodic patterns have been used to indicate speaker turn structure [10, 13, 25, 38].

In our experiment, we tailored the Purdue Prosodic Feature Extraction (PPFE) [17] tool's output to obtain prosodic features suitable for floor control shift detection. Although the PPFE tool's default output was designed to support sentence unit (SU) detection, its feature computation is quite general. In previous studies, similar prosodic features have been successfully utilized for detecting topic segmentation, dialog acts, and utterance boundaries [34].

The prosodic features used for floor control shift detection are related to various aspects of prosody, including F_0, energy, and duration. However, as described in Section 3.2, we developed a left-to-right model to predict FCS using multimodal features happening previous to an SU boundary. Therefore, in contrast to the prosodic features used for SU detection, the prosodic features used for FCS detection are: (1) computed at each SU ending boundary, and (2) computed only using the audio prior to each SU ending boundary. The prosodic features include:

- **Duration Features:** Final lengthening is an useful indicator of a floor control shift. The drawl on the final syllable tends to indicate turn-yielding [10]. We extract several duration related features, e.g., the duration of the last word before an SU boundary, the duration of the last phone, and the duration of the last vowel or stressed vowel in a multisyllabic word, as well as their normalizations.

- **F_0 Features:** F_0 variation has been reported to signal floor control shifts in previous studies. For example, to yield control of the floor, a rising pitch contour, which indicates a question SU, can provide an important cue. To model F_0 variation, we extract features about the range, movement, and slope of the F_0 contour:

  Range features: The range features reflect the pitch range of a word or a 0.2 second window relative to the speaker-specific baseline F_0 value. Examples of such range features include the minimum, maximum, mean, and last F_0 values for the SU boundary.
  Movement features: The movement features reflect the variation of F_0 values. From the stylized F_0 contours for the voiced regions of the word preceding the SU boundary, the minimum, maximum, mean, and the starting and ending stylized F_0 values are computed.
  Slope features: The slope features reflect the variation of the F_0 contour. Slope patterns are in the format of a sequence of "f", "r", representing a falling slope and a rising slope. Such slope patterns are computed on the word or a 0.2 second window preceding the SU boundary.

- **Energy Features:** When yielding control of the floor, the current floor holder's voice tends to taper off. To track the variation of speech energy, similar to the F_0 features, a variety of energy-related range features, movement features, and slope features are computed from the energy contour.

- **Other Speakers' Overlaps**: Because multiple speakers interact in a meeting, the speech provided by other speakers affects the current speaker's floor control management. For example, when another speaker talks simultaneously to compete with or interrupt the current speaker, it is more likely that the current speaker will give up control of the floor. To model the impact of

[3]Some SUs are not complete sentences. For example, answers to questions can be noun phrases, or some sentences may be cut off due to replanning or contention for the control of the floor.

other participants' speech, we computed the summation of durations of other participants' spoken words that overlap with the current floor holder's speech within a 1.0 second window preceding the end of the SU. The obtained duration of the overlapped speech is normalized by the number of meeting participants to reduce the influence of the different number of meeting participants.

Compared with related research on using prosodic cues for turn change detection [23, 33], our prosody feature extraction based on the PPFE toolkit provides richer prosodic features that more comprehensively reflect prosodic properties.

5.3 Visual Features

Visual cues have been found to play an important role for floor control management. Based on previous studies of floor control management, we extract several visual features described below from a window preceding the SU boundary with a length of 1.0 second that was estimated on the development data set, which will be described in Section 6.

Kendon [20] observed that the termination of gestures by the person controlling the floor is often used for yielding control of the floor. When a speaker is still making gestures at the end of an SU, he/she may still intend to continue speaking and therefore will maintain control of the floor after the end of the SU. If some other speakers make hand gestures, such as raising their hands, they may be requesting control of the floor. Such gestures from other speakers may trigger the current floor holder to yield control. The gesture features are computed as follows:

- **HOLDER_GES_TIME:** is the percentage of the floor holder's gesturing time before each SU ending boundary. If the current floor holder spends a considerable percentage of time gesturing at the end of an SU, he/she may be more likely to continue controlling the floor after the end of the SU.

- **OTHERS_GES_TIME:** is the percentage of other participants' gesturing time before each SU ending boundary. The value is normalized based on the number of meeting participants. This feature is expected to account for other meeting participants' floor grabbing gestures, which may trigger the current floor holder to yield control of the floor.

Previous studies [2, 19] suggest that gaze plays an important role in floor control management. The person who is the gaze target of the current floor holder is more likely to be the next floor holder when the floor control shifts. A special event, mutual gaze break, which was described in Section 2, has been reported to appear in dialogs and meetings [28]. The gaze features are computed as follows:

- **HOLDER_GAZE_TIME:** is the percentage of the floor holder's gazing time at other participants, rather than at other locations, during the window before each SU boundary. When the current floor holder does not gaze at any other participant, it is less likely that he/she intends to yield control of the floor.

- **OTHERS_GAZE_TIME:** is the percentage of other participants' time gazing at the current floor holder before each SU boundary. The value is normalized based on the number of participants to reduce the influence of the number of participants on this value. To grab control of the floor more smoothly, the subsequent floor holder will often look at the current floor holder in order to establish eye contact.

- **NUM_MUTUAL_GAZE:** is the number of mutual gazes established between the current floor holder and other meeting participants. Mutual gaze patterns are important "devices" used by the current and next floor holders to coordinate the shift of floor control. Therefore, existence of mutual gaze patterns close to the end of an SU should provide evidence about a floor control shift. Using the gaze annotations, we are able to identify mutual gaze between the current floor holder and other participants (i.e., time intervals when the current floor holder and a participant look at each other momentarily).

6. EXPERIMENT

6.1 Data Setup

Given the limited size of the multimodal corpus used in our experiment, we first split the entire data set containing 14 speakers into a held-out development set and a data set for training and testing. The data for the speaker C in the Jan07 meeting and the speaker F in the Mar18 meeting comprise the development set, which is used to set some of the parameters needed by the machine learning algorithms we use. In order to test on more instances, over the data from the remaining 12 speakers, we conduct a leave-one-out evaluation procedure, that is, we iteratively use data from 11 speakers to train the statistical model and test the obtained model on the remaining speaker's data. Among 1749 inter-SU boundaries used for evaluation (from 12 speakers), 65.52% have *Keep* FCS labels. Therefore, baseline performance can be calculated by predicting that there is always a *Keep* FCS, giving an ERR of 100% and a CER of 34.48%.

6.2 Statistical Models

Floor control shift detection is based on three somewhat independent knowledge sources: lexical, prosodic, and visual cues. The task can be generalized as follows:

$$\hat{E} = arg \max_{E} P(E|W, F, V)$$

Given that E denotes the inter-SU floor control shift event sequence (**Keep** or **Change**), W denotes the corresponding lexical cues, F denotes the prosodic feature vector, and V denotes the corresponding visual features related to gesture and gaze, the goal is to find the floor control shift event sequence that has the greatest probability given the observed multimodal features.

Recently, conditional modeling approaches were successfully used in SU detection using audio cues [24], as well as multimodal cues [5]. We would expect that conditional modeling approaches would also be effective for combining audio and visual features for floor control shift detection. Hence, we use the Maximum Entropy (ME) [3] and Conditional Random Fields (CRFs)[21] approaches to build statistical models for floor control shift detection.

Compared to SU events, for which decisions are made for every word boundary, floor control shift decisions are made for every SU boundary. Hence, there is a greater problem

with data sparsity. Also, there is a greater variance among participants' behaviors for managing control of the floor. For example, participants in the three VACE meetings have different patterns of looking at subsequent floor holders during floor control shifts [6]. To address this variability, several machine learning approaches have been proposed and used in previous studies. For example, Levow [23] used boosting ensemble learning to predict turn changes. Boosting (and algorithms based on it) has been reported to give competitive performance on a wide variety of data sets in several experimental studies [9, 14]. Hence, we also used boosting ensemble learning to build a statistical model for floor control shift detection.

For Maximum Entropy (ME) modeling in our experiments, we use the Maxent toolkit designed by Zhang [40]. The L-BFGS parameter estimation method is used, with Gaussian-prior smoothing [8] to avoid overfitting. The Gaussian prior is estimated on the held-out development data set. The rationale behind the use of Gaussian priors is to force the learned parameters to be distributed according to a Gaussian distribution. This prior expectation penalizes parameters that drift away from their mean prior value (the mean is usually 0). Because the Maxent toolkit is more effective with categorical features, the numeric features, e.g., prosodic features about $F0$, were converted to categorical features using Fayaad and Irani's MDL discretization method [11], which was implemented in the WEKA [39] machine learning package. The following ME models were implemented for our experiment:

- **ME Speech Model:** uses the lexical and prosodic features described in Section 5.1 and Section 5.2, respectively.
- **ME Visual Model:** uses the gesture and gaze features described in Section 5.3.
- **ME Multimodal Model:** uses all of the lexical, prosodic, and visual features.

For Conditional Random Fields (CRFs) modeling, we use the Java based package Mallet [26]. Similar to training an ME model, Gaussian smoothing is utilized to avoid overfitting. The Gaussian prior is estimated on the held-out development data set, and the numeric features are converted to categorical features using WEKA [39]. Viterbi decoding is used for these models. The following CRF models were implemented for our experiment:

- **CRF Speech Model:** uses the lexical and prosodic features described in Section 5.1 and Section 5.2, respectively.
- **CRF Visual Model:** uses the gesture and gaze related features described in Section 5.3.
- **CRF Multimodal Model:** uses all of the lexical, prosodic, and visual features.

For Boosting ensemble learning modeling, we use the AdaBoost.M1 boosting method implemented in WEKA [39]. Parameters of the AdaBoost model, e.g., the ensemble learning iteration number, are estimated on the held-out development data set. The following AdaBoost models were implemented for our experiment:

- **AdaBoost Speech Model:** uses the lexical and prosodic features as described in Section 5.1 and Section 5.2, respectively.
- **AdaBoost Visual Model:** uses the gesture and gaze related features described in Section 5.3.
- **AdaBoost Multimodal Model:** uses all of the lexical, prosodic, and visual features.

6.3 Results

Table 3 reports on the performance of the speech, visual, and multimodal models using the three machine learning approaches and various feature configurations. The first row represents the baseline error rates obtained by always predicting that there is no floor control shift around the SU boundaries.

Using only speech features, the three models (ME, CRF, and AdaBoost.M1) all achieve significantly lower error rates than the baseline performance. When using speech features, the CRF modeling approach achieves a better performance than the ME modeling approach, and the AdaBoost.M1 modeling approach achieves further improvement over the CRF but this difference is not statistically significant.

The accuracy of floor control shift detection using speech features is fairly low. This is not surprising because the floor control detection is quite challenging for several reasons. First, the underlying scheme to control the floor in conversations is quite varied; floor control is based on the interactions among participants. Therefore, a speaker may produce different floor control patterns depending on the people involved in the conversation. Second, in our current study, we only utilized speech features extracted locally previous to SU ending boundaries. Although high-level pragmatic information (e.g., topic structure) is likely to play an important role for making decisions about floor control shifts, this high-level information is challenging to model given the small amount of data available in this study. Third, since the classification is conducted on inter-SU boundaries, the number of available training instances is small, creating a challenge for building our models.

Using only visual features, the three models (ME, CRF, and AdaBoost) all achieve significantly lower error rates than the baseline, but somewhat higher than the corresponding speech models. Clearly, eye gaze and hand gesture features provide cues that are effective for floor control shift detection. Their usefulness is consistent with previous psycholinguistic and conversation analytic research findings.

When combining speech and visual cues together, the three modeling approaches (ME, CRF, and AdaBoost) all achieve significantly improved performances (sign-test using $p < 0.05$) over the models using only speech cues. This suggests that the combination of audio and visual information significantly improves the accuracy of floor control shift detection compared to either model alone. Using the combined features, both CRF and AdaBoost modeling approaches achieve significantly lower error rate than ME (sign-test using $p < 0.05$). However, although AdaBoost modeling approach achieves a lower error rate than CRF, this error reduction is not statistically significant.

For ME and AdaBoost modeling approaches, the models trained on visual features have higher DEL error rates but lower INS error rates compared to the models trained only on speech features. However, the CRF visual model

Model	DEL (%)	INS (%)	ERR (%)	CER (%)
Baseline	100.00	0.00	100.00	34.48
ME Speech	58.37	26.70	85.07	29.35
ME Visual	67.50	21.72	89.22	30.78
ME Multimodal	54.73	26.04	**80.77**	**27.86**
CRF Speech	63.85	18.41	82.26	28.38
CRF Visual	67.99	20.07	88.06	30.38
CRF Multimodal	56.55	19.90	**76.45**	**26.37**
AdaBoost Speech	51.91	27.53	79.44	27.40
AdaBoost Visual	66.34	19.57	85.90	29.63
AdaBoost Multimodal	50.58	23.22	**73.80**	**25.46**

Table 3: Floor control shift detection using speech features, visual features, and the combination of both types of features by Maxent, CRF, and AdaBoost models. (Bold fonts show that the multimodal model has a statistically significantly lower error rate than the corresponding speech model for floor control shift detection ($p < 0.05$).)

has higher DEL and INS error rates compared to the CRF speech model. This may be due to the fact that CRF uses information from the sequence; however, visual cues are not always available in the context. For all three modeling approaches, the models trained on both speech and visual features achieve the lowest DEL error rates among models trained on speech, visual, and both types of features. Therefore, some floor control shifts missed by speech models are compensated for based on visual cues.

Table 4 reports on FCS detection performance using majority voting among FCS models built using ME, CRF, and AdaBoost approaches. Compared to the best multimodal model (AdaBoost), the majority voting of the ME, CRF, and AdaBoost models produces a lower error rate (although it is not statistically significant according to the sign-test ($p < 0.05$)). The voting approach achieves the lowest INS error rates among models built using ME, CRF, and AdaBoost approaches.

Model	DEL (%)	INS (%)	ERR (%)	CER (%)
ME	54.73	26.04	80.77	27.86
CRF	56.55	19.90	76.45	26.37
AdaBoost	50.58	23.22	73.80	25.46
Voting	55.89	16.58	72.47	25.00

Table 4: Voting using the multimodal ME, CRF, and AdaBoost models

7. DISCUSSION

In this paper, we described our investigations on using visual cues (i.e., eye gaze and hand gesture) to automatically detect floor control shifts in multi-party conversations. Based on knowledge about a variety of multimodal cues (e.g., lexical, prosodic, and nonverbal cues) and their roles in signaling floor control shifts, we designed a set of multimodal features and implemented models for floor control shift detection. The extracted features were useful for predicting floor control shifts. Two conditional modeling approaches (ME and CRF) were utilized to implement the models. The CRF modeling approach outperformed ME for floor control shift detection. In addition, to address the variance of speaker's multimodal behaviors in floor control shifts, the boosting ensemble learning approach was used and was found to produce the lowest error rate in all conditions.

To our knowledge, this study is the first to utilize visual cues in an automatic detection task related to floor control and turn-taking structure. Our experimental results on the VACE multimodal meeting data suggest that visual cues play an important role for predicting floor control shifts in that they significantly improve the accuracy of floor shift detection when compared to a speech-only model. However, there is more work that needs to be done. Increasing the amount of training data and integrating longer-range features will be important for improving performance further. It is also important to examine the effect of using signal processing techniques to automatically derive gaze and gesture features. In addition, we will extend this study done using transcribed SUs to one using estimated SUs.

Acknowledgments

The authors thank all of our VACE team members for their efforts in producing the VACE multimodal meeting corpus. 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 of ARDA.

8. REFERENCES

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

[2] G. Beattie. The regulation of speaker turns in face-to-face conversation: Some implications for conversation in sound-only communication channels. *Semiotica*, 34:55–70, 1981.

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

[4] A. Cassell, T. Nakano, T. Bickmore, C. Sidner, and C. Rich. Non-verbal cues for discourse structure. In *Proceedings of the Conference of Annual Meeting on Association for Computational Linguistics Linguistics (ACL)*, pages 106–115, Toulouse, France, 2001.

[5] L. Chen. *Incorporating Nonverbal Features into Multimodal Models of Human-to-Human Communication*. PhD thesis, Purdue University, West Lafayette,IN, August 2008.

[6] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Q. Huang, and F. Quek. A multimodal analysis of floor control in meetings. In *Proceedings of the Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, Washington, DC, USA, May 2006.

[7] L. Chen, T. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. Xu, J. Tu, Z. Huang, M. Harper, F. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. S. Huang. VACE multimodal meeting corpus. In *Proceedings of the Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, 2005.

[8] S. Chen and R. Rosenfeld. A Gaussian prior for smoothing maximum entropy models. Technical report, Carnegie Mellon University, 1999.

[9] T. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[10] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

[11] U. Fayyad and K. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8:87–102, 1992.

[12] L. Ferrer, E. Shriberg, and A. Stolcke. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, 2002.

[13] C. Ford and S. Thompson. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the managment of turns. In T. Ochs, Schegloff, editor, *Interaction and Grammar*. Cambridge Univ. Press, 1996.

[14] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1996.

[15] D. Gatica-Perez. Analyzing human interaction in conversations: A review. In *Proceedings of IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006.

[16] C. Goodwin. *Conversational Organization: Interaction Between Speakers and Hearers*. Academic Press, 1981.

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

[18] N. Jovanovic, R. Akker, and A. Nijholt. Addressee identification in face-to-face meetings. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, April 2006.

[19] A. Kalma. Gazing in trials - A powerful signal in floor appointment. *British Journal of Social Psychology*, 31:21–39, 1992.

[20] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[21] J. Lafferty, A. McCallum, and F. Pereira. Conditional random field: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

[22] E. L. Lehmann. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

[23] G. Levow. Turn-taking in mandarin dialogue: Interactions of tones and intonation. In *Proceedings of the SIGHAN Workshop*, 2005.

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

[25] J. Local and J. Kelly. Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies*, 9:185–204, 1986.

[26] A. McCallum. Mallet: A machine learning toolkit for language. `http://mallet.cs.umass.edu`, 2005.

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

[28] D. G. Novick. Models of gaze in multi-party discourse. In *Proceedings of Computer Human Interface (CHI) Workshop on the Virtuality Continuum Revisted*, Portland, OR, April 2005.

[29] D. G. Novick, B. Hansen, and K. Ward. Coordinating turn-taking with gaze. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1996.

[30] E. Padilha and J. Carletta. Nonverbal behaviours improving a simulation of small group discussion. In *Proceedings of the First International Nordic Symposium of Multi-modal Communication*, 2003.

[31] T. Rose, F. Quek, and Y. Shi. MacVissta: A system for multimodal analysis. In *Proceedings of the International Conference on Multimodal Interface (ICMI)*, 2004.

[32] H. Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organisation of turn taking for conversation. *Language*, 50:696–735, 1974.

[33] D. Schlangen. From reaction to prediction experiments with computational models of turn-taking. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 2006.

[34] 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.

[35] S. Strassel. *Simple Metadata Annotation Specification*. Linguistic Data Consortium, 5.0 edition, 2003.

[36] S. Thede and M. Harper. A second-order hidden markov model for part-of-speech tagging. In *Proceedings of the Conference of Annual Meeting on Association for Computational Linguistics Linguistics (ACL)*, Baltimore, MD, 1999.

[37] R. Vertegaal, G. Veer, and H. Vons. Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface*, 2000.

[38] A. Wichmann and J. Caspers. Melodic cues to turn-taking in english: Evidence from perception. In *Proceedings of the SIGDial Workshop on Discourse and Dialogue*, 2001.

[39] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.

[40] L. Zhang. Maximum Entropy Modeling Toolkit for Python and C++. `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`, 2005.