Paragraph-based Prosodic Cues for Speech Synthesis Applications

Paragraph-based Prosodic Cues for Speech Synthesis Applications

Mireia Farrús[1], *Catherine Lai*[2], *Johanna D. Moore*[2]

[1]N-RAS Research Centre, Universitat Pompeu Fabra, Barcelona, Spain
[2]School of Informatics, University of Edinburgh, Edinburgh, UK
mireia.farrus@upf.edu, clai@inf.ed.ac.uk, j.moore@ed.ac.uk

Abstract

Speech synthesis has improved in both expressiveness and voice quality in recent years. However, obtaining full expressiveness when dealing with large multi-sentential synthesized discourse is still a challenge, since speech synthesizers do not take into account the prosodic differences that have been observed in discourse units such as paragraphs. The current study validates and extends previous work by analyzing the prosody of paragraph units in a large and diverse corpus of TED Talks using automatically extracted F0, intensity and timing features. In addition, a series of classification experiments was performed in order to identify which features are consistently used to distinguish paragraph breaks. The results show significant differences in prosody related to paragraph position. Moreover, the classification experiments show that boundary features such as pause duration and differences in F0 and intensity levels are the most consistent cues in marking paragraph boundaries. This suggests that these features should be taken into account when generating spoken discourse in order to improve naturalness and expressiveness.

Index Terms: discourse unit, prosodic cue, paragraph boundary, speech synthesis.

1. Introduction

Over the last decade, automatically generated speech has improved significantly, especially in terms of voice quality and expressiveness. However, although a good deal of effort has been spent on building synthetic voices from large multi-paragraph speech databases [1, 2, 3], multi-sentential synthesized speech still suffers from a high degree of unnaturalness. Current Text-To-Speech (TTS) systems generally attempt to improve naturalness and expressiveness by increasing the prosodic variability based on inferred affective states [4, 5]. However, it appears that simply varying sentence prosody, without accounting for paragraph structure, does not necessarily improve naturalness [6]. Instead, it seems that a more discourse structure aware approach is needed to make real improvements in speech synthesis.

Prosodic changes related to discourse in both conversational and read speech have long been observed in phonetic studies. For instance, pitch resets—higher F0 and increased F0 range—are usually found at the beginning of the discourse unit, while pitch declination has been observed across the discourse unit, ending with low F0 [7]. Discourse structure has also been associated with energy and timing features such as pausing, pre-boundary lengthening, and speech rate variations [8, 9]. However, most linguistically oriented studies have been based on small corpora, with relatively small numbers of speakers, and in limited domains. Moreover, few studies have quantified the predictiveness or robustness of these prosodic relationships.

In this paper, we extend previous findings in three aspects. First, we confirm the presence of prosodic cues for paragraph structure in a large, varied corpus of semi-spontaneous speech. To this end, we analyzed a corpus consisting of more than 1300 TED talks and 1100 different speakers. Second, we analyze prosodic patterns in terms of automatically extracted F0, intensity, and timing features of sentences with respect to paragraph position, as well as across paragraph boundaries. Finally, we perform SVM classification experiments to investigate how consistently and robustly specific prosodic cues appear at paragraph boundaries. Taking the results into account could potentially help determine which cues should be implemented in a speech synthesizer in order to improve naturalness and expressiveness of multi-sentential discourses.

The structure of this paper is as follows. Section 2 overviews related work on prosody and discourse structure. Section 3 describes our data set and experimental setup. Results of statistical analyses and classification experiments are described in Sections 4 and 5. We discuss the implications of these results for speech synthesis in Section 6 and conclude in Section 7.

2. The Prosody of Discourse Segments

Prosodic marking of discourse boundaries has been identified in a wide range of instrumental phonetic studies. Early work focused on sentence and paragraph breaks [10, 11], but studies have since investigated prosodic features related to topic changes [12, 13, 14, 15], as well as boundaries related to hierarchical discourse structure [16, 17]. The results of these investigations generally lead to similar conclusions: a speaker's pitch declines through both intra- and supra-sentential segments, with low boundary tones and laryngealization signalling finality, and mid-level pitch interpreted as continuation or floor holding [11, 12, 18, 19]. Pitch resets are also often observed after discourse boundaries, i.e., increases in pitch level and pitch range at the beginning of the new segment [20, 21, 22]. Similar energy reset patterns have been reported impressionistically [11] or in terms of RMS amplitude [20, 23, 24]. Besides pitch and energy based features, pause length and pre-boundary lengthening have also been identified as boundary indicators [10, 8, 15].

The accumulated results suggest that prosodic boundaries share similar features across discourse levels. So, for example, we expect to see similar prosodic features at sentence internal phrase boundaries and at topic boundaries. Moreover, we expect those prosodic features to be more pronounced for larger units. In line with this, [21] found pause, pitch reset, and boundary tones to be correlated with number of annotators who identified the position as a boundary. Better annotator agreement has also been observed for discourse segmentation when audio is available [21, 23]. In this vein, [25] show that simple rule

based manipulations of F0 topline and baseline and pauses can improve naturalness ratings of synthesized news items.

Although pitch reset and suprasentential pitch declination have been consistently reported as boundary cues, these results have often been based on a mixture of qualitative descriptions of prosodic patterns and signal based measurements, e.g., [11, 18, 24]. This makes the robustness of actual measurable features for signalling boundaries somewhat unclear. Even in more quantitative work, the way prosodic concepts are measured varies from study to study. Such terminological differences can lead to different conclusions about the utility of features. For example, [20] and [21] use F0 maximum as a proxy for pitch range, while other studies use topline/baseline or specified quantile differences [25, 14]. Studies that use the former measure found 'pitch range' to have a stronger relationship with discourse structure than those that use the the latter.

Using measureable, operational definitions of features is important to understanding how robust prosodic features are for signalling discourse units. Similarly, we'd like to know how well these results generalize to larger sets of speakers and speaking styles. Linguistically oriented studies generally only examine small amounts of data, usually from a restricted domain (e.g., reading aloud of constructed examples or small domain task-oriented dialogues). However, a number of studies have successfully employed prosody and timing features for discourse segmentation using larger data sets, e.g., topic segmentation of broadcast news [26, 27, 28]. Unsurprisingly, features in these studies are usually based around the idea of prosodic reset and differences in pitch ranges, but quantified directly from the speech signal. These segmentation algorithms tend to focus on boundary features, e.g., prosodic statistics of the words immediately adjacent to the boundary [26, 29] rather than declination or other features of the larger discourse unit. For example, [26] find pause duration and F0 range of the word preceding the boundary to be the most discriminative feature for topic boundaries in broadcast news.

Automatic discourse segmentation studies usually aim at detecting at high level discourse units, e.g., changes in news stories or tasks drawn from a pre-defined set [30, 31]. These changes are often signalled by more abrupt, 'disjunctive' boundaries [32], which we would expect to occur infrequently while being more prosodically marked. From a speech synthesis point of view, we would like to be able to generate prosody at a more fine-grained level of discourse. Paragraphs are a good match for this line of inquiry as they are often available in the texts we wish to synthesize while topic segmentations and detailed discourse relations are not (e.g., TTS for audio books). Moreover, their reality as a discourse unit has been established [33], but they are still reasonably theory neutral.

Since discourse level prosodic changes will affect how sentence internal prosody is interpreted, we would like to know how sentence prosody varies within paragraphs, as well as how prosody changes at discourse boundaries. In particular, we would like to know if we can map paragraph positions to intrinsic properties of sentences themselves (cf., [18, 25, 24]), or whether we always need to look at relative differences. In general, we would like to identify robust aspects of paragraph prosody that can be applied to longer spoken discourses to increase naturalness, intelligibility, and expressiveness. In the current work, we focus on spoken monologues since, at this stage, we would like to avoid conflation with prosodic features related to turn-taking/floor-holding [11, 18]. The following sections describe experiments examining the paragraph patterns formed by automatically extracted prosodic features and their relation to notions of declination and prosodic reset.

3. Experimental Setup

3.1. Data

In this study, we examine a set of 1365 TED (Technology, Entertainment, Design) talks published before 2014[1]. The data set includes 1156 different speakers of English with various accents, which means that some of the speakers present more than one talk. These talks span a wide variety of topics ranging from science and technology, to international development, to the fine arts. Talks are 15 minutes long on average. Most talks have one main speaker, although guests and audience members occasionally speak in some talks. Each talk is manually transcribed, including punctuation and paragraph breaks. While there are no hard rules for determining paragraphs, transcribers do attend to the audio stream when determining paragraph breaks.[2] Altogether, the data set includes 151820 sentences and 20953 paragraphs, with an average of 7 sentences per paragraph.

TED talks are well known for being polished and entertaining. We consider these talks semi-spontaneous speech since the material is prepared in advance. They are generally well structured and delivered so as to be engaging, convincing, and easy to follow in spoken form. As such, they present desirable properties for a speech synthesis system to model. Impressionistically, speakers vary greatly in style of their delivery. So, we expect features that are indicative of paragraph breaks in this data set to be robust across a range of lecturing styles.

3.2. Feature extraction

3.2.1. Sentence Alignment

TED transcripts come with broad subtitle timings that do not necessarily correspond to sentences in the transcript. To obtain sentence timings, we first obtain precise word timings through Viterbi forced alignment using an automatic speech recognition system. Word timings are then used to automatically obtain sentence boundaries.

3.2.2. Prosodic Features

F0 and intensity contours were extracted using Praat at 10ms intervals with linear interpolation and octave jump removal for F0 [34]. For F0, parameter settings were automatically determined using the method described in [35]. F0 and intensity values were normalized over talks so that zero values represent speaker mean values: intensity measurements (I) were normalized by subtracting the speaker mean for the talk, while F0 values were converted to semitones relative to speaker mean F0 value (Hz), i.e., log scaled to better represent pitch perception.

We calculated aggregate statistics over sentence units: mean, standard deviation, maximum, minimum, median, and slope. We also look at the difference between the first and last words of the sentence (fldiff) and the difference between the 99th and 1st quantile values (range). To capture contextual changes, we measured the difference between the target and the following sentence (sdiff), as well as the difference between the last word of the target sentence and the next sentence (lndiff). For these features, we measure the difference from the word or sentence occuring later, so positive difference values indicate a reset to a higher level. In the current work, we only look at differences in means.

[1] `http://www.ted.com`

[2] p.c. TED translation team.

	First	Middle	Last
F0 (semitones)			
max.norm.F0	8.43	7.59	7.40
mean.norm.F0	-1.07	-1.68	-2.14
median.norm.F0	-1.37	-1.90	-2.41
min.norm.F0	-10.48	-11.12	-11.63
range.norm.F0	17.70	16.47	17.04
sd.norm.F0	4.46	4.18	4.26
slope.norm.F0	-2.37	-2.09	-2.00
fldiff.norm.F0	-6.28	-5.12	-5.78
intensity (dB)			
max.norm.I	12.74	12.54	12.52
mean.norm.I	0.40	0.32	-0.22
median.norm.I	2.29	2.36	1.68
min.norm.I	-19.92	-20.22	-20.85
range.norm.I	30.80	30.22	30.63
sd.norm.I	8.47	8.30	8.38
slope.norm.I	-1.11	-0.94	-0.92
fldiff.norm.I	-3.32	-2.80	-3.34
timing (words/s)			
spk.rate	3.10	3.28	3.12

Table 1: *Sentence feature statistics by paragraph position.*

In addition to F0 and intensity features we measured the speaking rate (spk.rate) as the number of words in the sentence divided by the sentence duration, and the pause duration (pause.dur) before the next sentence.

3.3. Statistics and classification

To understand the relationship between paragraph breaks and prosodic features we first performed a statistical analysis of individual prosodic features in sentences that appear in first, middle, and last positions of paragraphs. One-sentence paragraphs are treated as paragraph initial [25]. We then performed classification experiments to get a better idea of how well features can be used to discriminate these classes.

The SVM classification experiments were performed by using 10-fold cross-validation. The SVM models were built with LibSVM [36] as integrated in the Weka machine learning toolkit [37], using the C-SVC approach with a RBF kernel, setting the cost parameter to $C = 1$ and gamma to $\gamma = 0$, and using the same parameters in all folds. The size of the *no break* outnumbered the *break* class with an approximate ratio of seven to one. In order to improve the classifiers given this class imbalance, downsampling was performed [38] by randomly selecting the same number of instances that we had for the *break* class.

4. Feature Statistical Analysis

In this section we present the statistics for the sentence and boundary based prosodic features described in Section 3. The differences between the three paragraph positions (first, middle, last) were found to be statistically significant for the individual prosodic features (t-test, $p < 0.05$) except for sd.norm.F0, slope.norm.F0, max.norm.I, slope.norm.I when comparing paragraph middle and last sentences.

Table 1 shows the mean values of sentence-based prosodic features in different paragraph positions. Some clear recurrent patterns can be seen in these sentence-based features. These are exemplified in Figure 1. The first pattern can be characterized by a decrease in feature values through the paragraph. This is

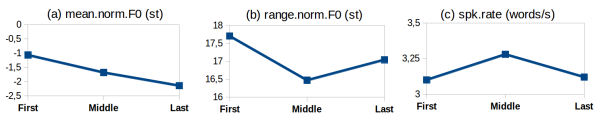

Figure 1: *Different patterns for sentence-based prosodic features: (a) mean.norm.F0, (b) range.norm.F0, and (c) spk.rate.*

	no break	break
sdiff.norm.F0 (st)	-0.22	0.91
lndiff.norm.F0 (st)	5.01	7.02
sdiff.norm.I (dB)	-0.13	0.45
lndiff.norm.I (dB)	2.73	3.88
pause.dur (words/s)	0.64	1.80

Table 2: *Boundary feature statistics by sentence type.*

shown for mean.norm.F0 in Figure 1a, but was also found in all features representing F0 and intensity level (i.e., mean, maximum, minimum and median) except for median.norm.I. This pattern confirms that declination in prosodic level occurs over paragraph units.

The second pattern (Figure 1b) applies to variation-related features such as range, sd, and the absolute values for fldiff and slope. Here we see reduced values in the middle position, suggesting range compression in the paragraph medial sentences. Note, on first glance, the slope values follow the declination pattern (1a). However, the difference between middle and last sentences was not statistically significant (t-test, $p > 0.05$). Thus, we see a general reduction in the range of variation after the first sentence of a paragraph. Interestingly, however, we do not see relative pitch range compression in paragraph final sentences relative to middle sentences.

Finally, the third pattern, represented by spk.rate in Figure 1c (but also observed for median.norm.I) sees significant increases in the middle position. For speaking rate, this suggests a word lengthening effect in the initial and final positions, which is consistent with pre- and post-boundary lengthening associated with prosodic phrases within sentences [39].

Table 2 shows mean values for boundary-based features obtained looking at sentences in initial and middle positions (*no break*) versus final position (*break*). The differences in boundary features in both *no break* and *break* classes are all statistically significant (t-test, $p < 0.05$). While a decrease of F0 and intensity is evident when there is no change of paragraph, a clear rise is encountered at the paragraph break in terms of F0, intensity and pause duration. That is, we see clear evidence of prosodic resets over paragraph breaks.

5. Classification Experiments

In this section we present the SVM classification accuracy obtained for all the individual prosodic features (Table 3) and combinations of sets of features: the two and three best performing features, boundary and sentence-based features, and the whole set feature set (Table 4). The area under ROC curve (AUC) is also provided, as well as the feature performance ranking.

Unsuprisingly, the results reflect the fact that difference features are more discriminative of boundaries than sentence-based features—four out of five boundary features are ranked in the top positions. Feature classification results in isolation do not show high accuracy (the best performance is achieved by pause

Feature	Accuracy (%)	AUC	rank
sentence-based features			
max.norm.F0	51.02	0.510	19
mean.norm.F0	53.65	0.536	8
median.norm.F0	54.14	0.541	5
min.norm.F0	52.75	0.527	11
range.norm.F0	51.32	0.513	17
sd.norm.F0	50.63	0.506	21
slope.norm.F0	53.11	0.531	10
fldiff.norm.F0	52.13	0.521	15
max.norm.I	50.97	0.510	20
mean.norm.I	53.98	0.540	7
median.norm.I	54.01	0.540	6
min.norm.I	52.25	0.523	14
range.norm.I	51.47	0.513	16
sd.norm.I	51.10	0.511	18
slope.norm.I	52.68	0.527	12
fldiff.norm.I	50.45	0.505	22
spk.rate	52.60	0.526	13
boundary features			
sdiff.norm.F0	56.73	0.567	2
lndiff.norm.F0	56.65	0.566	3
sdiff.norm.I	54.51	0.545	4
lndiff.norm.I	53.52	0.535	9
pause.dur	**62.09**	0.621	1

Table 3: *SVM classification of individual prosodic features.*

Feature set	Accuracy (%)	AUC
2 best	62.87	0.629
3 best	62.80	0.628
5 boundary	61.26	0.613
17 sentence	54.17	0.542
22 all	56.73	0.567

Table 4: *SVM classification of selected feature sets.*

duration with 62% accuracy, baseline 50%). This is almost certainly due to the large variance in each class. Nevertheless, the results again indicate that prosodic level resets are more indicative of paragraph ends than range changes: median values were the most predictive sentential features, while standard deviations were less predictive.

6. Discussion

The current study aimed to determine which prosodic features robustly signal paragraph structure in a large and varied corpus. The statistical analysis highlighted many significant differences over the different paragraph positions, especially in features indicating prosodic level of the target sentences. As in previous studies, these features indicated steady declination in prosodic level over the paragraph. In addition, the variation-related features had lower values in middle positions, suggesting range compression for both pitch and intensity in the middle of the paragraph, with no evidence of pitch range compression specific to the end of paragraphs. Additionally, speaking rate tended to increase mid-paragraph suggesting that first and last sentences are prosodically marked in terms of speaking rate.

In general, these findings suggest that paragraphs do have a basic, identifiable suprasentential prosodic structure that we can describe in terms of relative changes in F0, intensity, and timing. The idea that there are utterance intrinsic features to paragraph position is supported by the classification experiments for pitch level features. As suggested by the statistical analysis, range type features fill the last positions in the classification results. This suggests that these variation-related features mark paragraph internal structure rather than boundaries. Furthermore, these results indicate that using F0 maximum as a proxy for range or level can cloud our understanding of how discourse structure is manifested prosodically.

Overall, pause duration appears to be the most robust predictor of paragraph breaks. However, performance is still low, especially compared to previous topic segmentation results. One reason for this may be the subtle relationship between topic and paragraph transitions in TED talks. Another reason is that prosody conveys more than just discourse segment information. From a structural perspective, more analysis of discourse relations, finer grained topic changes, and sentence information structure, for example, are necessary to gain a fuller understanding of the relationship between prosody and discourse. From a TTS perspective, however, the current findings suggest that we should be able to immediately employ paragraph declination, pause, and prosodic reset features to improve the naturalness of longer synthesized discourses. Moreover, while current approaches often try to directly link prosodic level with text sentiment or emotion, the current findings suggests that prosodic levels are indicative of a discourse structure. So, controlling for these structural changes should help remove existing confounds in developing realistic models of prosodic expressiveness.

7. Conclusions

Paragraphs in spoken discourse carry a variety of information. Prosodically, we have seen that they are characterized by general properties like declination and prosodic reset even in a very stylistically diverse corpus. Characterizing prosodic patterns of paragraphs should be useful for generating more natural and expressive speech. Beyond this, the findings of this study are also applicable for paragraph segmentation in automatic transcription systems. However, to truly capitalize on prosodic knowledge for recognition or generation we need to account for other sources of variation such as emotion and information structure. Furthermore, we need to understand how they interact with higher level discourse structure.

In this study, our goal was to determine which prosodic features consistently describe paragraph structure in order to improve prosody in speech synthesis. However, it is clear that we need to implement these findings in a text-to-speech synthesis system and perform perception experiments to really validate them. Future work will also include trying other classifiers based on random forests or sequential models such as long short term memory (LSTM) algorithms or conditional random fields (CRF) to better understand how we can combine features, contextual information, and the sequential nature of communication.

8. Acknowledgements

We'd like to thank Peter Bell for providing the word timings. Part of this work has received funding from the *EU's Horizon 2020 Research and Innovation Programme* under the GA H2020-RIA-645012. The first author is partially funded by the Spanish Ministry of Economy and Competitivity through the *Juan de la Cierva* program and a *José Castillejo* mobility grant.

9. References

[1] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *Proceedings of Interspeech 2007*, 2007, pp. 2901–2904.

[2] J. Y. Zhang, A. R. Toth, K. Collins-Thompson, and A. W. Black, "Prominence prediction for super-sentential prosodic modeling based on a new database," in *ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2004.

[3] K. Prahallad, A. Black, and R. Mosur, "Sub-Phonetic Modeling For Capturing Pronunciation Variations For Conversational Speech Synthesis," in *Proceedings of ICASSP 2016*, 2006.

[4] M. Charfuelan, "MARY TTS HMMbased voices for the Blizzard Challenge 2012," in *Blizzard Challenge Workshop 2012*, 2012.

[5] L. Chen, M. Gales, N. Braunschweiler, M. Akamine, and K. Knill, "Integrated automatic expression prediction and speech synthesis from text," in *Proceedings of ICASSP 2013*, 2013, pp. 7977–7981.

[6] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proceedings of Interspeech 2015*, 2015.

[7] J. Cole, "Prosody in context: a review," *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, Feb. 2015.

[8] C. L. Smith, "Topic transitions and durational prosody in reading aloud: production and modeling," *Speech Communication*, vol. 42, no. 34, pp. 247–270, Apr. 2004.

[9] M. Swerts and R. Geluykens, "Prosody as a Marker of Information Flow in Spoken Discourse," *Language and Speech*, vol. 37, no. 1, pp. 21–43, Jan. 1994.

[10] I. Lehiste, "Some Phonetic Characteristics of Discourse," *Studia Linguistica*, vol. 36, no. 2, pp. 117–130, 1982.

[11] J. Kreiman, "Perception of sentence and paragraph boundaries in natural conversation," *Journal of Phonetics*, vol. 10, no. 2, pp. 163–175, 1982.

[12] S. Nakajima and J. F. Allen, "A Study on Prosody and Discourse Structure in Cooperative Dialogues," *Phonetica*, vol. 50, no. 3, pp. 197–210, 1993.

[13] M. Swerts and M. Ostendorf, "Prosodic and lexical indications of discourse structure in human-machine interactions," *Speech Communication*, vol. 22, no. 1, pp. 25–41, Jul. 1997.

[14] C. de Looze and S. Rauzy, "Automatic detection and prediction of topic changes through automatic detection of register variations and pause duration," in *Proceedings of Interspeech 2009*, 2009, pp. 2919–2922.

[15] M. Zellers and B. Post, "Fundamental frequency and other prosodic cues to topic structure," *Proceedings of IDP 2009*, 2009.

[16] G. Möhler and J. Mayer, "A discourse model for pitch-range control," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[17] H. den Ouden, L. Noordman, and J. Terken, "Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports," *Speech Communication*, vol. 51, no. 2, pp. 116–129, 2009.

[18] R. Geluykens and M. Swerts, "Prosodic cues to discourse boundaries in experimental dialouges," *Speech Communication*, vol. 15, no. 12, pp. 69–77, Oct. 1994.

[19] C. de Looze, I. Yanushevskaya, A. Murphy, E. O'Connor, and C. Gobl, "Pitch Declination and Reset as a Function of Utterance Duration in Conversational Speech Data," in *Proceedings of Interspeech 2015*, 2015.

[20] B. Grosz and J. Hirschberg, "Some intonational characteristics of discourse structure," in *Proceedings of ICSLP*, 1992, pp. 429–432.

[21] M. Swerts, "Prosodic features at discourse boundaries of different strength," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 514–521, Jan. 1997.

[22] C.-Y. Tseng, Z.-Y. Su, C. Chang, and C.-h. Tai, "Prosodic Fillers and Discourse Markers-Discourse Prosody and Text Prediction," in *Proceedings of TAL 2006*, 2006, pp. 27–29.

[23] J. Hirschberg and C. H. Nakatani, "A Prosodic Analysis of Discourse Segments in Direction-giving Monologues," in *Proceedings of ACL'96*, 1996, pp. 286–293.

[24] R. Herman, "Phonetic markers of global discourse structures in English," *Journal of Phonetics*, vol. 28, no. 4, pp. 466–493, Oct. 2000.

[25] A. Sluijter and J. Terken, "Beyond Sentence Prosody: Paragraph Intonation in Dutch," *Phonetica*, vol. 50, no. 3, pp. 180–188, 1993.

[26] E. Shriberg, A. Stolcke, D. Hakkani-Tr, and G. Tr, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 12, pp. 127–154, Sep. 2000.

[27] J. Hirschberg and C. H. Nakatani, "Acoustic indicators of topic segmentation," in *Proceedings of ICSLP 1998*, Sydney, 1998.

[28] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals, "Maximum entropy segmentation of broadcast news," in *Proceedings of ICASSP 2005*, 2005.

[29] G.-A. Levow, "Assessing Prosodic and Text Features for Segmentation of Mandarin Broadcast News," in *Proceedings of SpeechIR'04*, 2004, pp. 28–32.

[30] P.-Y. Hsueh and J. D. Moore, "Combining multiple knowledge sources for dialogue segmentation in multimedia archives," in *Proceedings of ACL 2007*, 2007.

[31] G.-A. Levow, "Prosodic cues to discourse segment boundaries in human-computer dialogue," in *Proc. of SIGdial*, 2004.

[32] M. Zellers and B. Post, "Combining Formal and Functional Approaches to Topic Structure," *Language and Speech*, vol. 55, no. 1, pp. 119–139, Mar. 2012.

[33] C. Sporleder and M. Lapata, "Broad Coverage Paragraph Segmentation Across Languages and Domains," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 2, pp. 1–35, Jul. 2006.

[34] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2001.

[35] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction." *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2291–2291, Oct. 2010.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[38] Y. Liu, N. V. Chawla, M. P. Harper, E. Shriberg, and A. Stolcke, "A study in machine learning from imbalanced data for sentence boundary detection in speech," *Computer Speech & Language*, vol. 20, no. 4, pp. 468–494, Oct. 2006.

[39] D. Byrd, J. Krivokapi, and S. Lee, "How far, how long: On the temporal scope of prosodic boundary effectsa)," *The Journal of the Acoustical Society of America*, vol. 120, no. 3, pp. 1589–1599, 2006.