The Effect of Speech Disfluencies on Turn-Taking

Lucy Li[1] **Kartik Sawhney**[2] **Divya Saini**[2]

[1]Symbolic Systems Program and [2]Computer Science Department
Stanford University, Stanford, CA
{lucy3, sainid, kartiks2}@stanford.edu

Abstract

Though modern dialogue systems have improved in processing language, the fluidity and rhythm of back-and-forth dialogue is still lacking. In particular, humans have the ability to predict end of turns accurately and quickly, and they are also comfortable with speaker overlaps in some situations, such as affirmative reinforcement (*uh-huh*) of the present speaker. In this present paper, we investigated the role of filler words and discourse markers in indicating turn continuations and endings, using the Switchboard Corpus. We confirm that pause duration alone is a poor predictor of end-of-turn, while acoustic and simple lexical features improve performance to 80% accuracy.

1 Introduction

The naturalness of an interactive speech system such as a chatbot is an important metric of performance evaluation. Chatbots aim to have accurate language understanding and generate coherent responses, but there is more to a conversation than its content. The organization of dialogue, such as turn-taking, is also a significant factor for an artificial system to learn. Dialogue between humans is fast and messy, with overlaps and hesitations, yet people still manage to communicate. The average length of a pause between turns in natural dialogue is 200 ms, suggesting that response planning and end-of-turn prediction must occur quickly and early, and with minimal use of between-turn pause duration (Levinson, 2016). However, many dialogue systems, especially simplistic ones, rely mostly on pause thresholds (Stolcke, 2002).

Automatic speech recognition systems have come a long way not only in recognizing actual words, but also in eliminating filler words and disfluencies to generate a more understandable transcript (Seltzer et al., 2013; Rao et al., 2007). However, these disfluencies can provide important contextual information, such as utterance planning and dialogue acts (Clark and Tree, 2002). They can be key to identifying the nature of an utterance, allowing one to infer speakers' intentions and avoid improper interruptions in conversation.

We wish to investigate how filler words (*yeah*, *um*) and discourse markers (*well*, *you know*) help people hold, cede, or take the floor in a dialogue[1]. The results of project would in turn help speech systems better interact with people, provide automated tools for conversation analysis, and enrich current linguistic understanding of turn-taking.

2 Related Work

In traditional linguistics, the field of conversation analysis characterizes speech as a joint activity between or among people that requires coordination, planning, and negotiation. People are able to pinpoint when a turn will end with high accuracy, and interestingly, Garrod and Pickering (2015) argue that they rely more on content and speech rate rather than pitch to make these predictions. This can be contrasted with the usage of prosodic features in automatic speech processing to improve end-of-turn detection Stolcke (2002). Others have focused on the relationship between turn-taking and timing. Ten Bosch et al. (2005) used timing information provided by transcripts to calculate overlap and pause duration distributions in telephone and face-to-face conversations, showing that overlaps occur more often in the former and 71% of pauses in telephone conversations are between speaker changes.

Commonly considered to be signifiers of sen-

[1]All code can be found on Github.

tence planning, filler words and discourse markers are most common in spontaneous speech or longer sentences (Beliao and Lacheret, 2013; Ten Bosch et al., 2005). Categorization of a word or phrase into either of these categories is up for debate. For example, some consider *but* and *oh* to both be discourse markers while others treat the former as a conjunction and the latter as filler word (Schiffrin, 2001). There have also been efforts in discerning the differences among fillers. For example, *um* is used for bigger delays in speaking than *uh* (Clark and Tree, 2002). Shriberg (1996) mapped the distribution of disfluencies in the Switchboard Corpus, showing that fillers, edits, and repetitions are most likely to occur in the beginning of an utterance. So, though these kinds of words and phrases are treated as "fillers," they are not homogeneous in purpose or placement.

End-of-turn detection is a structural segmentation task in natural language processing, and models in the past have utilized lexical and prosodic features. HMM, maximum entropy, and conditional random field models have been popular in work on sequence tagging and disfluency detection. The lexical features used were often word or part-of-speech n-grams, and prosodic features included information about duration, pause, intonation, and energy (Ostendorf et al., 2008; Liu et al., 2006; Schlangen, 2006). Despite variance in speaker style, Ostendorf et al. (2008) showed that boundary separation features such as F0, duration, and energy features have similar patterns among speakers, while pausing behavior is more inconsistent. Schlangen (2006) used only prosodic features of a single word with lexical features before making a decision to wait or take a turn, and achieved 68% accuracy using a decision tree classifier.

3 Approach

As discussed, it seems that multiple factors are at play in proper turn-taking. In this present paper, we hypothesize that words commonly used as fillers and discourse markers provide turn-taking information. In particular, we wish to investigate the ability of prosodic features to differentiate the various roles disfluencies play in speech, and confirm that timing information alone is not enough for characterizing turn-taking phenomena.

3.1 Data

Few corpora provide information on turns in dialogue, but some have been annotated with related information, such as dialogue acts, disfluencies, and detailed word-based time stamps. One of the most popular corpora in speech and conversation research is the Switchboard-1 Telephone Speech Corpus, consisting of approximately 2,400 telephone calls between strangers. We chose a telephone corpus due to its absence of face-to-face interaction, which would allow us to focus on turn-taking using only speech-based cues, making the task more relevant to modern chatbots. We treat these human-human conversations as the gold standard for turn-taking, though it is important to keep in mind that humans are also susceptible to errors and misunderstandings. Some examples of this include one speaker unintentionally interrupting another or not taking a turn during an opportunity to do so. Detecting such cases and mitigating their effects is a possible future line of work in this area.

We used manually corrected word alignments created by the Institute for Signal and Information Processing[2] and for disfluency annotations, we used Penn Treebank-3[3]. Word alignments include timings for silence, noise, and individual words for each speaker. The disfluency annotations labeled filler words and discourse markers separately. These two transcriptions of Switchboard were not consistent, therefore necessitating an alignment step. Words in the text that necessitated insertions or deletions when the two transcriptions were compared were discarded.

It is important to note that not all occurrences of words such as *so*, *and*, and *like* are disfluencies. For example, *Like, I don't really like the beach* has *like* playing both the role of a disfluency and a content word. To simplify the problem, we analyze all occurrences of *like* and do not disambiguate these cases. In addition, these annotations treat structural connectives such as *but* and *or* not as discourse markers, but as a separate class of conjunctions.

3.2 Vocab

Penn Treebank transcriptions of Switchboard alternate between speaker A and speaker B, so that each line is an approximation of a turn in the con-

[2] ISIP.

[3] The guide for this can be found here.

	all features	only filler words	only word and pause durations	only pause duration
SVM	0.75 (+/- 0.01)	0.74 (+/- 0.01)	0.66 (+/- 0.01)	0.49 (+/- 0.02)
Random Forest Classifier	0.76 (+/- 0.01)	0.74 (+/- 0.01)	0.69 (+/- 0.01)	0.61 (+/- 0.00)

Table 1: Summarized in this table are the accuracy results from the two baseline models. All features include lexical values, pause duration, and word duration in the feature vector, while the other versions refer to the inclusion of specific subsets of those features in the vector.

starters		enders	
word	count	word	count
yeah	771	uhhuh	644
uhhuh	711	yeah	465
and	484	laughter	201
oh	240	right	168
i	224	know	115
well	212	it	109
right	176	and	85
you	162	that	74
laughter	151	uh	74
uh	141	there	49
but	119	um	46
so	91	oh	45
thats	81	so	41

Table 2: Above is a consolidated count of the most common turn starters and turn enders as found in the Penn Treebank transcriptions of Switchboard. Out of 5299 total turn starters and 5262 total turn enders, many filler words have high frequency counts. This suggests that filler words have high potential as signals for starting and ending a turn.

versation. As a preliminary look into our data, we extracted the words that most commonly appear at the beginning and endings of each line. These can be found in Table 2, and suggest that filler words have high potential as signals for starting and ending a turn.

The words we focused on were a set of words labeled as fillers or discourse markers in the disfluency annotated dialogue, with a few additions. They consisted of the following: *actually*, *anyway*, *anyways*, *cool*, *hm*, *huh*, *huhuh*, *hum*, *like*, *nice*, *now*, *oh*, *okay*, *right*, *say*, *see*, *so*, *sure*, *that's right*, *that's true*, *true*, *uh*, *uhhuh*, *um*, *well*, *wow*, *yeah*, *yep*, *yes*, *you know*, *you see*, and *uhhum*. The spelling of many filler words, such as *uh* and *um* were not consistent across the two transcripts. For our analysis we used the spellings found in the time-stamped word transcriptions, though we changed *umhum* to *uhhuh* to improve our alignment with the disfluency transcription.

4 Experiments

As a baseline experiment, we used timing information to predict whether a word was a continuation or an ending of a turns. We used the definition used by Ten Bosch et al. (2005), where a turn is a stretch of one or more utterances that are not interrupted by another speaker. This definition is reasonable but does not necessarily account for cooperative affirmations or other brief interruptions that overlap with the current speaker's turn, but it was a good starting definition for our preliminary experiment. Like Ten Bosch et al. (2005), we used only transcription-based information to analyze turn-taking. Each of these words was represented by a feature vector, which had its duration, the duration of the pause in the conversation following it, and binary values that indicated the lexical label of the disfluency. Durations were normalized by their averages. We used an SVM and a random forest classifier to analyze this data and observed that pauses alone were not the best indicators, leading us to continue our work in our intended direction.

For our main experiment, we extracted audio clips of our words of interest. We used the disfluency annotated transcript to define our turns, where each speaker's line was a turn, unless their utterance continued to the next line. For example, in the following Penn Treebank-3 excerpt, A continues their turn over B's interruption:

```
A.72: {C but } we just went shopping / {
    C and } we came back [ with , + {F uh
    , } with , ] {F uh , } sweets , {D you
    know , } chocolate covered peanuts /
    {C and } —
B.73: Ugh. /
A.74: — {F uh , } we came back with
    sweets . / We didn't bring all the
    healthy food back too . /
```

We used corresponding transcriptions and annotations to label the extracted sound bytes as

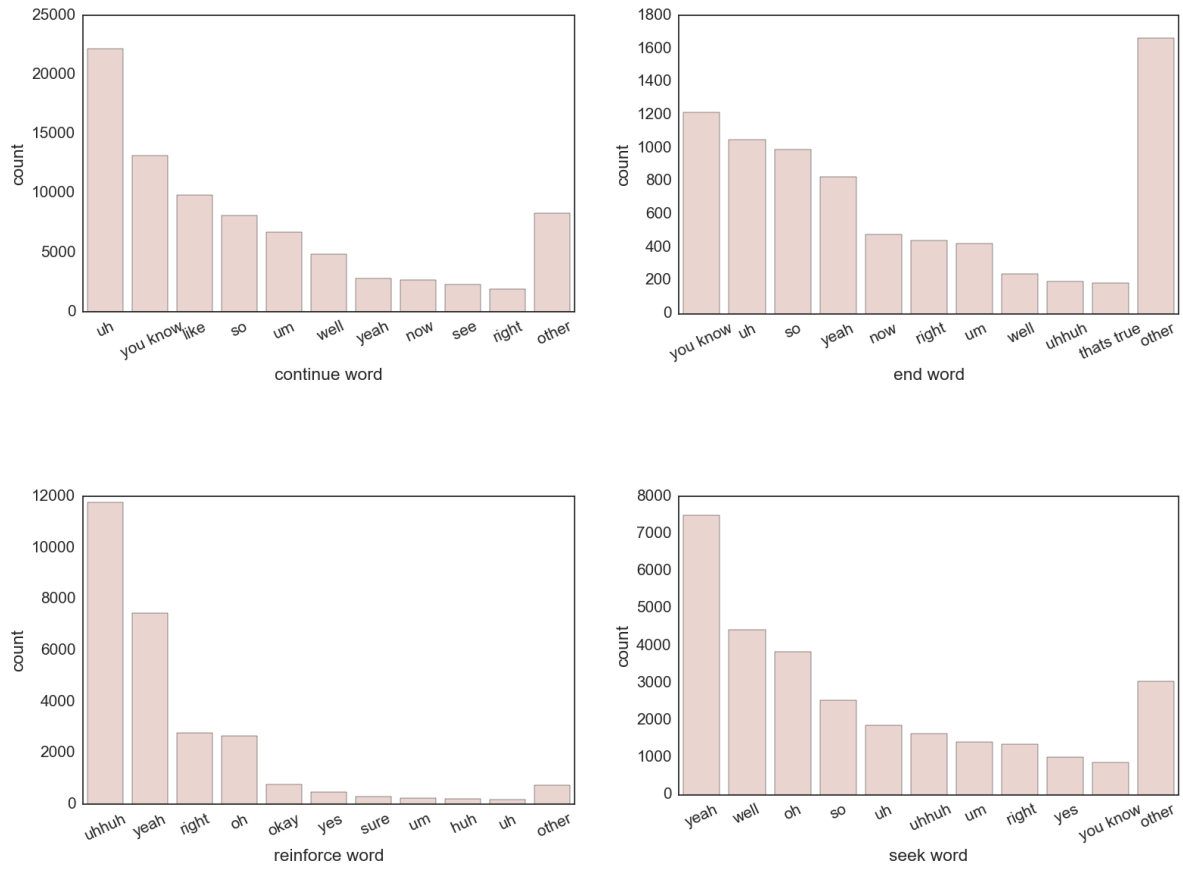

Figure 1: We had 7,721 enders, 27,664 reinforcers, 83,157 continuers, and 29,502 seekers in our labeled dataset. Shown are counts of the top ten words in each category. Note that the vertical axes do not have the same upper bound.

continuers, enders, seekers, and reinforcers. The first three classes are based on the placement of a word within a turn, whether it be at the middle, ending, or beginning, respectively. A reinforcement is where a speaker makes an affirmative or backchannel response (*yeah* or *uh-huh*) during another speaker's turn. We used dialogue act annotations of Switchboard to determine which lines were backchannels[4].

We extracted 34 features, including 12 MFCC features and 12 Chroma features as well as spectral and energy features[5], from overlapping windows in the audio clips of each word. Some timestamps were incorrect, such as starting times occurring after ending times, in which case the word was discarded. We used these features as an input to a GRU RNN sequence classifier with one hidden layer, but prior to the softmax layer, we appended a one hot feature vector indicating the lexical value of that example. We had two RNN classifiers, one that separated the data into all four classes, and another that predicted two classes. In the two class version, we reclassified reinforcers as enders and seekers as continuers.

RNNs are usually used for processing sequences, such as text. Take a sequence $x_1, x_2, ..., x_t, ..., x_T$. Each hidden state h_t of an RNN can be represented as $h_t = f(h_{t-1}, x_t)$ where h_{t-1} is the previous hidden state, x_t the sequence input at that time step, and f is a gated recurrent unit (GRU) (Cho et al., 2014). For us, each time step is a window of audio features. The function f has four stages, where W_z, W_r, W_h, U_z, U_r, and U_h are weights:

$$z_t = \sigma(W_z x_t + U_z h_{t-1})$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$

$$\widetilde{h}_t = \tanh(r_t \circ U_h h_{t-1} + W_h x_t)$$

$$h_t = (1 - z_t) \circ \tilde{h}_t + z_t \circ h_{t-1}$$

[4]These can be found here with a corresponding manual.
[5]Detailed in this Python package.

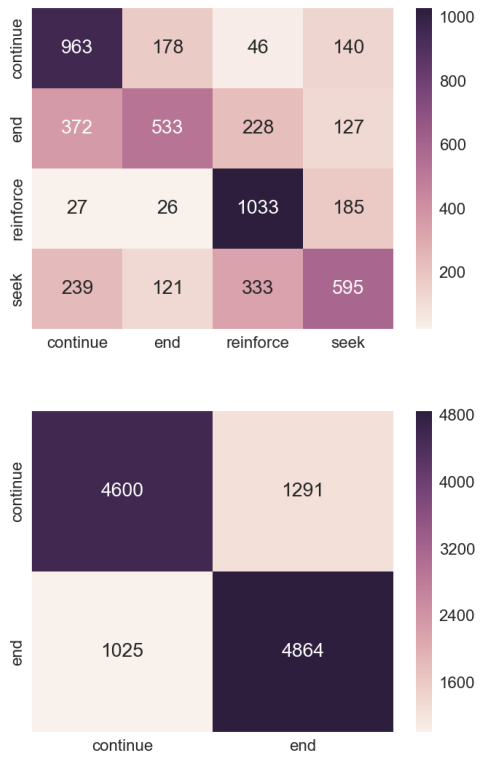

Figure 2: Heat maps of RNN sequence classifier confusion matrices for four classes and two classes on the test set. Actual labels are along the side, and predicted labels along the bottom.

Our RNN had a tendency to overfit, so the output of the hidden layer had only three dimensions and we included dropout with a 40% probability of keeping a connection. In a sequence classifier, only one output y is generated, with $y = softmax(Wh_T + b)$ where W is the weight and b is a bias term. We used cross entropy loss, which can be defined as a sum over all examples j:

$$J = -\sum_j y_j log(\hat{y_j}).$$

5 Results

5.1 Baseline Model

For this model, we had an even split of continuation and end-of-turn examples, 329,272 each. Our results can be found in Table 1. The performance of pause duration alone, or in conjunction with word duration, is much lower than lexical features alone and for our linear classifier, no better than prediction at random. The length of the average pause within a turn is 0.146 with a standard deviation of 0.326, and the length of the average pause between turns is 0.139 with a standard deviation of 0.328. So, there is quite a bit of overlap, making pause duration a poor predictor of end-of-turns. This also suggests that natural conversations reflect people's abilities to detect turns quickly and their dislike of silence.

Our baseline performance demonstrates that filler words and discourse markers are good indicators of turn continuation and ending, and a combination of all features yields up to 76% accuracy with a random forest classifier. In particular, *yeah* had the highest importance in the random forest classifier among all filler words and discourse markers. The duration of the word itself, which is linked to prosody, had a higher absolute weight or importance than pause duration in all classifiers that included it.

Model	train	val	test
RNN: 2-Class	82.00%	80.30%	80.34%
RNN: 4-Class	64.00%	59.70%	60.71%

Table 3: Training, validation, and test accuracies for our RNN sequence classifiers.

5.2 RNN Model

We examined the makeup of our labeled data. Figure 1 shows counts of the top ten words in each category. Our words of interest appeared as continuers the most often and enders the least often. Some words play multiple roles, such as *yeah*, while others are more exclusive, such as *uhhuh*. The most common reinforcers also seem to be common turn seekers. It may be that reinforcing the other speaker's completed utterance is an effective route to grabbing the floor. It is also interesting that many turn endings are ones that seem to indicate an incomplete utterance, such as *uh* and *so*. In the disfluency annotated transcription, some turns were ended by an incomplete utterance, where a speaker abandons their train of thought.

Table 3 and Figure 2 show the results of our two RNN sequence classifiers. Our four class model did most well at accurately predicting continuations and reinforcements, while it misclassified many enders as continuers and many seekers as reinforcers. Figure 3 shows three-dimensional PCA embeddings created from RNN outputs before the final softmax layer. For four classes, though there is somewhat a reinforce cluster and a seek clus-

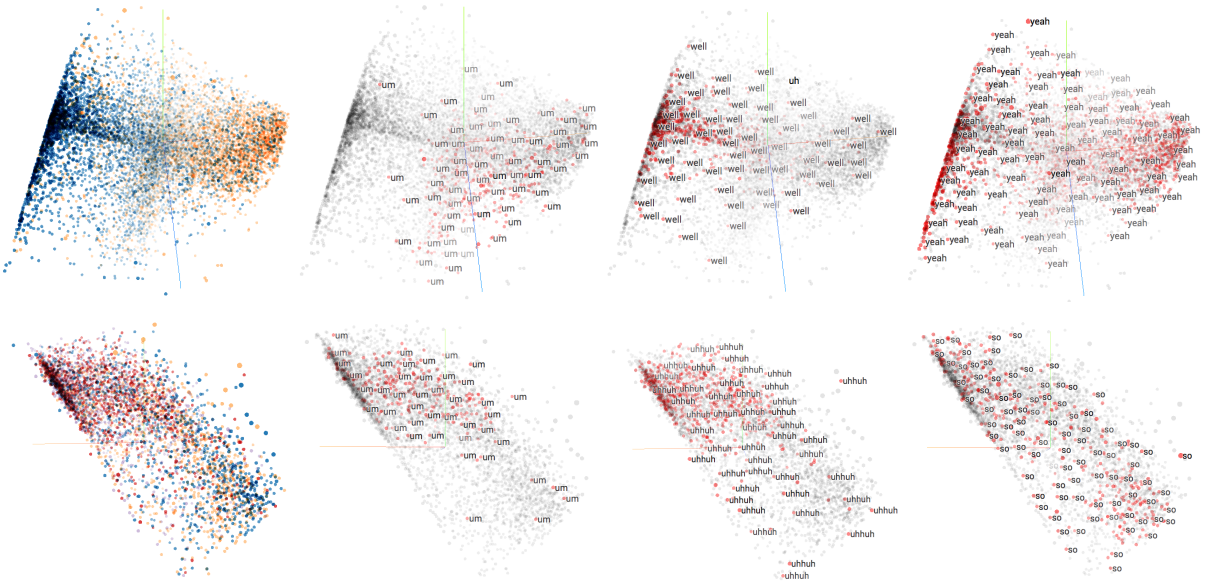

Figure 3: These representations are learned by RNNs from sequences of audio features.
Top leftmost: Each example is colored with blue as continue and yellow as end. The following three figures show that *um* tends to cluster on one end, *well* on the other, and *yeah* everywhere.
Bottom leftmost: Each example is colored with red as reinforce, blue as seek, yellow as continue, and purple as end. The following three figures show that *um* and *uhhuh* tend to cluster near the top left, while *so* is more scattered.

ter, there is also quite a bit of overlap among the classes. For two classes, the separation between continuers and enders is very clear. Some words tend to form clusters in the learned outputs, while others, such as *yeah* and *so*, scatter themselves across the embedding space.

6 Conclusions

Our models relied on a very short window of information to predict end-of-turns and other turn-taking behaviors. We focused on filler words and their duration, lexical, and acoustic features, yet we were able to reach 80% accuracy on our best model. Although it may be difficult to distinguish continuers, reinforcers, seekers, and enders, we can at least know whether it is proper to speak up or not by predicting continuers and enders. More context, such as words prior to these words of interest, will likely improve our results even further.

We also showed that a simple model reliant on just the presence of certain words achieved much higher end-of-turn prediction accuracy than pause thresholds. This is significant because we often believe that the easiest way to predict end-of-turn is by using pauses. This may be true, since once someone has finished making a point, they will pause to await the chatbot's response. However, if we wanted the chatbot to start formulating a response earlier or respond with a more reasonable speed, focusing on discourse markers and fillers may be an effective alternative. The additional challenge our solution presents would be a speech recognition one, where these words must be recognized almost immediately after the speaker has uttered them.

Other future investigations for predicting end of turns include looking at structural connectives and incorporating syntactic context. It may be that one could prepend a RNN sequence classifier with a convolutional neural network for extracting features from a spectrogram, though this may be excessive. It may also be interesting to predict other discourse phenomena using filler words and discourse markers, such as dialogue acts, confidence, awkwardness, and naturalness.

References

Julie Beliao and Anne Lacheret. 2013. Disfluency and discursive markers: when prosody and syntax plan discourse. In *DiSS 2013: The 6th Workshop on Disfluency in Spontaneous Speech*. volume 54, pages 5–9.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* .

Herbert H Clark and Jean E Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1):73–111.

Simon Garrod and Martin J Pickering. 2015. The use of content and timing to predict turn transitions. *Frontiers in psychology* 6:751.

Stephen C Levinson. 2016. Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences* 20(1):6–14.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing* 14(5):1526–1540.

Mari Ostendorf, Benoît Favre, Ralph Grishman, Dilek Hakkani-Tur, Mary Harper, Dustin Hillard, Julia Hirschberg, Heng Ji, Jeremy G Kahn, Yang Liu, et al. 2008. Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine* 25(3).

Sharath Rao, Ian Lane, and Tanja Schultz. 2007. Improving spoken language translation by automatic disfluency removal: Evidence from conversational speech transcripts. *Training* 6370(46300):6–50.

Deborah Schiffrin. 2001. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis* 1:54–75.

David Schlangen. 2006. From reaction to prediction: Experiments with computational models of turn-taking. *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking* .

Michael L Seltzer, Dong Yu, and Yongqiang Wang. 2013. An investigation of deep neural networks for noise robust speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, pages 7398–7402.

Elizabeth Shriberg. 1996. Disfluencies in switchboard. In *Proceedings of International Conference on Spoken Language Processing*. Citeseer, volume 96, pages 11–14.

Luciana Ferrer Elizabeth Shriberg Andreas Stolcke. 2002. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody .

Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication* 47(1):80–86.