# Floor Holder Detection and End of Speaker Turn Prediction in Meetings

*Alfred Dielmann, Giulia Garau, and Hervé Bourlard*

Idiap Research Institute - Rue Marconi 19 - 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne - 1015 Lausanne, Switzerland
a.dielmann@gmail.com, giuliagarau@yahoo.com, herve.bourlard@idiap.ch

## Abstract

We propose a novel fully automatic framework to detect which meeting participant is currently holding the conversational floor and when the current speaker turn is going to finish. Two sets of experiments were conducted on a large collection of multiparty conversations: the AMI meeting corpus. Unsupervised speaker turn detection was performed by post-processing the speaker diarization and the speech activity detection outputs. A supervised end-of-speaker-turn prediction framework, based on Dynamic Bayesian Networks and automatically extracted multimodal features (related to prosody, overlapping speech, and visual motion), was also investigated. These novel approaches resulted in good floor holder detection rates (13.2% Floor Error Rate), attaining state of the art end-of-speaker-turn prediction performances.

**Index Terms**: multiparty conversation, floor control, speaker turn, non-verbal features, Dynamic Bayesian Network.

## 1. Introduction

This work automatically analyses multiparty conversations, predicting: which participant is currently holding the conversational floor (i.e. the owner of the current speaker turn), and when the current speaker turn is going to end. This is achieved by considering only non-verbal participant behaviours [1]. Several downstream applications would benefit from automatically detecting the current floor holder and predicting its change. Human computer interfaces, such as spoken dialogue systems and Embodied Conversational Agents, would be able to improve their engagement in a conversation. Floor control modelling can be exploited by mediated communication applications, such as: "virtual video directors", and teleconference multicasting systems. In addition it could be employed to facilitate speech understanding tasks, such as: automatic summarisation, topic detection, and automatic role recognition.

Sacks et al. [2] observed that during a conversation speakers usually talk one-at-a-time, i.e. speaker overlaps are common but brief. Speakers take turns while trying to minimise the gap or overlap between adjacent turns, so that fluent conversations are formed. Listeners are thus able to roughly predict the end of the current speaker turn in order to time their own start [3]. A turn taking model was formulated by Sacks et al. [2] in order to describe the floor control process. Each speaker turn is composed by one or more Turn Constructional Units (TCUs). These are grammatically and prosodically complete utterances, often marked by a lowering of pitch and energy towards their end [3]. TCUs are followed by Transition Relevance Places (TRPs), points in the conversation where conversational floor holder changes are more likely. TRPs often correspond to unfilled pauses, and are marked by turn-yielding cues [3, 4] such as "the speaker gazing back up to an interlocutor", gestures, and posture shifts. The literature on floor control and end-of-speaker-turn prediction initially focused on dyadic conversations, such as telephone conversations [5]. Only recently the interest shifted towards a more challenging task: modelling multiparty conversations such as meetings [6, 7]. Schlangen [5] investigated the use of syntactic and prosodic features for end-of-speaker-turn detection/prediction on telephone conversations; human end-of-turn prediction performances were also reported. Chen and Harper [6] investigated end-of-speaker-turn detection on VACE meetings using prosodic, lexical, syntactic, and visual cues. These feature sets were extracted from manual orthographic, gestural, and Visual Focus of Attention (VFoA) annotations. Turn detection was performed at the Sentence Unit level comparing three statistical models. De Kok and Heylen [7] addressed the end-of-turn prediction task on AMI meetings (at a frame level) using a similar set of manual annotations (Dialogue Acts, head gestures, and VFoA) along with prosodic features. A Conditional Random Fields model was adopted to this end.

At first glance, speaker turn detection may be confused with speech activity detection [8] and speaker diarization [9]. However the latter tasks provide a fine grained representation of the multiparty conversation (tailored for Automatic Speech Recognition applications), aimed at identifying the temporal boundaries of each: silence, word, and utterance. Instead speaker turn detection aims at a coarse grained representation where: each speaker turn frequently includes multiple utterances (TCUs) from the same speaker, often separated by long pauses (TRPs). Backchannels and feedbacks from other speakers are included into the current speaker turn [5], as they help regulating the turn taking process without being part of the main conversational exchange [2].

In this work two sets of experiments were performed on the AMI meeting corpus (Section 2). Unsupervised speaker turn detection was performed using a sequential approach (Section 3). Speech activity detection and speaker diarization offer fine grained conversation segmentations; probabilistic models and acoustic features are then employed to post-process them, forming speaker turns. A supervised approach for on-line joint floor holder detection and end-of-speaker-turn prediction was also investigated (Section 4). A Dynamic Bayesian Network model relates the current floor holder with a multimodal feature collection (including: prosody, overlapping speech, and visual activities). Compared to previous works, in particular [6, 7], the proposed approaches are novel in the following aspects:

- Floor holder detection and end-of-speaker-turn detection/prediction are performed jointly; previous systems focused only on speaker turn detection/prediction.
- Fully automatic approach; manual annotations in terms of words, gestures, and Dialogue Acts were previously adopted as observable features or to facilitate their extraction [5, 6, 7].

- We focus on non-verbal communication [1]: our aim is to develop text-independent low-level approaches that do not rely on manual or automatic orthographic transcriptions.
- Multiple participants are modelled jointly, rather than processing each speaker/microphone channel independently.
- Results are reported for the first time using Multiple Distant Microphones, a challenging but highly portable audio recording setup (which only requires a few table-top microphones).

## 2. Meeting Data and Annotations

Our experiments are based on the scenario subset of the AMI meeting corpus [10]: a collection of 138 meetings (72 hours) elicited using a scenario. Four meeting participants, playing different roles in a team, took a product development project from beginning to completion. The aim of this collection was to obtain a multimodal record of the complete communicative interaction between the meeting participants. To this end, three meeting rooms were instrumented with: 4 Individual Headset Microphones, 8 Multiple Distant Microphones forming a table-top microphone array, 4 close-up and 2 room-view video cameras. Manual orthographic transcriptions and Dialogue Act (DA) annotations are available for the entire collection [10].

**Manual floor annotation** 12 randomly selected meetings (4 for each recording site) were manually annotated in terms of speaker turn segments, also identifying the conversational floor holder for each segment. The resulting annotation showed a good human inter-annotator agreement [11] of $K = 0.9$: 90% better than one could have expected as simply due to chance.

**DA derived floor annotation** An alternative speaker turn / floor holder annotation was also derived from the reference Dialogue Act annotation (under the assumption of non-schismatic conversations). To this end, DA units such as backchannels and DA fragments which are fully included into longer DA segments are ignored (i.e. they do not constitute speaker turns [2]); DA segments are extended in order to include their trailing silences (i.e. the current speaker holds the conversational floor until someone grabs it); segments belonging to the same speaker are joined together. The resulting segmentation is considered as a proxy for the reference speaker turn / floor holder annotation. The agreement between the resulting proxy annotation and the 12 manually annotated meetings is $K = 0.9$. Therefore the DA derived floor annotation is comparable to the manual floor annotation, and can be used in its lieu. Note that the DA derived floor annotation is available for all the 138 AMI meetings.

## 3. Unsupervised Experiments

Unsupervised joint floor holder and end-of-speaker-turn detection were performed on the 12 manually annotated AMI meetings (Section 2) using a two step approach: acoustic segmentation, followed by segments' regrouping. During the first step a low-level segmentation of the conversation is obtained through Automatic Speaker Diarization (Section 3.0.1) or Speech Activity Detection (Section 3.0.2). The resulting segments are then post-processed and merged in order to form speaker turns. To this end, two alternative approaches were developed: Probabilistic Segment Filtering (Section 3.0.3) and Joint Maximum Cross-correlation Filtering (Section 3.0.4).

### 3.0.1. *Speech Activity Detection*

Speaker activities were estimated from each Individual Headset Microphone using the SHOUT toolkit [8]. Automatic detection of speech, silence, and audible non-speech (sound) is performed in five steps: acoustic features extraction (12 Mel Frequency Cepstral Coefficients and Zero Crossing Rate features); rough speech/non-speech segmentation using out of domain acoustic models; training of accurate recording specific speech/silence/sound models, using the initial speech/non-speech segmentation; merging of sound and speech models when they are found to be equivalent (according to the Bayesian Information Criterion); estimation of the final segmentation using Viterbi decoding.

### 3.0.2. *Automatic Speaker Diarization*

Automatic speaker diarization aims at identifying individual speaker interventions on a single track audio recording. This fully unsupervised audio segmentation technique is able to learn a statistical model for the voice of each speaker, without any prior knowledge about the number and the identities of the participants. The speaker diarization system adopted in our experiments is based on the Hidden Markov Model (HMM) agglomerative clustering approach proposed in [9]. 19 MFCCs are extracted from the raw audio recordings every 10 *ms.* and modeled using Gaussian Mixture Models (GMMs). An ergodic HMM with one state for each audio cluster is used to model the conversation, enforcing a minimum duration constraint of 2 seconds for each segment. Starting from a large number of audio clusters (30), the most similar ones (according to the BIC) are iteratively merged, keeping the overall number of GMM parameters constant. The merging process is stopped when the HMM likelihood starts decreasing (comparing successive iterations). The outlined diarization framework attained an overall cluster purity [9] of $k = 0.64$ on the mix of 4 Individual Headset Microphone signals (IHM-Mix), and a purity of $k = 0.52$ on beamformed Multiple Distant Microphones (MDM, Section 3.1).

### 3.0.3. *Probabilistic Segment Filtering*

The speaker diarization output (Section 3.0.2) results in a larger number of audio clusters (9–15) than the actual number of meeting participants ($n = 4$). The resulting segmentation includes: spoken segments uttered by a single meeting participant, interleaved with shorter segments characterised by non-vocal-sounds, noise, and overlapping speech. In order to recover the underlying speaker turn structure, it is desirable to remove all these short noisy segments, including them into their surrounding spoken segments. To this end we employed a Gaussian Naïve Bayes classifier trained on features such as: normalised segment length, proportion of the recording represented by that audio cluster label, proportion of the segment classified as voiced during pitch estimation (Section 4.0.1). Each meeting recording is processed with an individually trained classifier. Data from the 4 smallest and largest automatically detected audio clusters provided evidence for the two target classes (i.e. noisy and spoken segments). All audio segments are re-classified as spoken or noisy using the learned models, and then merged accordingly. Note that probabilistic segment filtering, requiring more than 4 input clusters, could not be applied to the Speech Activity Detection output (Section 3.0.1).

### 3.0.4. *Joint Maximum Cross-correlation (JMXC) Filtering*

Joint Maximum Cross-correlation features, initially proposed in [12] to address microphones' crosstalk in multichannel voice activity detection, can be employed to detect the most active meeting participant (i.e. speaker) $S_i, i \in [1, n = 4]$ on each automatically obtained audio segment $\Delta_t$:

Table 1: *Unsupervised floor holder (Floor Error Rate percentage) and end-of-speaker-turn detection (Precision, Recall, and F1-score) performances on 12 manually annotated meetings.*

| Unsupervised Setup | FER | Prec. | Rec. | F1 |
|---|---|---|---|---|
| MDM Diarization | 57.1 % | 0.23 | 0.55 | 0.32 |
| + Prob. segment filtering | 48.0 % | 0.30 | 0.34 | 0.32 |
| IHM-Mix Diarization | 34.5 % | 0.27 | 0.51 | 0.35 |
| + Prob. segment filtering | 28.5 % | 0.37 | 0.42 | 0.39 |
| + JMXC filtering | 20.4 % | 0.43 | 0.40 | 0.41 |
| IHM Speech Activity D. | 17.3 % | 0.35 | 0.47 | 0.40 |
| + JMXC filtering | 16.9 % | 0.41 | 0.50 | 0.45 |

$$\Xi_i(\Delta_t) = \sum_{\substack{j=1 \\ j\neq i}}^{j=n} \log_{10}\left(\xi_{ij}(\Delta_t)\right) = \sum_{\substack{j=1 \\ j\neq i}}^{j=n} \log_{10}\left(\frac{\max \phi_{ij}(\Delta_t)}{\phi_{jj}(\Delta_t)}\right).$$

$\phi_{ij}(\Delta_t)$ represents the cross-correlation (estimated on raw acoustic signals) between Individual Headset Microphone channels $i$ and $j$; $\xi_{ij}$ estimates to what extent speaker $S_j$ is responsible for the cross-correlation peak $\max \phi_{ij}(\Delta_t)$ relative to channel $i$ and speaker $S_i$. Participant $S_i$ is speaking if $\Xi_i(\Delta_t) > 0$ (a geometric interpretation for $\xi_{ij}$ and $\Xi_i$ can be found in [12]). Therefore we assume $L_t = \arg\max_i\left(\Xi_i(\Delta_t)\right)$ as the most active speaker in $\Delta_t$. Adjacent segments $\Delta_{t-1}$, $\Delta_t$ sharing the same speaker label $L_{t-1} = L_t$ are then joined together, aiming at reconstructing the reference speaker turn segmentation.

### 3.1. Experimental Setup and Numerical Results

Experimental results of unsupervised floor holder detection are shown in table 1. They are reported in terms of *Floor Error Rate*, intended as the percentage of the recording length where the conversation floor holder was incorrectly detected. Table 1 also reports the end-of-speaker-turn detection performances in terms of: *precision* (probability that an automatically detected end-of-speaker-turn corresponds to a reference turn-end with a tolerance of ±0.5 seconds); *recall* (probability that a reference turn-end is automatically predicted); and *F1-score* (unweighted precision and recall harmonic mean).

Numerical experiments were performed on three different configurations: speaker diarization (Section 3.0.2) of the microphone array beam-forming output (applying J. Ajmera *Beamformit 2.0* toolkit to the 8 Multiple Distant Microphone acoustic signals); speaker diarization of the 4 Individual Headset Microphone channels Mix; Speech Activity Detection (Section 3.0.1) using the 4 Individual Headset Microphones. On the MDM setup, probabilistic segment filtering improves floor holder detection but not the overall end-of-speaker-turn detection performances (when compared to the baseline MDM diarization system). Instead on the IHM-Mix diarization setup, probabilistic segment filtering is effective on both tasks, resulting in a 6% FER and a 4% F1-score absolute improvement. JMXC filtering clearly outperforms probabilistic segment filtering (14% FER and 6% F1-score improvement) on the IHM-Mix diarization setup, being also effective on the IHM Speech Activity Detection (SAD) setup.

Although the IHM-SAD configuration provides the best detection performances on both tasks, by requiring access to each IHM channel and prior knowledge about the total number of speakers, this is the most constrained setup. In contrast probabilistic segment filtered MDM diarization, not only results in the most unobtrusive recording condition, but also avoids prior assumptions on the meeting setup.

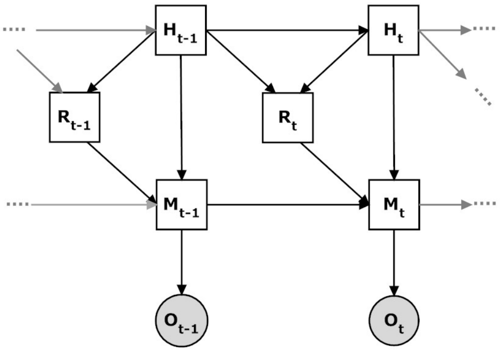

Figure 1: *DBN model for joint floor holder detection and end-of-speaker-turn prediction; discrete hidden random variables $H, R, M$ are represented by unshaded square nodes; observable feature vectors $\mathbf{O}$ correspond to shaded circles.*

## 4. Supervised Experiments

Supervised experiments of joint floor holder detection and end-of-speaker-turn prediction were performed by modelling a collection of multimodal features (Section 4.0.1) through a generative Dynamic Bayesian Network (DBN) model (Section 4.0.2). While sequential unsupervised experiments of section 3 aim at off-line processing; the supervised framework outlined in this section is suitable for on-line applications. To this end, the use of looking forward features such as pauses was avoided [5, 7]. The Viterbi decoding of the speaker independent DBN model runs twice faster than realtime, on a single core processor with 1Gb of memory.

#### 4.0.1. Multimodal Features

Three feature families were automatically extracted every 30 *ms.* from each Individual Headset Microphone and each individual close-up camera:

**Prosodic features** including pitch contour F0 (estimated using the *entropic get_f0* tool), Root Mean Square signal Energy, and syllabic Rate Of Speech (ROS) directly estimated from the acoustic signals without a transcription of what was said [13].

**Overlapping speech features** based on Joint Maximum Cross-correlation $\Xi_i(\Delta_t)$, $i = 1, ..., 4$ (Section 3.0.4) were extracted from non-overlapping 30 *ms.* long audio segments $\Delta_t$.

**Participant visual activities** including motion intensities and X,Y coordinates of the center of motion, were estimated from the luminance differences between adjacent video-frames.

The resulting feature sets are concatenated in a single multidimensional observable feature vector (early feature integration).

#### 4.0.2. Dynamic Bayesian Network Model

The DBN model depicted in figure 1 was adopted to predict the sequence of floor holders from the three multimodal feature families outlined in section 4.0.1. This ergodic two-level Hidden Markov Model represents the sequence of speaker-turn-holders through the Markov chain constituted by nodes $H_{0:T}$, which is responsible for a second hidden Markov chain formed by sub-state nodes $M_{0:T}$. Therefore each speaker turn $H_{t1:t2} : H_{t1} = H_{t1+1} = ... = H_{t2-1} = H_{t2}$; $t_1 < t_2$; $t_1, t_2 \in [0, T]$; $|H| = 4$ is decomposed into a sequence of sub-

Table 2: *Supervised floor holder detection and end-of-speaker-turn prediction, testing our DBN model and 4 different feature setups (Section 4.0.1) on 12 manually annotated AMI meetings. All results are significantly different (at a confidence level of p=0.001) according to the McNemar's significance test.*

| Feature Setup | FER | Prec. | Rec. | F1 |
|---|---|---|---|---|
| F0, Energy | 13.8 % | 0.56 | 0.42 | 0.48 |
| F0, Energy, Visual | 15.9 % | 0.52 | 0.36 | 0.42 |
| **F0, Energy, ROS** | **13.2** % | **0.57** | **0.44** | **0.50** |
| F0, Energy, ROS, JMXC | 14.0 % | 0.56 | 0.39 | 0.46 |

states $M_{t1:t2}$, aiming at modelling the temporal evolution of the current speaker turn $H_{t1:t2}$. Each sub-state $M_t$, $t \in [t_1, t_2]$ generates a single observable feature vector $\boldsymbol{O_t}$. Note that the mapping between sub-states $M$ and continuous feature vectors $\boldsymbol{O}$ is implemented through a 2-component GMM. A total of 12 sub-states ($|M| = 12$), shared by different floor holders, was adopted in our experiments (Section 4.1). Model parameters, including: prior probability vectors, transition matrices, sub-states, and GMMs; are learned from DA derived annotations (Section 2) during model's training. The deterministic binary reset node $R_t$ aims at reinitialising the sub-state variable $M_t$ when a floor holder change (i.e. the end of a speaker turn) is predicted. $R_t$ is set to zero during a speaker turn ($H_{t-1} = H_t$). A floor holder change ($H_{t-1} \neq H_t$) triggers the reset node $R_t = 1$, forcing the hidden sub-state node $M_t$ to be reinitialised in according to its prior probability distribution.

### 4.1. Experimental Setup and Numerical Results

The DBN model outlined in section 4.0.2 was trained on 123 AMI scenario meetings, employing the DA derived proxy annotation in terms of speaker turns (Section 2). The resulting model was tested on the same set of 12 manually annotated meetings used during the unsupervised experiments of section 3.1. Note that 3 AMI meetings were held out from the training set for hyperparameter optimisation. Joint floor holder detection and end-of-speaker-turn prediction performances are reported in table 2, using the evaluation metrics outlined in section 3.1 and the manual floor holder annotations of section 2. This novel generative approach infers the current floor holder from the multimodal feature vector and the DBN internal state, predicting for every 30 ms. if the current speaker turn ends at that instant.

Different feature setups were investigated following a forward search feature combination scheme (Table 2). Pitch and energy provided good floor holder detection performances, outperforming the unsupervised IHM-SAD setup (Table 1). However the adoption of automatically extracted visual features did not result in further improvements. Pitch, energy, and syllabic Rate Of Speech, proved to be the best feature combination on both tasks. Their adoption resulted in a low Floor Error Rate (13.2%) and in the best F1 end-of-turn prediction score ($F1 = 0.5$), which favourably compares to the state-of-the-art ($F1 = 0.48$) [7]. This feature set successfully fulfils the two-fold task of modelling participants' speech activities and Turn Constructional Unit prosodic completion [3].

## 5. Conclusions

In this paper two automatic systems for the joint floor holder detection and end-of-speaker-turn detection/prediction on multiparty conversations were investigated. The first approach is fully unsupervised: speaker diarization and speech activity detection outputs are post-processed using probabilistic models and cross-correlation measures, in order to detect speaker turns. The second approach aims at supervised on-line end-of-turn prediction. A multimodal feature set and a DBN model are employed to detect (every 30 *ms.*) the current floor holder, and to predict when the current speaker turn is going to end. Moreover we showed that the manual floor holder annotation can be reliably derived from the reference Dialogue Act annotation, facilitating the development of supervised approaches.

Differently from previous works, both approaches are fully automatic (i.e. no manually derived annotations are used during feature extraction) and model all participants/microphone channels jointly; unsupervised speaker turn detection was also performed on Multiple Distant Microphones. Numerical experiments on the AMI meeting corpus showed that the DBN based supervised system outperforms the unsupervised approach. The current floor holder can be accurately detected with a Floor Error Rate as low as 13.2%. The end of the current speaker turn can be reliably predicted from prosodic cues, attaining state of the art performances.

## 6. Acknowledgements

This work was supported by the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2).

## 7. References

[1] D. Gatica-Perez, "Automatic Nonverbal Analysis of Social Interaction in Small Groups: a Review," *Image and Vision Computing, Special Issue on Human Naturalistic Behavior*, vol. 27, pp. 1775–1787, Nov. 2009.

[2] H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language*, vol. 504, pp. 696–735, 1974.

[3] E. Padilha and J. Carletta, "Nonverbal Behaviours Improving a Simulation of Small Group Discussion," in *Proc. of nordic symposium on multimodal communication*, Nov. 2003, pp. 93–105.

[4] V. Petukhova and H. Bunt, "Who's next? Speaker-Selection Mechanism in Multiparty Dialogue," in *Proc. of D'aholmia Workshop on the Semantics and Pragmatics of Dialogue*, June 2009.

[5] D. Schlangen, "From Reaction to Prediction Experiments with Computational Models of Turn-Taking," in *Proc. of Interspeech*, Sept. 2006.

[6] L. Chen and M. P. Harper, "Multimodal Floor Control Shift Detection," in *Proc. of ICMI-MLMI*, Nov. 2009.

[7] I. de Kok and D. Heylen, "Multimodal End-of-Turn Prediction in Multi-Party Meetings," in *Proc. of ICMI-MLMI*, Nov. 2009.

[8] M. Huijbregts, "Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled," PhD Thesis, Twente University, 2008.

[9] J. Ajmera, "Robust Audio Segmentation," PhD Thesis, Ecole Polytechnique Federale de Lausanne, 2004.

[10] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[11] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Education and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[12] K. Laskowski and T. Schultz, "A Geometric Interpretation of Non-Target-Normalized Maximum Cross-channel Correlation for Vocal Activity Detection in Meetings," in *Proc. of NAACL-HLT*, April 2007, pp. 89–92.

[13] N. Morgan and E. Fosler-Lussier, "Combining Multiple Estimators of Speaking Rate," in *Proc. of IEEE ICASSP*, May 1998.