Incorporating Prosodic Boundaries in Unsupervised Term Discovery

Bogdan Ludusan[1], *Guillaume Gravier*[2], *Emmanuel Dupoux*[1]

[1]LSCP - EHESS/ENS/CNRS, Paris
[2]IRISA - CNRS, Rennes

bogdan.ludusan@ens.fr, guillaume.gravier@irisa.fr, emmanuel.dupoux@gmail.com

Abstract

We present a preliminary investigation on the usefulness of prosodic boundaries for unsupervised term discovery (UTD). Studies in language acquisition show that infants use prosodic boundaries to segment continuous speech into word-like units. We evaluate whether such a strategy could also help UTD algorithms. Running a previously published UTD algorithm (MODIS) on a corpus of prosodically annotated English broadcast news revealed that many discovered terms straddle prosodic boundaries. We then implemented two variants of this algorithm: one that discards straddling items and one that truncates them to the nearest boundary (either prosodic or pause marker). Both algorithms showed a better term matching F-score compared to the baseline and higher level prosodic boundaries were found to be better than lower level boundaries or pause markers. In addition, we observed that the truncation algorithm, but not the discard algorithm, increased word boundary F-score over the baseline.

Index Terms: term discovery, prosody, prosodic boundary

1. Introduction

During their first year of life, human infants extract word-like units from continuous speech without supervision [1]. In parallel, unsupervised term discovery (UTD) algorithms are increasingly used within speech technology [2, 3, 4, 5]. In both cases, the task consists in finding repetitive patterns, while using as input only the speech signal. An examination of how infants are solving this task may reveal useful strategies that could be implemented into UTD algorithms.

Many researchers have pointed out that prosody is an important cue that helps infants to segment continuous speech. Newborns are sensitive to the acoustic cues correlated with the presence or absence of phonological phrase boundaries in otherwise identical stretches of speech (e.g. /mati/ in "mathématicien" (mathematician) versus "panorama typique" (typical panorama) [6, 7]. Nine-month-olds use these cues to posit breaks within sentences [8, 9]. Ten- and thirteen-month old infants use these boundaries to constrain word recognition, i.e. they fail to recognize a string that straddles a phonological phrase boundary [10]. Similarly, adults use these boundaries to constrain online lexical cognition, i.e. they do not produce false alarms on word forms that straddle a phonological phrase boundary[11].

Unsupervised term discovery, so far, does not use prosodic information. The existing algorithms are based on computing a similarity score between stretches of speech signal, usually done by means of dynamic time warping (DTW). The proposed systems return a list of matched pairs [2, 3] and/or a library of clusters of discovered terms [2, 4, 5]. Some systems scan the entire corpus for repetitions [2, 3], while others only scan a small time buffer and match the signal against an incrementally built library of terms [4, 5]. The terms discovered with UTD systems have already been proven useful in a number applications like keyword spotting [12], topic segmentation [13], or document classification [14].

Based on the findings regarding the role of prosodic boundaries in speech processing for both children and adults, we investigated the use of such boundaries in unsupervised term discovery. The current study uses manually annotated prosodic boundaries in order to establish the upper boundary of their impact on the discovery task. The rest of the paper is organized as follows: the UTD system used in the experiments and the evaluation method are presented in section 2, while a short description of the corpus employed in this study is given in section 3. Two experiments are illustrated in section 4, in which we varied the strategy for using prosodic information. The paper concludes with a discussion of the results obtained and some possible paths to follow.

2. Methods

2.1. System Presentation

An open source system for spoken term discovery, MODIS [15], was employed for the experiments. It is based on a generic approach to mining repeating sequences, tolerant to term variability [5], and it uses a limited search buffer, making it more psychologically realistic than systems performing an exhaustive search.

MODIS takes a speech signal as input (represented as either MFCCs or posteriorgrams) and delivers a library of repeated terms as its output. Term discovery is based on the notion of seed fragments. A seed fragment is a stretch of signal segmented from the input stream and searched for in a fixed-length buffer ahead of the seed using a segmental variant of the DTW algorithm. If a match for the seed is found, the seed is extended to find the maximal length matching pattern and, if it exceeds a minimal term length, it is stored in the library of terms. This library is used as a long term memory to search for repeating terms: Each new seed considered is first matched against entries in the library before searching for self-repetitions in the buffer. Potential re-occurrences detected by DTW are validated using self-similarity matrix (SSM) comparison.

The key parameters of the seeded discovery algorithm are the seed and term lengths and a set of similarity thresholds used to validate template comparison (higher thresholds means potentially more variability at the expense of precision).

2.2. Evaluation Method

We evaluated two aspects of the discovered terms: the matching quality and the word boundary quality. For the matching

10.21437/SpeechProsody.2014-176

quality, we used a similar method to [16], which transforms the speech chunks corresponding to the found terms into a symbolic representation, based on the phonetic transcription of the speech signal. Then, the precision and the recall are determined based only on the strings of phonemes corresponding to the obtained terms. For a formal definition of the measures used for the evaluation, see Muscariello et al study [16]. The following steps are performed during the evaluation process:

- All phonemes falling inside the time interval corresponding to the found term are concatenated. A phoneme is considered to belong to the term if at least 50% of its duration falls within the term.
- For each class of terms, a centroid is computed, defined as being the string with the lowest normalized edit distance from all the other strings belonging to the class.
- The precision is calculated as the percentage of class members, out of the total number of tokens in the class, falling within a certain distance from the class centroid. The neighbourhood threshold was set to 0.2.
- Next, the recall is determined as being the percentage of how many strings belonging to the centroid neighbourhood were found, from the total number of occurrences of those strings in the whole corpus.

The second measure we evaluated, the word boundary quality, comes from the field of natural language processing (NLP) and it can be a useful measure when the terms discovered by the UTD systems are used in a downstream application. The boundary quality was computed by comparing the set of discovered term boundaries to the set of gold word boundary of the corpus, as done in [17]. We expect a very low recall on this metric, since UTD systems do not attempt to exhaustively segment a corpus, contrary to NLP systems, that perform term discovery based on text input.

3. Materials

The materials used in this paper are a subset of the Boston University Radio News Corpus (BU corpus) [18], which contains news stories recorded by 7 professional speakers. Out of the whole corpus, around 3.5 hours of data are annotated prosodically for phrase breaks and accent tones. The prosodic annotation is based on the ToBI system [19] for American English, which uses a 5-level scheme for prosodic boundaries of increasing strength, starting with cliticized word boundaries (level 0) and ending with intermediate phrase boundaries (level 3) and intonational phrase boundaries (level 4).

We used only levels 3 and 4 as we are interested in the effect of prosodic boundaries that can, in principle, be detected in an unsupervised fashion. We removed the recordings for which these two levels were missing and those with no phone-level segmentation, because this type of annotation was necessary for the evaluation procedure. Thus, for our experiments, we used about 3 hours of data, including 6 speakers (3 males, 3 females) distributed into 403 files. In these materials there were a total of 6059 intonational phrase boundaries and 2731 intermediate phrase boundaries annotations.

4. Experiments

We propose to investigate the usefulness of prosodic boundary information for term discovery, by integrating this type of information in MODIS in two experiments. In Experiment 1, we examined the idea of using boundary information to prune away

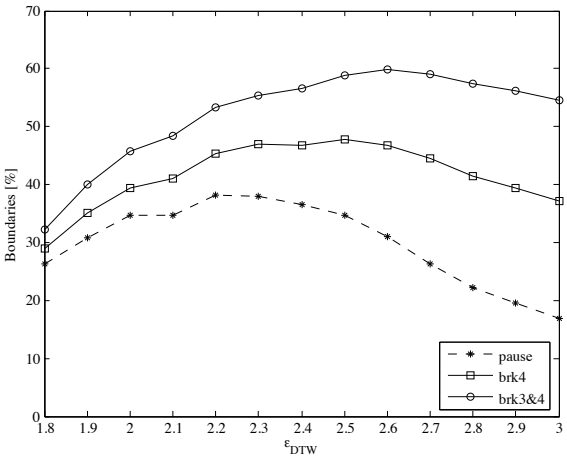

Figure 1: *Percentage of terms found by the baseline system overlapping pause markers, level 4 boundaries and level 3 and 4 boundaries.*

the terms found to be straddling a prosodic boundary ($discard$). In Experiment 2, we tested the option of truncating the terms instead of discarding them ($truncate$). In both experiments, besides intonational phrase boundaries ($brk4$) and intonational plus intermediate phrase boundaries ($brk3\&4$) we also tested another cue which denotes finality - the silent pause ($pause$). Pauses were chosen as they correlate well with prosodic boundaries, but are easier to extract automatically. In order to perform a fair comparison to the case when prosodic boundaries are used, the pauses were extracted from the manual transcription and were defined as a time instant (the beginning of the pause). We considered to be a pause all silent regions of speech having a length of at least 200 ms, resulting in 2723 pause boundaries. The pause markers correspond mostly to level 4 boundaries, but there are some which indicate level 3 boundaries.

The speech signal was represented by standard spectral features: 12 MFCCs plus energy and their first and second order temporal derivatives. The system used a seed length of 250 ms, a 90 second future buffer when searching for terms and it accepted a candidate as a term if it was at least 500 ms long. A found term was represented by its median occurrence, i.e. the token closest to all the other ones in terms of a dissimilarity score and SSM checking was also employed. In the two experiments done we varied the similarity threshold used by the DTW algorithm (ϵ_{DTW}) in the range [1.8, 3.0] and we reported the results for all the values.

4.1. Experiment 1

For the baseline system, spoken term discovery was only constrained by file boundaries (403 markers in total), which were processed by the algorithm in the same manner as the prosodic boundaries (here, discarded). Figure 1 shows the percentage of terms found with the baseline system which straddle a level 4 break, a level 3 or 4 break or a pause marker. One can observe that up to 60% of the terms straddle either a level 3 or a level 4 prosodic boundary. Given that prosodic boundaries match with constituent boundaries, it is likely that the straddling terms will be less meaningful to downstream applications. In addition, it is probable that such terms are purely coincidental, and therefore correspond to low quality clusters of word fragments. We will

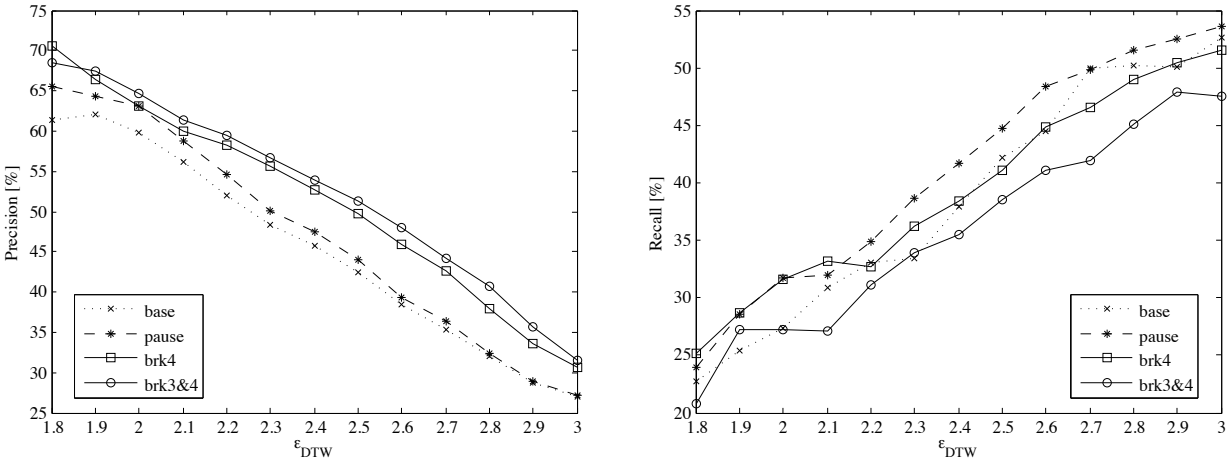

Figure 2: *Term matching precision (left panel) and recall (right panel) obtained for different ϵ_{DTW} values and various types of information used: baseline, pause information, level 4 boundaries, and level 3 and 4 prosodic boundaries.*

evaluate this premise next.

In order to measure the change in performance after the removal of straddling terms, compared to the baseline, we used the precision and recall, as computed with the method introduced in section 2.2. The results obtained with our baseline, and the same system employing prosodic or pause boundaries to discard straddling terms are illustrated in Figure 2. The left panel shows the precision obtained, while the right panel presents the recall rate. The baseline is represented with a dotted line, the system taking advantage of pause markers with a dashed line, and the results obtained with the prosodic boundary informations by a continuous line. The square represents the results for $brk4$, while the circle illustrates the $brk3\&4$ system.

It can be seen that, by adding extra boundary information into the system, besides file boundaries, the terms found are consistently more accurate. The average increase in precision is 6.0% for $brk4$, 7.3% for $brk3\&4$, and 1.8% for $pause$. In terms of recall, the system using intonational phrase boundaries gives similar performances to the baseline, which, in turn, performs better than the system having knowledge of both types of prosodic phrase boundaries, but worse than when pauses are known.

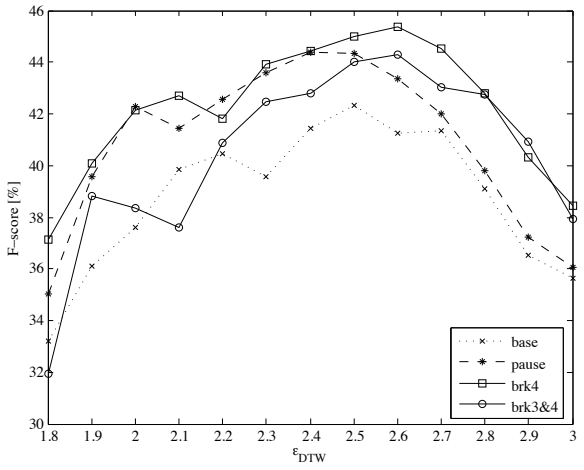

Figure 3: *Term matching F-score obtained for different values of ϵ_{DTW} and various type of information used: baseline, pause information, level 4 boundaries, and level 3 and 4 prosodic boundaries.*

We also looked at the F-score curve over the different ϵ_{DTW} values tested (Figure 3). When comparing the baseline with the other three systems, one sees a consistent improvement throughout the range of values investigated, resulting in an average F-score increase of 3.4% for $brk4$, 1.6% for $brk3\&4$ and 2.1% for the $pause$ system. We observed similar $brk4$ or worse $brk3\&4$ performance than for pauses at low values of the DTW threshold, but consistently better results for boundaries at high values of ϵ_{DTW}. This demonstrates that, when more heterogeneous terms are found, the boundaries tend to help discriminate better between occurrences of different terms, showing that prosodic boundaries encode more information than the pauses.

For an overall comparison of performance, we have summarized in Table 1 the two measures used for evaluation: the goodness of the obtained terms and the word boundaries discovered by the terms. The number in each cell represents the average precision, recall and F-score computed over all DTW threshold values (ϵ_{DTW}). The term matching measurements illustrated in the table mirror well the results displayed in Figures 2 and 3, showing a better performance than the baseline. Still, in terms of of word boundaries, the systems employing boundary information are penalized by a low recall and have lower F-scores, even if the word boundaries found are much more accurate.

	Term			Word Boundary		
	P	R	F	P	R	F
Baseline	45.3	38.5	38.8	22.8	2.8	4.7
Pause	47.1	40.9	40.9	24.1	2.5	4.2
Break 4	51.3	39.2	42.2	24.1	2.1	3.6
Break 3&4	**52.6**	35.8	40.4	24.4	1.7	3.0

Table 1: *Average precision, recall and F-score for discovered terms and word boundaries respectively, when the discard method is used.*

4.2. Experiment 2

In the previous experiment, we have observed that the method used for incorporating boundary information tends to persistently decrease the recall rate. This was due to the fact that the approach used consisted in discarding all terms found straddling a boundary. In this last experiment we wanted to compare this method to a different one which does not discard the terms spanning over several prosodic units, but shortens them ($truncate$). In this method, a term straddling a boundary would be truncated so it would include only the speech signal belonging to the prosodic unit having the highest overlap in time with the found term. The minimum term length constraint is applied after the truncation procedure.

	Term			Word Boundary		
	P	R	F	P	R	F
Baseline	45.2	42.4	40.3	23.0	2.9	4.9
Pause	46.5	**43.1**	41.6	26.3	**3.2**	**5.2**
Break 4	49.1	41.6	**42.4**	27.5	3.1	**5.2**
Break 3&4	51.3	38.5	41.9	**28.3**	3.0	5.1

Table 2: *Average precision, recall and F-score for discovered terms and word boundaries respectively, when the $truncate$ method is used.*

Table 2 illustrates the results obtained with this approach of incorporating prosodic information for the various types of information added. It contains the same type of information as Table 1. As we expected, the approach which truncates terms straddling a boundary gives a better recall than $discard$, at the expense of a slightly lower precision. In terms of F-score, an overall increase in performance is observed for all conditions.

The results in Table 2 show a small advantage in terms of word boundary F-score when boundary information is used. We detailed these results in Figure 4, by plotting the F-score over the range of ϵ_{DTW} values tested. It seems that for lower values of the DTW threshold, the baseline performs slightly better than the other systems, but, as the system becomes more permissive, the boundary information becomes more useful for discriminating words. Thus, it encourages the use of prosodic breaks information in conjunction with higher DTW thresholds for improved performance of UTD systems.

Interestingly, we found that the system using intonational phrase boundaries generally outperforms the system using smaller breaks (intermediate boundaries), which correspond to phonological phrase boundaries within the prosodic hierarchy [20]. Smaller breaks give better precision but this, in turn, is compensated by a worse recall which yields a slightly lower F-score, also for the $truncate$ case. This stands in contrast to findings in psycholinguistics studies where smaller breaks do seem to play a role, both for processing online speech in adults and for boosting speech segmentation in babies [11, 8, 9, 10]. We speculate that this may be due to the fact that our system imposes a 500 ms limit on the size of the discovered terms, which could affect proportionally more the breaks 3 and 4 compared to breaks 4. In order to prove this, we looked at the length of fragments delimited by boundaries. We discovered that when level 4 breaks are considered, 0.6% of the total speech fragments are shorter than 500 ms, the minimum term length employed in this study. This proportion increases to 3.7% when both level 3 and level 4 boundaries are taken into account, but it is equal to 0% for pauses. It means that for a percentage of the corpus no terms can be found. In this case, it would be interesting to investigate whether a lower limit for minimum term length will boost the recall rate, while not affecting too much the precision.

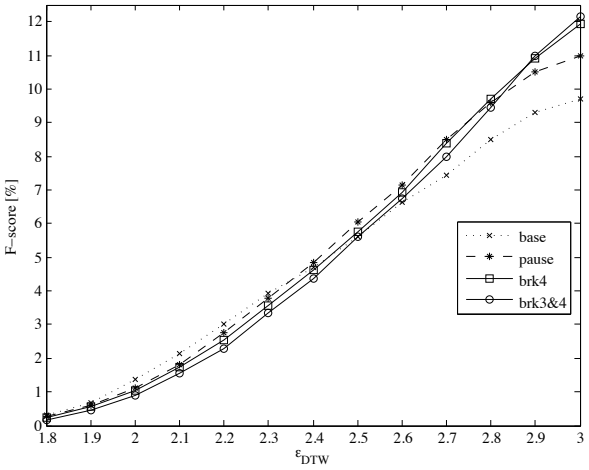

Figure 4: *Word boundary F-score obtained for different values of ϵ_{DTW} and various type of information used: baseline, pause information, level 4 boundaries, and level 3 and 4 prosodic boundaries.*

5. Conclusions

We have presented in this paper a preliminary study regarding the usefulness of prosodic boundaries for unsupervised term detection. Our findings show that boundary information, either intonational boundaries or intermediate and intonational boundaries, increases the performance of the system. The better results obtained are mainly due to increasingly accurate found terms, reflected in an improved precision. We have also compared prosodic boundaries against pauses, a prosodic cue which is generally easier to detect automatically. The system employing pauses outperformed the one using both level 3 and 4 boundaries, but it behaved worse than the system having knowledge of level 4 boundaries. We have discovered this advantage of pauses over level 3 and 4 boundaries to be due to a much lower recall, caused in part by constraints imposed for the length of the terms found.

The results we obtained encourage us to further continue our investigation by planning to use in a future study automatically extracted prosodic boundaries. In order to achieve this, we are currently focusing on methods of prosodic boundary detection based exclusively on acoustic cues. A second direction to pursue would be extending the study to several other languages. While unsupervised term discovery was applied until now only to less than a handful of languages we would expect prosodic boundary information to bring a consistent improvement in any language.

6. Acknowledgements

The research leading to these results was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG) and the Fondation de France. It was also supported by ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC. The authors would like to thank Armando Muscariello for providing them with the evaluation code.

7. References

[1] J. Saffran, R. Aslin, and E. Newport, "Statistical learning by 8-month old infants," *Science*, vol. 274, pp. 1926–1928, 1996.

[2] A. Park and R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[3] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Proc. of INTERSPEECH 2010*, 2010, pp. 1676–1679.

[4] R. Flamary, X. Anguera, and N. Oliver, "Spoken WordCloud: Clustering recurrent patterns in speech," in *Proc. of Int. Workshop on Content-Based Multimedia Index*, 2011, pp. 133–138.

[5] A. Muscariello, G. Gravier, and F. Bimbot, "Unsupervised motif acquisition in speech via seeded discovery and template matching combination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2031–2044, 2012.

[6] A. Christophe, E. Dupoux, J. Bertoncini, and J. Mehler, "Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition," *Journal of the Acoustical Society of America*, vol. 95, pp. 1570–1580, 1994.

[7] A. Christophe, J. Mehler, and N. Sebastián-Gallés, "Perception of prosodic boundary correlates by newborn infants," *Infancy*, vol. 2, no. 3, pp. 385–394, 2001.

[8] P. Jusczyk, D. Kemler-Nelson, K. Hirsh-Pasek, L. Kennedy, A. Woodward, and J. Piwoz, "Perception of acoustic correlates of major phrasal units by young infants," *Cognitive Psychology*, vol. 24, no. 2, pp. 252–293, 1992.

[9] L. Gerken, P. Jusczyk, and D. Mandel, "When prosody fails to cue syntactic structure: 9-month-olds' sensitivity to phonological versus syntactic phrases," *Cognition*, vol. 51, no. 3, pp. 237–265, 1994.

[10] A. Gout, A. Christophe, and J. Morgan, "Phonological phrase boundaries constrain lexical access II. Infant data," *Journal of Memory and Language*, vol. 51, no. 4, pp. 548–567, 2004.

[11] A. Christophe, S. Peperkamp, C. Pallier, E. Block, and J. Mehler, "Phonological phrase boundaries constrain lexical access: I. Adult data," *Journal of Memory and Language*, vol. 51, no. 4, pp. 523–547, 2004.

[12] A. Muscariello, G. Gravier, and F. Bimbot, "Zero-resource audio-only spoken term detection based on a combination of template matching techniques," in *Proc. of INTERSPEECH 2011*, 2011, pp. 921–924.

[13] I. Malioutov, A. Park, R. Barzilay, and J. Glass, "Making sense of sound: Unsupervised topic segmentation over acoustic input," in *Proc. of ACL 2007*, 2007, pp. 504–511.

[14] M. Dredze, A. Jansen, G. Coppersmith, and K. Church, "NLP on spoken documents without ASR," in *Proc. of EMNLP 2010*, 2010, pp. 460–470.

[15] L. Catanese, N. Souviraà-Labastie, B. Qu, S. Campion, G. Gravier, E. Vincent, and F. Bimbot, "MODIS: an audio motif discovery software," in *Proc. of INTERSPEECH 2013*, 2013, Software available online at `https://gforge.inria.fr/projects/motifdiscovery/`.

[16] A. Muscariello, G. Gravier, and F. Bimbot, "Variability tolerant audio motif discovery," in *Proc. of Int. Multimedia Modeling Conf. on Advances in Multimedia Modeling*, 2009.

[17] R. Daland and J. Pierrehumbert, "Learning diphone-based segmentation," *Cognitive Science*, vol. 35, no. 1, pp. 119–155, 2011.

[18] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[19] 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. of ICSLP 1992*, 1992, pp. 867–870.

[20] M. Nespor and I. Vogel, "Prosodic structure above the word," in *Prosody: Models and measurements*, pp. 123–140. Springer, 1983.