Leveraging Prosody for Punctuation Prediction of Spontaneous Speech

Jenny Yeonjin Cho[1], *Sara Ng*[2], *Trang Tran*[3], *Mari Ostendorf*[1]

[1]University of Washington, Department of Electrical and Computer Engineering
[2]University of Washington, Department of Linguistics
[3]University of Southern California, Institute for Creative Technologies

{yeonjinc,sbng,ostendor}@uw.edu,ttran@ict.usc.edu

Abstract

This paper introduces a new neural model for punctuation prediction that incorporates prosodic features to improve automatic punctuation prediction in transcriptions of spontaneous speech. We explore the benefit of intonation and energy features over simply using pauses. In addition, the work poses the question of how to represent interruption points associated with disfluencies in spontaneous speech. In experiments on the Switchboard corpus, we find that prosodic information improved punctuation prediction fidelity for both hand transcripts and ASR output. Explicit modeling of interruption points can benefit prediction of standard punctuation, particularly if the convention associates interruptions with commas.

Index Terms: Automatic punctuation, speech recognition, prosody

1. Introduction

Punctuation is a crucial part of many languages, such as English, as inappropriate placement or use of punctuation marks can semantically alter the intention of the writer. Punctuation is not verbalized in spoken language; sentence structure is communicated via prosodic cues, such as pauses, duration lengthening, and intonation associated with phrase structure and utterance intent. In written text, punctuation serves as a proxy for prosodic information. Spontaneous speech transcribed by automatic speech recognition (ASR) systems often lacks the punctuation marks that one would expect in written text, resulting in transcripts that can be difficult to comprehend correctly. This is especially true when lack of punctuation is compounded with automatic transcription errors. The automatic prediction of punctuation for transcribed speech is therefore important to represent the structure of the spoken utterance.

In more formal contexts or for read speech, punctuation can be reasonably well predicted from language context alone (i.e. without attending to prosodic features), particularly with powerful neural language models. Perhaps for that reason, most recent punctuation prediction work has not used prosody. However, for conversational speech, which does not adhere to written grammatical structure and often includes disfluencies, it may be that prosodic cues are more helpful. In addition, prosodic cues may help compensate for confusions associated with ASR errors, though the prosodic features themselves may be sensitive to ASR errors. Pauses are reasonably reliable, but speakers use pauses for multiple reasons, including hesitation and in putting special emphasis on a word. Further, punctuation is not always associated with a pause.

In this paper, we propose a mechanism for incorporating prosodic features into a neural punctuation prediction model, building on prior work in spoken language processing. Working with conversational speech, our work explores two questions. First, to what extent does prosody (beyond pauses) improve standard punctuation prediction and are findings impacted by ASR word errors? Second, we explore explicit prediction of interruption points associated with disfluencies. There is no standard convention for punctuation associated with the interruption point (IP) and it is often unmarked. What types of confusions does this introduce, and might it be useful to explicitly represent interruption points in the punctuation set? In experiments with the Switchboard corpus, we find that there is a gain from using acoustic-prosodic cues beyond pauses, both for manual and automatic transcripts. Commas are the most difficult to accurately predict, and explicit modeling of interruption points does not improve overall performance.

2. Related Work

2.1. Automatic Punctuation Prediction

Researchers have been exploring methods for automatic punctuation prediction for many years [1, 2, 3, 4], since it benefits readability for humans as well as automatic language processing. Recent work has leveraged neural models. Different variants of RNNs are explored in [5, 6], and CNNs are also used in [5]. Recent studies primarily leverage pre-trained transformers [7, 8], which is an approach this work also takes. A comparison of all these architectures is provided in [9].

There are some attempts to recover punctuation in other languages (see [10] for Hungarian, or [11, 12, 13] for Chinese), however English datasets are more commonly used. The work on English has involved a variety of speech styles, including audio books [14, 8], broadcast speech [6], TED Talks [7], medical dictation [9], and conversational speech [5, 8, 9]. Our focus will be on conversational speech. There is no standard set of punctuation marks. The most common set is {comma, period, question mark}, which is used in [5]. The punctuation set from [9] includes only {comma, period}. In contrast, [8] predicts punctuation marks from the set {full stop, comma, question mark, exclamation mark, semicolon, double-dash, ellipsis}. In our work, we consider expansions of the basic set to handle two conversational speech phenomena: incomplete sentences and disfluency interruption points. Many studies only consider hand transcripts; results on automatic transcripts are presented in [6, 9], both showing lower F1 scores for ASR.

While several early studies explore the use of prosodic features in punctuation prediction [2, 3, 14], most recent work relies solely on the speech transcripts. A notable exception is [6], which leverages features similar to the work here, but within a hierarchical RNN framework. The key difference in our approach is the neural architecture (transformer+CNN) and inclusion of pause and duration features.

2.2. Prosody in Spoken Language Processing

Prosody has been used in many spoken language processing tasks, most notably segmentation, parsing, disfluency detection, and dialog act (DA) recognition. In a long line of work, prosody was shown to improve topic segmentation [15, 16], sentence boundary detection [17, 18, 19], and turn segmentation [20]. [21] leveraged automatically predicted prosodic labels (i.e. ToBI [22]) in a statistical parser, achieving improvements in both parsing and disfluency detection. Similarly, in [23], prosody was shown to benefit joint parsing and word recognition, especially when sentence boundaries were unknown. More recent work in parsing [24] modeled raw acoustic features and showed the benefit of prosody especially in disfluent sentences and attachment error corrections. Most works in DA focused on sentence-level classification of a DA given a known (segmented) utterance. In earlier work, the use of prosody was shown to be beneficial, specifically in distinguishing questions from statements, and backchannels from agreements [25, 26]. Using a similar approach to [24], [27] showed that prosody benefited joint segmentation and DA classification, where prosody and dialog history seem to be complementary—prosody benefits segmentation while history benefits classification.

Many of these studies, however, mostly relied on hand transcripts. For ASR outputs, [28] applied a CNN on segment-level MFCCs, and improved accuracy over using only ASR outputs. A joint DA segmentation and classification system with an acoustic-to-word model is described in [29], but it was not clear where performance most suffered by using imperfect transcripts. Parsing on ASR transcripts using prosody was shown to yield more gains than parsing with only text information, as reranking helps recover function words, and seems to favor grammatically correct utterances [30].

3. Methods

3.1. Task and Data

The dataset in our work is Switchboard (SWBD) [31], a collection of spontaneous telephone speech between strangers prompted to talk about a specific set of topics. SWBD has been widely used for a number of speech transcription tasks, including parsing, disfluency detection, dialog act recognition, sentence segmentation, and of course, speech recognition. Our work builds on a system originally designed for joint dialog act segmentation and recognition, so we use the portion of SWBD that was annotated with dialog acts. For training, tuning and testing the different models, we use the split commonly used in dialog act classification, which are defined in [32]. Since the test set is not fully annotated with disfluencies, experiments with interruption point prediction are reported on the subset that is annotated, referred to as the "IP test." Table 1 shows statistics of the different dataset splits.

Table 1: *Data statistics of SWBD*

Split	# Dialogs	# Turns	# Sentences	# Tokens
train	1.1K	107K	194K	1.4M
dev	21	1.6K	3.2K	25K
full test	19	2.4K	4.1K	29K
IP test	14	1.7K	2.9K	21K

There are multiple transcriptions of the SWBD data. In this work we use the transcripts associated with disfluency annotations[1] when available, so as to be able to investigate inclusion of interruption points as a punctuation category. However, we use the utterance times associated with the more careful Mississippi State transcriptions [33], which have been aligned to the earlier transcripts used in dialog act and disfluency annotation.

For the punctuation task, a sample is a speaker turn, which is the concatenation of successive utterances from a speaker up until the other speaker takes the floor (ignoring overlap for backchannels). Backchannels are treated as a full turn. Utterances that have no words (e.g. laughter) are removed. For speech recognition, turn times are based on the start and end times of the first and last utterance, respectively. Unlike some text-based models of SWBD, we do not separate contractions into two tokens. This convention is more consistent with the patterns observed in the parallel acoustic features.

The ground truth punctuation labels are extracted from SWBD with some modifications, as illustrated in Figure 1. The disfluency span markers [...] and annotations of coordinating conjunctions {C ... } and discourse cue words {D ... } are ignored in this study. The transcription conventions included commas around all filled pauses (denoted as {F...}) and at interruption points (denoted with +), which is non-standard and artificially biases the predictions, so these were removed. The slash (/) marks boundaries of sentence-like units, so it was associated with a period irrespective of the punctuation used with the slash. Specific implementation of these and other mappings are listed below.

- Commas preceding filled pauses are removed.
- Commas attached to *uh* and *um* are removed, except before *you know*.
- Exclamation points, ellipses, and words followed by a plain slash are associated with periods.
- Periods are assigned to the word preceding **./** and **./**.
- Interruption points are assigned to the word preceding **+**.
- Incompletes are assigned to the word preceding **-/**.

Two series of experiments are conducted that differ in handling of IPs. In the first series, IPs are associated with no punctuation, resulting in a 4-class punctuation set (comma, period, incomplete, question). In the second series, IPs are explicitly predicted, resulting in a 5-class set. Words without punctuation are labeled as "other" (O). Table 2 shows the token counts of the 4-class punctuation types.

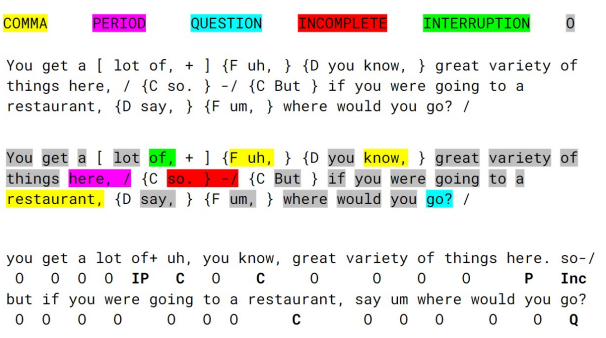

Figure 1: *Example of text preprocessing. The top figure shows the raw data, including original disfluency annotations; the middle figure shows the mapping of annotations to punctuation tags; and the bottom figure shows the resulting labels.*

[1] https://doi.org/10.35111/gq1x-j780

Table 2: *Counts of 4-class punctuation types: 'C,'=comma; 'P.'=period; 'Inc-'=incomplete; 'Q?'=question; 'O'=no punctuation. Roughly 4% of the "O" tokens correspond to IPs.*

Split	C,	P.	Inc-	Q?	O	Total
train	128K	127K	9.0K	7.8K	1.1M	1.4M
dev	3.2K	2.2K	144	92	19K	25K
full test	2.8K	2.7K	175	197	23K	29K
IP test	1.8K	1.9K	134	125	16K	21K

Punctuation is evaluated using macro F-scores. For the case when the transcripts are automatically generated, there may be words inserted or deleted. Following [9], if the automatic transcription deletes a word and that word is assigned a punctuation *and* the punctuation of the previously recognized word matches the deleted word's punctuation, then it is considered correct. The equations in (1) show how the precision and recall are computed for the prediction of question mark ("?") in automatically generated transcripts; metrics for other punctuations are computed similarly.

$$P = \frac{|TP(?)|}{|? \text{ in ASR}|} \quad , \quad R = \frac{|TP(?)|}{|? \text{ in reference}|} \tag{1}$$

3.2. Punctuation Prediction Model

The punctuation model is based on the dialog act recognition model from [27]. Specifically, this model is an extension of the best performing RNN encoder-decoder model with attention in [34], combined with the CNN module for learning acoustic-prosodic features as described in [35]. Unlike in [34], we do not use previous turn context labels, since punctuation prediction relies less on previous turns than dialogue act prediction. An overview of the model is presented in Figure 2.

Briefly, the encoder-decoder model takes a turn as input, each represented by $x = [x_1, \ldots, x_N]$, where x_i is the word-level input feature vector and N is the sequence length. The model learns to output the sequence of punctuation labels $y = [y_1, \ldots, y_N]$. An RNN encoder produces the hidden states $h = [h_1, \ldots, h_N]$, where $h_i = \text{RNN}(x_i, h_{i-1})$, and the RNN decoder computes hidden states $d_t = \text{RNN}([\tilde{y}_{t-1}; c_{t-1}], d_{t-1})$ where $\tilde{y}_{t-1}$ is the embedding associated with the label y_{t-1} and c_{t-1} is the context vector computed from the encoder hidden states of the whole sequence, i.e. $c_{t-1} = h\alpha_t$ where $\alpha_t = \text{softmax}(u_t)$, with $u_t = v^T \tanh(W_1 h + W_2 d_t + b_a)$, i.e. the additive attention function [36]. The predicted punctuation y_t is determined by $p(y_t|h, \tilde{y}_{1:t-1}) = \text{softmax}(W_s[c_t; d_t] + b_s)$, $W_1, W_2, v, W_s, b_a, b_s$ are all learnable parameters.

For the model which uses all prosody features, the input vectors $x_i = [e_i; \phi_i; s_i]$ are composed of word embeddings e_i, pause- and duration-based features ϕ_i, and learned energy/pitch (E/f0) features s_i, which taken together represent a prosodically contextualized word vector. The word embeddings e_i are pre-trained BERT embeddings [37] (BERT-base-uncased version), which have been shown to perform well on a variety of NLP tasks. Pause- and duration-based features ϕ_i are composed of both raw and categorical pause durations after each word (i.e. quantized raw durations to 6 categories as in [24]); word durations are normalized by the mean duration of the word in the training corpus vocabulary.

The acoustic-prosodic features s_i are learned via a CNN from energy (E) and pitch (f0) contours as described in [24]. Unlike in [24], where the convolution window is centered at the middle of the words, here we shift the window so that its center is located at the *end* of words in order to capture the f0/E towards the end. This is motivated by the phenomenon where speakers change the pitch and/or energy of their voice at the end of the word to communicate a prosodic boundary. The frame-level energy and pitch features are extracted using Kaldi [38] and normalized for each speaker side of the whole SWBD conversations. The frames corresponding to each word are then extracted based on word-level time alignments. Each sequence of f0/E frames corresponding to a time-aligned word (and potentially its surrounding context) is convolved with N filters of m sizes (a total of mN filters). The motivation for the multiple filter sizes is to enable the computation of features that capture information on different time scales. For each filter, we perform a 1-D convolution over the f0/E features with a stride of 1. Each filter output is max-pooled, resulting in mN-dimensional speech features s_i for word i. These prosody representations are jointly learned with the punctuation classification objective.

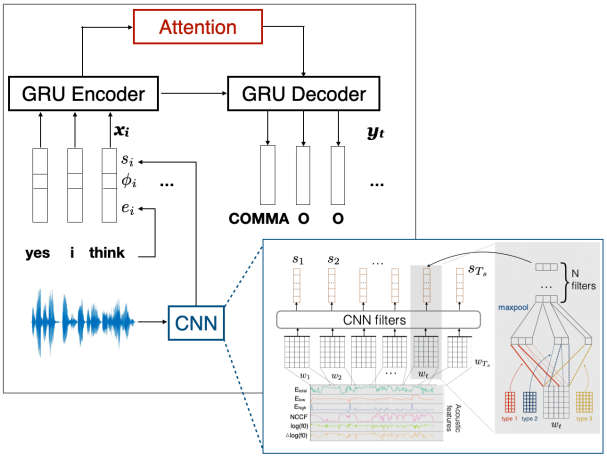

Figure 2: *Schematic of the punctuation model. Each turn u is encoded via embeddings of the BERT-tokenized text, (optional) pause and duration embeddings, and (optional) convolved acoustic features.*

In addition to a model trained on all features, we train a text-only model and a model where ϕ_i contains only the categorical pause feature. Neither of these comparison models encodes acoustic features. Our best model uses 12-dimensional pause embeddings; the CNN has $N = 32$ sets of filters of widths $[5, 10, 25, 50]$, i.e. $m = 4$, totaling 128 filters. The RNN is a uni-directional GRU [39], and our parameters were learned using Adam optimizer [40] with initial learning rate 0.0001, halving when the performance on the dev set does not improve every 3 epochs.

3.3. Automatic Speech Recognizer

We use an off-the-shelf ASR system, ASPiRE [41], a standard standard benchmark for ASR systems available in Kaldi's [38] model suite and trained on Fisher conversational speech data [42]. Briefly, the ASPiRE system was trained using a lattice-free maximum mutual information (LF-MMI) criterion, with computation efficiencies enabled by a phone-level language model and outputs at 1/3 the standard frame rate (one frame every 30 ms). The ASPiRE system has a reported word error rate (WER) of 15.6% on the Hub5 '00 evaluation set. The WER on our SWBD data is 21% and 24% for the development and test

sets, respectively.

4. Experiments

4.1. Standard Punctuation Prediction

Our experiments first address questions about the usefulness of prosody with the standard punctuation set used in most work: {period (P.), question mark (Q?), comma (C,)}, augmented by a marker for an incomplete sentence (Inc-). Table 3 gives results for the 4-class punctuation set on the full test set. For hand transcripts, the F0 and energy features give a small benefit over the pause in the macro scores due to improved detection of incompletes. Commas are the most difficult to predict, and they are likely more inconsistently used by human transcribers. Performance degrades for the ASR transcripts, as one would expect, but there is a particularly large drop for incomplete sentences, both for precision and recall. With ASR, prosody has a bigger impact, and the biggest impact is again for incompletes.

Table 3: *F1 scores for prediction of 4-class punctuation types on the full test set, using different features with hand vs. ASR transcripts.*

Hand	C,	P.	Inc-	Q?	Macro
text only	.60	.81	.80	.73	.736
pause only	.60	.82	.80	.74	.739
all features	.60	.82	.82	.73	.744
ASR	C,	P.	Inc-	Q?	Macro
text only	.53	.77	.49	.60	.597
pause only	.53	.77	.52	.62	.612
all features	.53	.78	.53	.62	.615

4.2. Explicit Interruption Punctuation

A second set of experiments looks at incorporating the interruption point (IP+) as an additional category. Table 4 gives results for the 5-class punctuation set on the IP test set. In the 4-class experiments, the interruption points were considered unmarked, so interruption detection that is reasonably high precision should not impact other classes much, which is the case for the manual transcripts. The precision drops quite a bit for the ASR transcripts (from greater than .7 to .4), but it mainly affects the unmarked cases. Interestingly, using the pause feature is not helpful when predicting from the 5-class punctuation set.

Table 4: *F1 scores for prediction of 5-class punctuation types on the IP test set, using different features with hand vs. ASR transcripts.*

Hand	C,	P.	Inc-	Q?	IP+	Macro
text only	.63	.81	.80	.79	.77	.761
pause only	.63	.82	.81	.78	.76	.759
all features	.65	.82	.82	.80	.78	.773
ASR	C,	P.	Inc-	Q?	IP+	Macro
text only	.56	.76	.49	.65	.54	.600
pause only	.56	.77	.47	.63	.54	.595
all features	.57	.77	.52	.65	.55	.611

Figure 3 shows the confusion matrix for the model using all features on the hand transcripts, where "O" corresponds to words without punctuation. The vast majority of confusions are associated with commas.

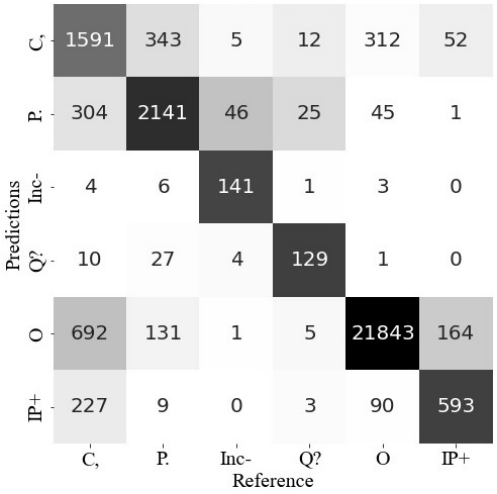

Figure 3: *Confusion matrix of the 5-class model using all features on hand transcripts for the IP test set.*

The high confusions for the comma and no-punctuation class for the 4-class model raises the question of whether results could be improved by explicitly modeling interruption points separately and then mapping them to either comma or no punctuation to obtain 4 classes. Table 5 shows that separate modeling of interruption points improves prediction when modeled as commas compared to as no punctuation.

Table 5: *Macro F1 score of 4-class punctuation prediction on the IP test set given hand transcripts, comparing the 4-class model to the result for different mappings from the 5-class result to 4 classes.*

	4-class (IP test)	5:4-class (O)	5:4-class (COMMA)
text only	.754	.758	.778
pause only	.757	.759	.778
all features	.768	.773	.793

5. Conclusions

As expected, automatically generated transcripts induce noise that contributes to the difficulty of punctuation prediction. Use of durational and acoustic features add computational cost, but do not degrade overall model performance. Prosodic features are most informative for prediction of incomplete (Inc-) boundaries, and in particular improve the macro F1 score for those tokens when coupled with automatically generated transcripts. Prediction of IPs is reasonably reliable with hand transcripts, but about as reliable as commas given ASR transcripts. Irrespective of conventions for punctuation at IPs, explicit modeling of interruption points can benefit prediction of standard punctuation.

6. References

[1] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: A lightweight punctuation annotation system for speech," in *Proc.*

ICASSP, 1998, p. 689–692.

[2] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001.

[3] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. ICSLP*, 2002.

[4] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *Proc. ICASSP*, 2009, p. 4741–4744.

[5] P. Zelasko, P. Szymanski, J. M. A. Szymczak, Y. Carmiel, and N. Dehak, "Punctuation prediction model for conversational speech," in *Proc. Interspeech*, 2018, pp. 2633–2637.

[6] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *Proc. ICASSP*, 2017, pp. 5700–5704.

[7] K. Makhija, T.-N. Ho, and E.-S. Chng, "Transfer learning for punctuation prediction," in *Proc. APSIPA Annual Summit and Conference*, 2019, pp. 268–273.

[8] R. Pappagari, P. Zelasko, A. Mikolajczyk, P. Pezik, and N. Dehak, "Joint prediction of truecasing and punctuation for conversational speech in low-resource scenarios," in *Proc. ASRU*, 2021, pp. 1185–1191.

[9] M. Sunkara, S. Ronanki, K. Dixit, S. Bodapati, and K. Kirchhoff, "Robust prediction of punctuation and truecasing for medical asr," in *Proc. Workshop on NLP for Medical Conversations*, 2020.

[10] A. Moró and G. Szaszák, "A prosody inspired RNN approach for punctuation of machine produced speech transcripts to improve human readability," in *Proc. CogInfoCom*, 2017.

[11] M. Fang, H. Zhao, X. Song, X. Wang, and S. Huang, "Using bidirectional LSTM with BERT for Chinese punctuation prediction," in *Proc. IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, 2019, pp. 1–5.

[12] Y. Guo, H. Wang, and J. Van Genabith, "A linguistically inspired statistical model for Chinese punctuation generation," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 9, no. 2, pp. 1–27, 2010.

[13] Y. Zhao, C. Wang, and G. Fu, "A CRF sequence labeling approach to Chinese punctuation prediction," in *Proc. Pacific Asia Conference on Language, Information, and Computation*, 2012.

[14] T. Levy, V. Silber-Varod, and A. Moyal, "The effect of pitch, intensity and pause duration in punctuation detection," in *Proc. IEEE Convention of Electrical and Electronics Engineers in Israel*, 2012, pp. 1–4.

[15] J. Hirschberg and C. Nakatani, "Acoustic Indicators of Topic segmentation," in *Proc. ICSLP*, 1998.

[16] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation," *Computational Linguistics*, vol. 27, pp. 31–57, 2001.

[17] J.-H. Kim and P. Woodland, "A Combined Punctuation Generation and Speech Recognition System and Its Performance Enhancement Using Prosody," *Speech Communication*, vol. 41, no. 4, pp. 563–577, 2003.

[18] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Comparing and Combining Generative and Posterior Probability Models: Some Advances in Sentence Boundary Detection in Speech," in *Proc. EMNLP*, Barcelona, Spain, 2004.

[19] J. Kolář, E. Shriberg, and Y. Liu, "Using Prosody for Automatic Sentence Segmentation of Multi-party Meetings," in *Text, Speech and Dialogue*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 629–636.

[20] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and Other Cues to Speech Recognition Failures," *Speech Communication*, vol. 43, pp. 155–175, 06 2004.

[21] J. G. Kahn, M. Lease, E. Charniak, M. Johnson, and M. Ostendorf, "Effective Use of Prosody in Parsing Conversational Speech," in *Proc. HLT/EMNLP*, 2005.

[22] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A Standard for Labeling English Prosody," in *Proc. ICSLP*, 1992.

[23] J. G. Kahn and M. Ostendorf, "Joint reranking of parsing and word recognition with automatic segmentation," *Computer Speech & Language*, vol. 26, no. 1, pp. 1–51, 2012.

[24] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proc. NAACL*, Jun. 2018, pp. 69–81.

[25] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Ess-Dykema, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[26] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van, and M. Meteer, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech," *Computational Linguistics (COLING)*, vol. 26, no. 3, pp. 339–374, 2000.

[27] T. Tran, "Neural Models for Integrating Prosody in Spoken Language Understanding," Ph.D. dissertation, University of Washington, 2020.

[28] X. He, Q. Tran, W. Havard, L. Besacier, I. Zukerman, and G. Haffari, "Exploring Textual and Speech information in Dialogue Act Classification with Speaker Domain Adaptation," in *Proc. Australasian Language Technology Association Workshop*, 2018.

[29] V.-T. Dang, T. Zhao, S. Ueno, H. Inaguma, and T. Kawahara, "End-to-End Speech-to-Dialog-Act Recognition," in *Proc. Interspeech*, 2020, pp. 3910–3914.

[30] T. Tran and M. Ostendorf, "Assessing the use of prosody in constituency parsing of imperfect transcripts," in *Proc. Interspeech*, 2021, pp. 2626–2630.

[31] J. J. Godfrey and E. Holliman, *Switchboard-1 Release 2*, Linguistic Data Consortium, 1993.

[32] D. Jurafsky, E. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13," University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO, Tech. Rep. 97-02, 1997.

[33] N. Deshmukh, A. Gleeson, J. Picone, A. Ganapathiraju, and J. Hamaker, "Resegmentation of SWITCHBOARD," in *Proc. ICSLP*, 1998.

[34] T. Zhao and T. Kawahara, "Joint Dialog Act Segmentation and Recognition in Human Conversations Using Attention to Dialog Context," *Computer Speech & Lang.*, vol. 57, pp. 108–127, 2019.

[35] T. Tran, J. Yuan, Y. Liu, and M. Ostendorf, "On the Role of Style in Parsing Speech with Neural Models," in *Proc. Interspeech*, 2019, pp. 4190–4194.

[36] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, 2015.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[38] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011.

[39] K. Cho, B. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proc. EMNLP*, 2014.

[40] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[41] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Proc. Interspeech*, 2016.

[42] C. Cieri, D. Graff, O. Kimball, D. Miller, and K. Walker, "Fisher English training speech part 1 transcripts, ldc2004t19," 2004.