INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer Meshorer, Peter A Heeman

Center for Spoken Language Understanding
Oregon Health & Science University, Portland, Oregon,USA
tmeshorer@hotmail.com, heemanp@ohsu.edu

Abstract

This paper explores using a summary of past speaker behavior to better predict turn transitions. We computed two types of summary features that represent the current speaker's past turn-taking behavior: relative turn length and relative floor control. Relative turn length measures the current turn length so far (in time and words) relative to the speaker's average turn length. Relative floor control measures the speaker's control of the conversation floor (in time and words) relative to the total conversation length. The features are recomputed for each dialog act based on past turns of the speaker within the current conversation. Using the switchboard corpus, we trained two models to predict turn transitions: one with just local features (e.g., current speech act, previous speech act) and one that added the summary features. Our results shows that using the summary features improve turn transitions prediction.

Index Terms: turn taking, conversation, speaker transition

1. Introduction

Turn management is an important component of everyday conversations. Studies on turn management in human to human conversation [1, 2] suggest that, to minimize gaps between turns and speaker overlap, listeners anticipate turn transitions. To anticipate transitions, conversants are believed to mainly use features that are derived from the last few utterances of the speaker: syntactic [1, 3], prosodic [4, 5, 6], and pragmatic [7]. To better engage in conversation with humans, turn management components of spoken dialogue systems (SDS) have also evolved from using simple thresholds on the silent time to training machine learning models [8] on local features (syntactic and prosodic).

While most of the current work suggests that listeners use features derived locally from the speaker's current utterance, this paper investigates whether features representing a summary of past speaker behavior can help. The suggested features are computed over multiple past turns of the current speaker. The features measure the *relative turn length* of the current turn and the *relative floor control* of the current speaker. We believe that the summary features represent an evolving model of the other conversant. For example, speakers who typically use long turns will likely use long turns in the future. Moreover, speakers with more control of the conversation floor will be less likely to yield the turn. As the conversational image of the speaker evolves with the conversation, the other conversant might adjust their turn taking behavior in response.

To test the effectiveness of the summary features, we used the NXT version of the Switchboard corpus [9, 10] to train random forest models [11]. We created two baseline models that only use local features: current dialog act, and current and previous dialog acts. We also trained a model on the summary features as well as a model that included both the local and the summary features. Our results show that using only the summary features improves prediction performance against the model that includes only the last dialog act, in both the area under the curve (AUC), 0.65 vs 0.63, and F1, 66.42% vs 54.97%. In addition, the model trained on all of the features (summary and local features) performed better than the local features model in both AUC, 0.82 vs 0.79, and F1, 74.87% vs 74.08%. The results show that using the summary features can help predict turn transitions.

The paper is organized as follows: Section 2 presents related work. Section 3 introduces the local and summary features. Section 4 describes the experiment. Section 5 shows the results obtained by training random forest models with and without the summary features. Finally, in Section 6 we present our conclusion.

2. Related Work

This section covers the related work in both human-human conversations (conversation analysis and psycholinguistics) and human-machine conversations (spoken dialogue systems).

2.1. Human-Human Conversations

Duncan [12] argued that speakers signal when they want the listener to take the turn and presented six signals used by the speaker to accomplish this: intonation, drawl on the final syllable, body motion, sociocentric sequence, drop in pitch or loudness, and syntax. Kendon [13] added gaze as a signal to turn transition. Our summary features complement the set of signals as suggested by [12].

Turn allocation was introduced in the seminal work by Sacks, Schegloff, and Jefferson [1], who observed that conversations are "one speaker at a time" and gaps between turns as well as speaker overlaps are kept to a minimum. To satisfy these constraints, Sacks et al. suggested an ordered set of rules for turn allocation: (a) current speaker selects the next conversant; (b) if the current speaker did not select, any of the listeners can self select; or (c) if neither of the previous two cases apply, the current speaker continues. For the first rule, Sacks et al. suggested that the current speaker uses adjacency pairs as the main apparatus for selecting the next speaker. Hence, we recognized the importance of dialog acts in turn allocation and chose them as the atomic turn components. In addition, our work might impact the second rule, in which the conversant self selects. While Sacks et al. suggested that the first starter is the next speaker, we suggest that a conversant might use the conversational image of the speaker and of themselves when self selecting. For example, a controlling speaker (with a high relative floor control score) has a better chance to gain control of the conversation floor when self selecting. The work on turn bidding is also re-

http://dx.doi.org/10.21437/Interspeech.2016-1409

lated [14], which suggested that each conversant measures the importance of their utterance when negotiating the right to the conversation floor.

In addition to the turn allocation system, Sacks et al. also suggested that turn construction units (TCU) should support projection of turn ends by the participants. The projectability attribute was later extended to other features of the speaker's utterance: (syntactic [1], prosodic [4] and pragmatic [4, 7]). Our work augments the local utterance features with summary features that can be used to improve projectablity.

Entrainment was presented in [15], which suggested entrainment of endogenous oscillators in the brains of the speaker and the listener on the basis of the speaker syllabus production. In their study, the speaker and the listener are counter phased such that speech overlaps and gaps are minimized. Although our work does not imply cyclic synchronization between speaker and listener, we do suggest that each conversant creates a conversation image of the other conversant and uses it during turn transition.

The importance of using dialog acts was emphasized by a very recent study of Garrod and Pickering [16]. The study suggested that turn production is a multi-stage process in which the listener performs simultaneous comprehension of the existing turn as well as production of the new turn content. They suggested that the first step in the process is dialog act recognition, which is done as soon as possible and acts as the basis for the listener's turn articulation and production. In our study we use dialog act as the main turn component.

2.2. Human-Computer Conversations

As recent advances in machine learning [17] reduce speech recognition error rates, the problem of turn taking in SDS rises in importance. Traditional SDS systems use a simple silence timeout approach to trigger turn transitions. This creates three issues [18]: first, the model might not be robust enough in a noisy environment (for example when driving); second, if the timeout is too short, the system might detect intra turn pauses (for example, the user pausing to think) as a turn transition and will cut into user's turn; and third, if the timeout is too long the system will wait too long to take the turn, resulting in large gaps between turns.

Recent studies tried to improve over the simple threshold model by using machine learning to train models based on features derived from the last utterance. As different studies use a variety of features, we will outline those that used counting features that are close to the summary features.

Arsikere et al. [19] focus on utterance segmentation in the context of incremental dialog system. Using the switchboard corpus, they used a decision tree algorithm to decide if a word is utterance final using various features and in particular the number of words in the turn so far. The usage of count features improves precision by 10% but has very low recall (7%), which might have occurred, according to the author, from turns with only one word.

Gravano and Hirschberg [8] used the Columbia games corpus in order to study the effectiveness of different turn transition cues. The authors define inter pausal units (IPU) as a maximum sequence of words surrounded by silence of more than 50 ms. A turn is the longest sequence of IPUs by the same speaker. One of the features studied is IPU duration in ms as well as number of words. As in our findings, the authors found that long IPUs are a good indication of upcoming turn changes (long IPUs might correlate with a speaker passing its average turn length). Moreover, as we show in Section 5, the authors found that combining multiple cues leads to better accuracy.

Raux and Eskenazi [20] performed a comprehensive study on features that inform turn changes. The study found that timing features, like turn duration and number of pauses, have relatively strong predictive power. While Raux and Eskenazi use features of the current turn, in our study we use the timing features for the turns that have occurred so far in the current conversation.

In more recent study, Nishitha and Rodney [21] used a model based on N grams of dialog acts to predict turn transitions. They trained a decision tree model using the switchboard data and tested bigram, trigram and 4 grams models of dialog acts with and without speaker id. They achieved an F1 measure of 0.67 for the trigram model. In this paper we based our baseline models on bigrams and trigrams of dialog acts. We also mapped the switchboard dialog acts from 148 dialog acts down to 9 in order to reduce data dimensionality. The prediction performance of our baseline model is comparable to their results.

3. Local and Summary Features

This section defines the local and summary features. The local features are based on pragmatics and consist of the current and previous dialog acts. The summary features are based on measurements of each speaker's behavior over over the preceding turns in the dialogue.

3.1. Local Features

We define a conversation as a sequence of dialogue acts $d_1 \dots d_N$, where d_i is uttered by speaker s_i. We write this as the following sequence:

$$\dots s_{i-2}, d_{i-2}, s_{i-1}, d_{i-1}, s_i, d_i \dots \qquad (1)$$

We denote whether there was a turn transition with y_i. A turn transition occurs when the speaker s_i is different from speaker s_{i-1}. Hence, (1) can be also be viewed as a sequence of dialog acts d_i followed by turn transitions y_i:

$$\dots d_{i-2}, y_{i-1}, d_{i-1}, y_i, d_i, y_{i+1} \dots \qquad (2)$$

In our first baseline model, we try to predict the turn transition value y_{i+1} based only on the latest dialog act d_i. In our second baseline model, we try to predict turn transition y_{i+1} based on the latest two dialog acts: d_{i-1} and d_i.

3.2. Summary Features

As discussed in the introduction, we introduce two types of summary features in this paper: relative turn length (rt_i) and relative floor control (rc_i). These features are used in predicting whether there is a change in speaker y_{i+1} after dialogue act d_i.

To compute the summary features, at dialogue act d_i, we denote S_i to be the set of complete turns of speaker s_i that are prior to the turn that d_i is in. Let t_i represent the turn so far that d_i is in, up to d_i but no subsequent dialogue acts. Let length(t) be the length of a turn or a partial turn in seconds (or words). To compute the *relative turn length* of turn t_i we first compute the average length of all the turns in S_i

$$avg\_t_i = \frac{\sum_{t \in S_i} length(t)}{|S_i|} \qquad (3)$$

The *relative turn length* summary feature of t_i, denoted as rt_i, measures the percent of the length of the turn t_i so far, relative to the average turn length up to t_i of the current speaker s_i (but not including t_i).

$$rt_i = \frac{length(t_i)}{avg\_t_i} \tag{4}$$

Note that we calculate two versions of rt_i: in seconds and in words. The purpose of rt_i is to let the listener, in predicting turn changes, take into account whether the current speaker is exceeding his or her average turn length.

The *relative floor control*, denoted as rc_i, measures the percent of time in which the current speaker controlled the conversation floor up to d_i. We again define S_i as above, and we define L_i to be the turns of the other conversant (the listener of d_i). We first compute the conversation length up to d_i denoted as c_i which excludes inter-utterance pauses.

$$c_i = \sum_{t \in S_i \cup L_i} length(t) \tag{5}$$

To compute relative floor control at d_i, we divide the floor time of the speaker s_i up to turn t_i by c_i:

$$rc_i = \frac{\sum_{t \in S_i} length(t)}{c_i} \tag{6}$$

Note that we calculate rc_i in seconds and in words. Participants can use the relative floor control as a means to determine if one speaker is controlling the conversation - a controlling speaker will probably be less inclined to give up the floor.

We use these two summary features in the *summary model* and *full model*, as described in the next section.

4. Evaluation

Figure 1 shows the experiment data pipeline. Data is imported from the NXT switchboard corpus [9] into a graph database [22]. Figure 2 shows the data structure as it is represented inside the graph database. For each conversation, the conversation entities (words, dialog acts and turns) are represented as edges between time points, which are represented as vertices. The structure leads to a direct computation of the summary features using the graph query language.

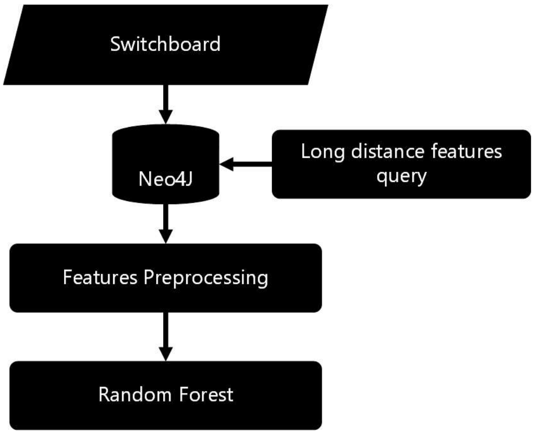

Figure 1: The experiment data pipeline

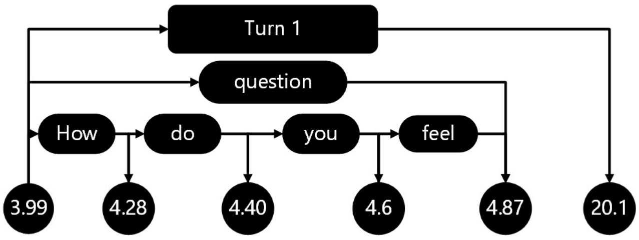

Figure 2: Conversation graph data model

Switchboard dialog acts	Dialog act classes
sd,h,bf	statement
sv,ad,sv@	statement - opinion
aa,aar̂	agree accept
%.%-,%@	abandon
b,bh	backchannel
qy,qo,qh	question
no,ny,ng,arp	answer
+	+
o@,+@	NA

Table 1: Mapping from dialog act to dialog act class

After computing the summary features, we perform the following data transformation:

- We exclude 11 dialogue acts that were coded in Switchboard as “other”.
- Since we believe that it takes a certain amount of time to build a stable conversational image, in evaluating our model, we removed all turns that occurred in the first part of each conversation. For this paper, we used an estimate of 120 seconds. This reduced the number of dialog acts from 50,633 to 37,508.
- To reduce data sparsity, we grouped switchboard dialog acts into dialog act classes. This reduced the number of dialog acts from 148 to 9 dialog act classes. See Table 1 for examples of the mapping.
- We added a binary y_{i+1} feature to each dialog act. As explained in Section 3, the variable is 1 if there is a turn change from dialogue act d_i to d_{i+1}.

To test the contribution of the summary features, we used a binary classifier with y_i as the outcome variable. We trained four models, which used the following sets of features:

baseline 1: Predict turn transition based only on the current dialog act label.

baseline 2: Predict turn transition based on the labels of the current and previous dialog acts.

summary model: Predict turn transition using just the summary features.

full model: Predict turn transition using the summary features and the current and previous dialog acts.

We used random forests to build the binary classifiers ($N = 200$) [23]. Random forests build an ensemble of decision trees during training, and during testing, each decision tree votes on the outcome. Like decision trees, it can account for interactions between variables, such as making greater use of the summary features when the current speech act is not a question. Random forests though are not as sensitive to overfitting and data fragmentation.

To find the optimal hyper parameters, we ran a grid search over the *max_features* and *max_depth* hyper parameters for each model. The hyper parameters search was done over $\{sqrt, log2, 10\}$ for *max_features* and $\{5, 7, 9\}$ for *max_depth*. When training the model, we used the optimal hyper parameters for each feature set.

We performed 10 fold labeled cross validations. We made sure that each conversation was entirely in a single fold. This way, each dialogue was entirely used for training or testing, but never for both at the same time.

5. Results and Discussion

We first analyze the results in terms of accuracy: how often the models correctly predicted whether a turn transition occurred or not, in other words, how often it predicts the correct value of y_{i+1}. Table 2 shows the results of training a random forest for each model. We see that using the summary features provides better accuracy than baseline 1, which only uses the current current dialog act (66.14% vs 60.26%). In addition, using the full model yields an improvement of over 1.58% in the result.

Model	Accuracy	AUC	hyper parameters
Baseline 1	60.26%	0.63	max_features=sqrt, max_depth=7
Baseline 2	74.43%	0.79	max_features=log2, max_depth=9
Summary	66.14%	0.65	max_features=sqrt, max_depth=5
Full	76.05%	0.82	max_features=10, max_depth=9

Table 2: Accuracy and AUC results

The effect can also be seen in Figure 3, which shows the ROC curves and the AUC for each model. We notice that the AUC of the summary model is better than baseline model (0.65 vs 0.63), and when adding the summary features to the local features, the full model, we see the AUC improves (0.82 vs 0.79). This suggests that while the discrimination facility of the summary features is lacking (AUC $<$ 0.7), adding them to a classifier that uses local features (full model) yields better results than the baseline.

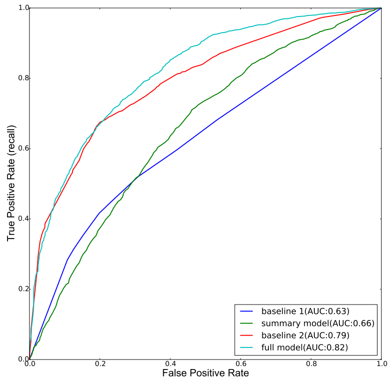

Figure 3: ROC curves and AUC of the different models

In addition to analyzing the results in terms of accuracy, we also analyze the results of the four models in terms of how well we predict that there is a change in speaker (i.e., y_{i+1} indicates that there was a turn switch). Table 3 shows the results in terms of recall, precision, and F1, which combines the two scores. Although baseline 1 has high precision, it has very low recall. Using only the summary model improves recall and decreases precision by less, leading to a higher F1 score and overall better performance. Using the full model improves precision, which means that dialog acts that were considered to lead to turn transitions are classified correctly. If we use the full model, we lose precision (over baseline 2 model), but gain recall, leading to the highest F1 score and the best performance.

Model	Precision	Recall	F1
Baseline 1	69.49%	45.52%	54.97%
Baseline 2	80.38%	68.80%	74.08%
Summary	64.55%	68.88%	66.42%
Full	76.17%	77.25%	74.87%

Table 3: Precision, recall and F1 results

6. Conclusions and Future Work

This paper explores the use of features that capture speakers' past turn-taking behavior in predicting whether their will be a turn transition. These summary features include (a) relative turn length: how the current turn under construction compares to the current speaker's average turn length; and (b) relative floor control: the percentage of time that the current speaker has held the floor. We include two versions of each, one based on time, and one based on number of words. Relative turn length should capture whether one or both of the speakers tends to hold the turn over multiple utterances, while relative floor control captures whether one speaker is dominating the conversation. Both of these factors should influence who will speak next.

In evaluating our model on data from the Switchboard corpus, we find that our summary features on their own do better than just using the previous speech act (accuracy of 66.14% vs 60.26%). We also find that when we add these features to a model that uses the last two speech acts, we also see an improvement (76.05% vs 74.43%). These results show the potential of modeling speakers' past turn-taking behavior in predicting upcoming turn-transitions. Better modeling of turn-taking should lead to more natural and efficient spoken dialogue systems.

In this work, the local features that we considered in our baseline model were just the last two speech acts. Other work on turn-taking prediction use a richer set of local features, such as syntactic [12, 1, 4, 3, 24, 19], prosodic [12, 4, 25, 6, 3, 26, 20, 27, 19], pragmatic [4, 16, 20], semantic [20] and non verbal [13]. In future work, it would be good to evaluate the contribution of our summary features with a richer set of local features.

In our work, we evaluated our model on the Switchboard corpus. In future work, it would also be good to evaluate our summary features on other corpora, especially task-based dialogues. Tasks in which there is a difference in the role of the user and speaker, such as in Trains [28], should benefit from modeling the past turn-taking behavior of each speaker.

More generally, the summary features introduced in this work represent just one aspect of the conversational image of the user. Future work should try to "summarize" other local features by creating the average value of a local feature over past turns. For example, we can compute relative speech rate, or relative use of stereotyped expressions.

7. Acknowledgements

This work was funded by the National Science Foundation under grant IIS-1321146.

8. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *language*, pp. 696–735, 1974.

[2] S. C. Levinson, "Turn-taking in human communication–origins and implications for language processing," *Trends in Cognitive Sciences*, vol. 20, no. 1, pp. 6–14, 2016.

[3] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: A cognitive cornerstone of conversation," *Language*, pp. 515–535, 2006.

[4] C. E. Ford and S. A. Thompson, "Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns," *Studies in Interactional Sociolinguistics*, vol. 13, pp. 134–184, 1996.

[5] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody," in *INTERSPEECH*, 2002.

[6] ——, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," in *ICASSP*, 2003, pp. 608–611.

[7] C. E. Ford, "At the intersection of turn and sequence," *Studies in Interactional Linguistics*, vol. 10, p. 51, 2001.

[8] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011.

[9] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue," *Language Resources and Evaluation*, vol. 44, no. 4, pp. 387–419, 2010.

[10] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, 1992, pp. 517–520.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

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

[13] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, pp. 22–63, 1967.

[14] E. O. Selfridge and P. A. Heeman, "Importance-driven turn-bidding for spoken dialogue systems," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala Sweden, 2010, pp. 177–185.

[15] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn-taking," *Psychonomic Bulletin & Review*, vol. 12, no. 6, pp. 957–968, 2005.

[16] S. Garrod and M. J. Pickering, "The use of content and timing to predict turn transitions," *Frontiers in Psychology*, vol. 6, 2015.

[17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[18] H. Arsikere, E. Shriberg, and U. Ozertem, "Enhanced end-of-turn detection for speech to a personal assistant," in *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.

[19] M. Atterer, T. Baumann, and D. Schlangen, "Towards incremental end-of-utterance detection in dialogue systems." in *COLING (Posters)*, 2008, pp. 11–14.

[20] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 9, no. 1, p. 1, 2012.

[21] N. Guntakandla and R. Nielsen, "Modelling turn-taking in human conversations." in *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, Stanford CA, 2015.

[22] J. Webber, "A programmatic introduction to neo4j," in *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, ser. SPLASH '12, New York, NY, USA, 2012, pp. 217–218.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] L. Magyari and J. P. De Ruiter, "Prediction of turn-ends based on anticipation of upcoming words," *Frontiers in Psychology*, vol. 3, p. 376, 2012.

[25] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1, pp. 127–154, 2000.

[26] B. S. Reed, "Units of interaction: intonation phrases or turn constructional phrases," *Actes/Proceedings from IDP (Interface Discours & Prosodie)*, pp. 351–363, 2009.

[27] R. Hariharan, J. Hakkinen, and K. Laurila, "Robust end-of-utterance detection for real-time speech recognition applications," in *ICASSP*, 2001, pp. 249–252.

[28] P. A. Heeman and J. F. Allen, "The Trains spoken dialog corpus," Linguistics Data Consortium, CD-ROM, April 1995.