INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Wav2ToBI: a new approach to automatic ToBI transcription

Wanyue Zhai[1], *Mark Hasegawa-Johnson*[2]

[1]Stanford University, USA
[2]University of Illinois Urbana-Champaign, USA

wzhai702@stanford.edu, jhasegaw@illinois.edu

Abstract

ToBI [1] is a prosody labeling system that transcribes American English prosody in terms of phonological tones and break indices. Previous works on automatic ToBI transcription require additional information such as word boundaries and use modular feature extraction with separately optimized feature detectors and classifiers [2]. We are interested in investigating if a neural network-based approach would also result in high performance on automatic ToBI transcription without additional information. In this paper, we investigate the problem of pitch accent detection and prosody boundary detection using the Wav2vec 2.0 model [3] with only acoustic information. Our model is trained on the Boston University Radio News Corpus and evaluated on both the Boston University Radio News Corpus and the Boston Directions Corpus. We show that it achieves an F1 score of 0.82 on pitch accent detection and 0.86 on phrase boundary detection. Code and model weights are available. [1]

Index Terms: Prosodic boundaries, Wav2vec2, ToBI-label generation

1. Introduction

The ToBI annotation standard [1] is a prosody labeling system that transcribes Standard American English prosody in terms of phonological tones and break indices. It has been used widely in research studying the prosodic correlates of syntax [4, 5], semantics [6], information structure [7], dialog struture [8], and segmental acoustics [9]. Systems related to ToBI have been designed to code the prosody of languages including Japanese [10], Korean [11], Castilian Spanish [12], Portuguese [13], and Catalan [14]. Recognition or generation of ToBI labels has been demonstrated to be useful for detecting automatic speech recognition errors [15], for reducing speech recognition error rates [16], and for improving the quality of text-to-speech [17].

There are two main tiers of labels in the ToBI standard: the break index tier and the tone tier. The break index tier describes the grouping of words by measuring the strength of association between each pair of consecutive words on a scale of 0 (strongly conjoining) to 4 (the most disjoint). Two levels of prosodic boundaries mark the boundaries of prosodic phrases: intermediate phrase boundary (label 3) and intonational phrase boundary (label 4).

The tone tier transcribes the intonation pattern of the utterance. The transcriptions are associated with the occurrence of two types of pitch events: the above-described phrase boundaries and accented syllables (pitch accents). Tones are phonologically distinct sequences of high (H), low (L), and down-stepped (!H) pitch excursions. An example of complete ToBI annotation from Columbia Games Corpus [18] is shown in Figure 1.

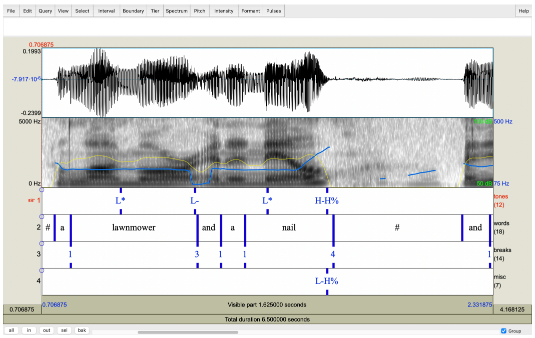


Figure 1: *Example ToBI annotation*

Since ToBI transcription requires a great deal of annotator effort, automatic ToBI transcription is considered an important field of research. Prosodic event detection is a subtask of automatic ToBI transcription, which is assigning a prosodic event (pitch accent or intonational phrase boundary) to the corresponding position of the speech. As we can see from the ToBI annotation standard, it is also an essential task for automatic ToBI transcription.

Previous works on automatic ToBI transcription often require additional information besides the acoustic signal, including, most frequently, information about word boundary times [2, 19]. Word boundary times are used to provide auxiliary information to a neural network [19], or to facilitate the use of linguistically-oriented hand-crafted feature extraction with separately optimized feature detectors and classifiers [2]. Such extra information and hand-crafted features may limit the application of ToBI to speaking styles with limited transcription resources. Therefore, we are interested in investigating if a neural network-based approach that does not rely on word boundary times would also result in high performance on automatic ToBI transcription. We investigate the problem of pitch accent detection and prosody boundary detection without word boundary times by using the Wav2vec 2.0 model [3] with only acoustic information. Our model is trained on the Boston University Radio News Corpus and evaluated on both the Boston University Radio News Corpus and the Boston Directions Corpus. We show that it achieves an F1 score of 0.82 for pitch accent detection and 0.86 for intonational phrase boundary detection.

2. Related work

The first publicly available automatic ToBI detection and classification system, AuToBI, was introduced by Rosenberg[2].

[1] https://github.com/reginazhai/Wav2ToBI.git

10.21437/Interspeech.2023-477

It calculates normalized pitch and intensity to train separate prosodic event detection and classification systems using logistic regression and SVMs. With the help of multiple alignment tools, Dominguez et al. [20] introduced PyToBI, which is a Toolkit for ToBI Labeling under Python. In the light of growth for neural networks, more studies on prosodic event detection using deep learning were introduced. Stehwien et al. [19] use acoustic features (signal frame energy, Mel spectrum, F0, voicing probability etc.) along with temporal features (position indicators) to perform pitch accent detection and intonational phrase boundary detection. They use a convolutional neural network (CNN) to learn these features. The model was trained and evaluated on two English corpora and one German corpus, in both same-corpus and cross-corpus evaluation paradigms. All of these systems use word boundary time as an auxiliary input to the prosodic event detector.

On the other hand, Vetter et al. [21] use Mel-frequency cepstral coefficient (MFCC) features and bidirectional long short-term memory neural network (BLSTM) for prosodic boundary detection and classification in a cross-lingual setting. The advantage of their model is that they only use unannotated speech at test time, without word boundary times. Kunešová et al. [22] also avoid the use of additional transciption by taking advantage of the rich representation learned from the Wav2vec 2.0 model to detect intonational phrase boundaries in Czech speech. This paper uses a similar approach as Kunešová et al., but focuses on English speech data and goes beyond the intonational phrase boundary detection by also detecting intermediate phrase boundaries and pitch accents.

3. Methodology

Our method use the Wav2vec 2.0 framework [3] to extract features from raw speech data. Wav2vec 2.0 is a self-supervised approach for automatic speech recognition (ASR). Because of the great amount of pretraining that it receives, it is able to perform well after a small amount of finetuning. Pretraining results in a relatively generic representation of the speech input, which can be used as features for tasks other than ASR.

Tones and break indices are often signaled by F0 (fundamental frequency), which may not be well represented in the output layer of Wav2vec 2.0. We therefore append an F0 measure (computed using Parselmouth [23], a python library for Praat [24] software) to the output layer of Wav2vec 2.0.

Detection of tones and break indices requires long-term dynamic context different from the context Wav2vec 2.0 is trained to extract. To fully learn the representation of each pitch accent and phrase boundary with their surrounding context, we therefore connect the output of Wav2vec 2.0 (with appended F0) to the input of a BLSTM. The entire network, including Wav2vec 2.0 and BLSTM, is then fine-tuned for the detection of prosodic events.

4. Experimental setup

4.1. Data

We use two widely known corpora for automatic ToBI labeling and evaluate the result in both within-corpus and cross-corpus evaluation to test for the robustness of the model when encountering an unfamiliar context.

The Boston University Radio News Corpus (BURNC) [25] consists of news professionally read on a public radio station by 7 announcers, and recordings by the announcers in the laboratory setting.

The Boston Directions Corpus (BDC) [26] consists of direction instructions spoken by 4 native Standard American English speakers. The directions are first recorded as spontaneous responses to requests for directions, then it was transcribed and later re-read by the same speakers in the lab. The read portion contains approximately 50 minutes of speech and 10818 words; the spontaneous portion contains approximately 60 minutes of speech and 11627 words.

Our systems are trained using BURNC, and tested on a read subset of BDC, in order to ensure generalizability of results. Rosenberg trained AuToBI [2] using BDC, and tested it using the Columbia Games Corpus [18]. Stehwien et al. [19] tested a number of within-corpus and cross-corpus evaluation paradigms; they found that BDC was harder to transcribe correctly than BURNC (accuracies were lower by 10% absolute, on average), and that cross-corpus evaluation was harder than same-corpus evaluation (by about 6% absolute). Our models are trained using the BURNC corpus (both lab news and radio news). Given the limited amount of data, the BURNC corpus was split into train and test set with a 4:1 ratio. Models were trained for 10, 20 and 30 epochs, then tested using the BURNC test set and three out of the four speakers from the read portion of BDC.

4.2. Network architecture

Wav2vec 2.0 is finetuned using clips of 20s with a step of 10s. The finetuned Wav2vec 2.0 model outputs a frame for every 20 ms. To align with the representation extracted from the Wav2vec 2.0 model, we use a frame size of 20 ms for the extraction of F0, then feed the feature matrix into the BLSTM for final detection. Since we have a relatively small training corpus, we use a small BLSTM with 128 or 256 hidden states.

Neural nets perform badly when trained with imbalanced data. In the case of speech, the majority of the corpus will not contain an indication of a prosodic event. To solve this, we compare two strategies. First, we take a similar approach as Kunešová et al. [22]: the reference labels during finetuning were provided using a fuzzy labeling function. During an interval ($\pm$ 0.16 s for pitch accent detection and $\pm$ 0.2 s for phrase boundary detection), the neural net target function linearly increases to reach 1.0 at the time of the corresponding prosodic event, then decreases to 0.0 by the end of the interval. The network is trained to reproduce this fuzzy target function with minimum mean squared error.

The pitch accent target function encodes the locations of any of the eight categories of pitch accents labeled in BURNC. The eight pitch accents are not equally frequent. H* accents are far more common than any other type, hence our system learns to detect H* accents better than any other type. We have tried training separate detectors for each accent type, and for various combinations of the under-represented accent types, but these alternatives cause detection accuracy of the under-represented accent types to decrease. Apparently BURNC does not contain enough examples of under-represented accent types to train effective separate neural net detectors, therefore we report results in which all accent types are pooled for both training and test.

Intonational and intermediate phrase boundaries have different acoustic correlates, but intermediate phrase boundaries provide information that can be helpful in intonational phrase boundary detection. Therefore, we also provide fuzzy labeling for labels with intermediate phrase boundary, but with a lower peak point (1.0 for intonational phrase boundary and 0.5 for in-

Table 1: *Pitch accent detection performance of systems trained and tested on BURNC data. CNN is the system of Stehwien et al. [19], which includes word boundary times as an auxiliary input, and scores correct detection if the detected accent is in the same word as the ground truth. NoFuzzy is our system without fuzzy targets; Wav2ToBI is our system with fuzzy targets.*

Model	Tolerance	Precision	Recall	F-1 Score
CNN [19]	word	0.86	0.90	0.88
NoFuzzy	0 ms	0.13	0.11	0.12
	40 ms	0.70	0.61	0.65
	80 ms	0.87	0.74	0.79
	100 ms	0.89	0.76	0.81
Wav2ToBI	0 ms	0.23	0.21	0.22
	40 ms	0.74	0.65	0.68
	80 ms	0.88	0.78	0.81
	100 ms	0.89	0.78	0.82

termediate phrase boundary). Targets for intermediate phrase boundaries also linearly increase to the peak point then decreases to 0.0.

Our second training strategy is also similar to the strategy of [22], but without fuzzy labeling. Consequently, during the interval (± 0.16 s for pitch accent detection and ± 0.2 s for phrase boundary detection), all labels for intonational phrase boundaries and pitch accents are marked as 1.0, and intermediate phrase boundaries as 0.5.

4.3. Post-processing

A proper ToBI label is a label paired with the timestamp, so we convert the frame-wise detection result to the correct format. For the fuzzy labeling approach, we look for the peaks from the output. A peak is defined as the highest point during a 100ms window. If the peak is higher than the threshold (0.75 when tested for only intonational phrase boundary; 0.8 and 0.4 when tested simultaneously for intonational and intermediate phrase boundary, respectively), we output the calculated time for the highest point during the window. For the second labeling approach, we look for short flat areas for the correct label. We keep track of intervals that starts from an increase in magnitude larger than 0.2, and ends with a decrease in magnitude larger than 0.2. If such an interval is longer than 100ms in duration, the middle time of the interval is counted as a detection.

Detected prosodic events are not perfectly synchronous with ground truth prosodic events. For this reason, if the timestamp predicted is within 100ms (five frames) of the correct label, we will count it as a correct output. We also report the scores for smaller tolerances, where a tolerance of “0ms” means that the detected and ground truth events occurred in the same 20ms frame.

5. Results

We evaluate each of our prediction results on the test set from BURNC corpus and the test set from BDC corpus using precision, recall and F-1 score.

5.1. Pitch accent detection

The results for pitch accent detection are shown in Tables 1 and 2. The model was trained with 128 hidden states with 30 epochs. As shown in Table 1, tolerance heavily affects the scores. The F-1 score for pitch accent detection can be as high as 0.82 when the tolerance is 100ms. Comparing the NoFuzzy and Wav2ToBI settings, we can see that the results for pitch detection with fuzzy labeling are typically better than those without fuzzy labeling by 0.02 absolute when tolerance is low.

Table 2: *Pitch accent detection (PA) and phrase boundary detection (PB) F-1 score of systems trained on the BURNC corpus and tested on BDC corpus. CNN is the system of Stehwien et al. [19]. NoFuzzy is our system without fuzzy targets; Wav2ToBI is our system with fuzzy targets.*

Model	Tolerance	PA	PB
CNN [19]	word	0.71	0.53
NoFuzzy	100 ms	0.58	0.76
Wav2ToBI	100 ms	0.72	0.79

Table 3: *Intonational phrase boundary detection performance of systems trained and tested on BURNC data. CNN is the system of Stehwien et al. [19], which includes word boundary times as an auxiliary input, and scores correct detection if the detected accent is in the same word as the ground truth. NoFuzzy is our system without fuzzy targets; Wav2ToBI is our system with fuzzy targets.*

Model	Tolerance	Precision	Recall	F-1 Score
CNN [19]	word	0.86	0.73	0.79
NoFuzzy	0 ms	0.10	0.10	0.10
	40 ms	0.64	0.64	0.64
	80 ms	0.84	0.83	0.83
	100 ms	0.87	0.86	0.86
Wav2ToBI	0 ms	0.22	0.21	0.21
	40 ms	0.67	0.66	0.66
	80 ms	0.84	0.82	0.83
	100 ms	0.86	0.84	0.85

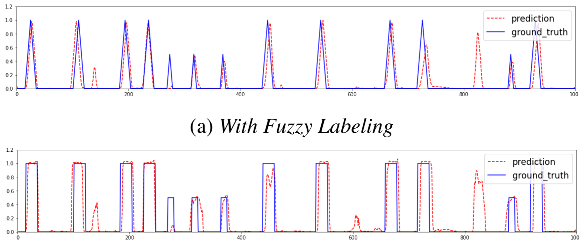
Although there were a lot of previous studies on automatic ToBI labeling systems and the subtasks related to it, most of them are performed with knowledge of word boundary times. Of these, the only result trained on BURNC and tested on both BURNC and BDC is that by [19], shown in Table 1. Their detections were considered correct if they occurred within the span of the same word as the ground-truth pitch accent; average word duration in BDC is 294ms. Though tested on the same corpora, our models and those of [19] were not tested on the same data. The models of [19] were trained and tested using only 5 of the 7 BURNC speakers in a 5-fold cross-validation, then tested cross-corpus on all BDC speech. Our models were trained and tested in a single 4:1 split of data from all 7 BURNC speakers, then tested cross-corpus on read speech from 3 BDC speakers.

5.2. Intonational phrase boundary detection

Intonational phrase boundary detection results are shown in Tables 2 and 3. Similar to pitch accent detection, intonational phrase boundary detection is also influenced by tolerance. There are few phrase boundaries that are detected with the exact time match, but an F-1 score of 0.86 is achievable with 100ms tolerance. Comparing the NoFuzzy and Wav2ToBI results, we

Table 4: *Comparison of Wav2ToBI results for the simultaneous detection of intermediate and intonational phrase boundaries (trained BURNC, tested BDC) with the cross-lingual results of Vetter et al. [21] (trained CSJ, tested BURNC)*

Model	**Break Type**	**Precision**	**Recall**	**F-1 Score**
Wav2ToBI	3	0.12	0.31	0.16
	4	0.53	0.75	0.62
Vetter et al [21]	3	0.63	0.01	0.01
	4	0.70	0.17	0.25



(a) *With Fuzzy Labeling*

(b) *Without Fuzzy Labeling*

Figure 2: *Example Output for Phrase Boundary Detection*

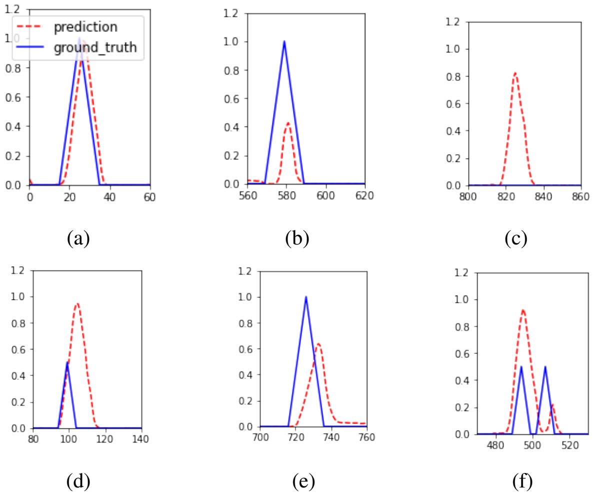
can see from Table 3 that fuzzy labeling yields F-1 improvements when tolerance is low but little improvement when tolerance is high. The improvement of Wav2ToBI relative to the CNN method of [19] is much larger for phrase boundaries than for pitch accents, possibly because Wav2vec 2.0 includes more information about long-term segmental phonetic context than do the features used by [19].

Most previous studies of ToBI phrase boundary detection use information about word boundary times. The only previously published system that does not take advantage of word boundary times, to our knowledge, is the cross-lingual study of Vetter et al [21], in which a system trained on the Corpus of Spoken Japanese (CSJ) is tested using BURNC. The cross-lingual system was tested for the detection of both intermediate (type 3) and intonational (type 4) phrase boundaries, and therefore, although cross-lingual evaluation is considerably harder than cross-corpus evaluation, their intermediate phrase boundary detection results can be used as an approximate baseline for the evaluation of ours. Table 4 shows that the cross-lingual system, when evaluated on BURNC with 80ms tolerance, has higher precision, lower recall, and lower F-1 than the Wav2ToBI system evaluated on BDC.

6. Discussion

We can visualize the types of information that the Wav2ToBI architecture is able to learn from a fuzzy target function. Fig. 2 shows the example output for phrase boundary detection. The blue line represents the ground truth label, while the the red line represents the predicted label. Fig 2a shows fuzzy labeling; we can observe the peaks for intermediate and intonational phrase boundaries. Fig 2b shows the target and predicted outputs without fuzzy labeling; we can see flat intervals of the peak value. Even when trained completely on the target without fuzzy labeling, the model still tends to output detection intervals with fuzzy onset and fuzzy offset.

We can also observe the different kinds of errors that the model makes when predicting the boundaries. Fig 3 shows the 6 main types of output when using fuzzy labeling on intonational phrase boundary detection. Fig. 3a shows the correct output, while the other five are different types of incorrect output. Observing from Fig 2a and Fig. 3a, we know that the correctly identified boundaries often overlap heavily with the ground truth results. Fig. 3b is a typical missed detection: the model outputs a peak, but the peak is below the detection threshold. Fig. 3c is a typical false alarm; the model output indicates that there is some evidence of a boundary in the acoustic signal, but human labelers did not hear this event as a prosodic phrase boundary. Fig 3d, Fig 3e and Fig 3f are examples when the model confuses intermediate phrase boundaries with intonational phrase boundaries. Fig 3f is a special case where two intermediate phrase boundaries are close to each other, so the model mistakes the two with an intonational phrase boundary. In this case, the higher *a priori* probability of intonational phrase boundaries, compared to the lower *a priori* probability of a rapid succession of intermediate phrase boundaries, may bias the model to detect the former in preference to the latter.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3: *Types of Output for Phrase Boundary Detection*

7. Conclusions and Future Work

In this paper, we explored a new approach towards ToBI labeling. We applied the Wav2vec 2.0 network using BLSTM networks for the task of prosodic event detection. We showed that Wav2vec 2.0 is able to provide a reasonably good representation of the speech corpus We also demonstrated that even without knowledge of word boundaries and word-level feature extraction, we can still achieve a high F-1 score for detecting pitch accents and intonational phrase boundaries that can be generalized to other corpora.

We have tried several methods to extend this work to the classification of pitch accents into the different ToBI pitch accent categories, and we've been forced to conclude that BURNC does not contain enough examples of the infrequent pitch accents to train a neural network detector. In the future we will explore hybrid methods, e.g., maximum-margin training criteria and other small-dataset methods, to see if such hybrid methods can give better performance for the detection of infrequent label categories in ToBI.

8. References

[1] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody." in *ICSLP*, vol. 2, 1992, pp. 867–870.

[2] A. Rosenberg, "Autobi-a tool for automatic tobi annotation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] N. Veilleux, "Computational models of the prosody/syntax mapping for spoken language systems," Ph.D. dissertation, Boston University, 1994.

[5] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. Am.*, vol. 90, no. 6, pp. 2956–2970, Dec. 1991.

[6] J. Hirschberg, "Intonational overload: Uses of the H* !H* L- L% contour in read and spontaneous speech," in *Ninth Conference on Laboratory Phonology (LapPhon)*, Urbana, 2004.

[7] M. Grice and M. Savino, "Can pitch accent type convey information status in yes-no questions?" in *Concept to Speech Generation Systems*, 1997.

[8] J. Hirschberg, D. Litman, and M. Swerts, "Identifying user corrections automatically in spoken dialogue systems," in *Internat. Conferen. Spoken Language Processing*, 2002.

[9] J. Cole, H. Choi, H. Kim, and M. Hasegawa-Johnson, "The effect of accent on the acoustic cues to stop voicing in radio news speech," in *Internat. Conferen. Phonetic Sciences*, 2003.

[10] J. J. Venditti, *Japanese ToBI labelling guidelines*. Columbus, Ohio: Ohio State University. Department of Linguistics, 1997.

[11] S.-A. Jun, "K-tobi (korean tobi) labelling conventions," *Speech Sciences*, vol. 7, no. 1, pp. 143–170, 2000.

[12] T. Face and P. Prieto, "Rising accents in Castilian Spanish: a revisino of Sp_ToBI," *Journal of Portuguese Linguistics*, vol. 6, no. 1, pp. 117–146, 2007.

[13] S. Frota, P. Oliveira, and M. Cruz, *P-ToBI: tools for the transcription of Portuguese prosody*. Lisboa: Laboratóorio de Fonética, 2015.

[14] P. Prieto, L. Aguilar, I. Mascaró, F. Torres-Tamarit, and M. Vanrell, "L'etiquetatge prosodìc Cat_ToBI," *Estudios de fonética éxperimental*, vol. XVIII, pp. 287–309, 2009.

[15] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic cues to recognition errors," in *Eurospeech*, 2001.

[16] K. Chen and M. Hasegawa-Johnson, "How prosody improves word recognition," in *Proceedings of Speech Prosody*, Nara, Japan, 2004, pp. 583–586.

[17] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth annual conference of the international speech communication association*, 2014.

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

[19] S. Stehwien, A. Schweitzer, and N. T. Vu, "Acoustic and temporal representations in convolutional neural network models of prosodic events," *Speech Communication*, vol. 125, pp. 128–141, 2020.

[20] M. Domínguez Bajo, P. L. Rohrer *et al.*, "Pytobi: a toolkit for tobi labeling under python," *Interspeech 2019; 2019 Sept 15-19; Graz, Austria. Baixas: ISCA; 2019. p. 3675-6.*, 2019.

[21] M. Vetter, S. Sakti, and S. Nakamura, "Cross-lingual speech-based tobi label generation using bidirectional lstm," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6620–6624.

[22] M. Kunešová and M. Řezáčková, "Detection of prosodic boundaries in speech using wav2vec 2.0," in *International Conference on Text, Speech, and Dialogue (TSD 2022)*. Springer, 2022, pp. 377–388.

[23] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/, 2021.

[25] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[26] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *Working notes of the AAAI spring symposium on empirical methods in discourse interpretation and generation*. Citeseer, 1995, pp. 106–112.