Neural Speech Synthesis with Enriched Phrase Boundaries

Marie Kunešová, Jindřich Matoušek

New Technologies for the Information Society and Department of Cybernetics,
Faculty of Applied Sciences, University of West Bohemia, Pilsen, Czech Republic
{mkunes,jmatouse}@ntis.zcu.cz

Abstract

Prosodic phrasing is one of the factors influencing the naturalness of synthesized speech. In this paper, we enrich the phonetic representation for neural speech synthesis with additional markers denoting the strength of phrase breaks between words. These markers are assigned to the training data automatically, using our previously introduced model for audio-based phrase boundary detection. We tested the approach with two different levels of resolution for the break indices – either ten distinct levels (P10) or only "ToBI-like" four levels (P4). Listening tests with two different speaker voices show a statistically significant preference among listeners for P10 or P4 over the baseline speech synthesis without these markers (P0), although which version is judged as better depends on the voice.

Index Terms: speech synthesis, phrasing, phrase breaks, wav2vec

1. Introduction

Text-to-speech (TTS) synthesis aims at generating human-like speech from input text. Modern *neural speech synthesis* models have dramatically improved the quality of synthetic speech. Many sophisticated architectures have been proposed, gradually replacing CNN/RNN-based models (e.g. [1, 2, 3]) with Transformer-based models [4, 5, 6], original autoregressive models [3, 4] with more powerful generative models (VAE-, GAN-, flow-, and diffusion-based ones [7, 8, 9]), and two-stage acoustic models/vocoders [3, 5] with fully end-to-end models [7, 10].

In addition to models' architectures, the representation of the input text is also an important factor. In the true end-to-end approach, raw text (i.e., letters or *graphemes*) is used as the input, mapping the input graphemes directly to their acoustic counterparts. However, since graphemes generally do not represent pronunciation, they need not correspond to the acoustic representation of the synthesized speech closely [11]. On the other hand, the phonetic representation of the input text in the form of a sequence of *phonemes* (or phones) is often used as it has a more direct relationship to the acoustic signal than graphemes, and thus, it approximates speech more closely than the graphemes [11, 12].

The input representation (either in the form of a text or phonetic labels) inherently contains important information about the *prosody* of the synthesized speech. Since prosody (i.e., intonation, speech tempo, phrasing, etc.) to a large extent influences the naturalness of the synthesized speech, speech synthesis models have to extract prosody-related information from the input, interpret it correctly, and ensure that the output speech contains appropriate prosodic characteristics. This is also the case with *prosodic phrasing* – appropriate phrasing (placing *phrase breaks* of suitable strength and length in the right places) influences the intonation and tempo of output speech and greatly increases the naturalness and also intelligibility of TTS systems [13].

For text-to-speech, phrase breaks are typically predicted from input textual/phonetic representation. Many different approaches have been studied during the last decades, including deterministic approaches based on punctuation marks, classification-based approaches with different sets of features, HMMs, and neural networks [14, 15, 16, 13, 17, 18]. Some of them aimed at automatic annotation approximating the well-known ToBI scheme [19, 20, 21, 22]. Other approaches detect phrase boundaries in the speech signal and are often used to provide speech corpora with audio annotation [23, 24, 25, 26]. Textual and acoustic information could also be combined [27, 28, 29].

In this paper, we approximate the phonetic representation inputting to a speech synthesis system even more closely to its corresponding speech by enriching it with additional markers denoting the strength of phrase boundaries. The markers are assigned to the training data automatically, using our previously introduced model for audio-based phrase boundary detection [26]. Two different phrase boundary strength scales were investigated – either ten distinct levels (P10) or only "ToBI-like" four levels (P4)[1]. Finally, we examine whether the enriched representation improves the naturalness of speech synthesized by a neural speech synthesis model.

The paper is organized as follows. In Section 2, we describe the phrase detector used to label phrase boundaries and the training data used to train it. The exact process of creating the labels from the model's predictions is explained in Section 3, where we also present the speech synthesis experiment and describe speech datasets used to train a speech synthesizer. Results and discussions are presented in Section 4. Finally, conclusions and future work are drawn in Section 5.

2. Phrase boundary detection

Our approach to labeling the strength of the phrase boundaries was inspired by an observation in our previous work [26], which focused specifically on the detection of *prosodic phrase* boundaries in the speech signal.

In the paper, we observed that even when our model was trained only to detect full prosodic boundaries (ToBI break index 4), it also predicted relatively high scores for *intermediate* boundaries (ToBI break index 3). In other words, the model seemed to detect some sort of acoustic feature that is common to both types, only pronounced to a different degree.

In our current work, we decided to extrapolate that idea to a larger scale, by reinterpreting the predictions of the model as a

[1] The P4 scheme resembles the ToBI scheme [19] with break indices 0 and 1 combined into a single category.

10.21437/Interspeech.2023-1552

measure of the strength of potential phrase boundaries between words. For this, we have taken our approach[2] from [26] and slightly adapted it for our current work.

Most of the details of the model and its training are kept identical to the original paper, but we briefly describe them here for convenience. The creation of the final labels from the model's outputs will be explained in Section 3.2.

2.1. Training data

The training data for the model were the same as in the original paper [26]: a set of recordings of Czech radio news bulletins, referred to as the News-Reading Speech (NRS) corpus. There are 12 recordings, each spoken by a different speaker, with a total length of 42 minutes (486 sentences, 6371 words). The data are annotated by phonetic experts [30] following the ToBI labeling guidelines [31], although only ToBI break indices 3 (intermediate phrase break) and 4 (full prosodic phrase break) are included. Of these, ToBI break indices 4 represent 22% of all breaks between words (including the ends of sentences), while 7% are marked as break index 3. The remaining 71% have neither label.

2.2. Model for phrase boundary detection

The model for phrase boundary detection is based on wav2vec 2.0 [32]. Wav2vec 2.0 (or "wav2vec2") is a self-supervised framework for speech representation which has recently gained popularity in a wide range of speech processing tasks [33, 34].

As in [26], we use the pre-trained wav2vec 2.0 base model "ClTRUS"[3] [35] and fine-tune it to predict a score from 0–1 for each audio frame (i.e. every 20 ms of speech, as usual for wav2vec2), indicating the presence or absence of a prosodic boundary in each frame. The input of the model is 16 kHz audio signal, zero-padded or split into chunks of 20 s.

We fine-tuned the model on the entirety of the NRS data (all 12 speakers), with 10 epochs. Unlike the original paper, which averaged outputs from multiple initializations for better consistency of results, we only used a single initialization. Otherwise, all settings were identical to [26].

The fine-tuning process uses a fuzzy labeling function, where ground-truth prosodic phrase boundaries (ToBI break index 4) are given the value 1, which linearly decreases to zero over an interval of 0.2 s on each side. Intermediate phrase boundaries (ToBI break index 3) are labeled in the same way, but with a maximum value of 0.5 (an example of this labeling can be seen in Figure 1). Consequently, the phrase boundary detection is done as a regression task, rather than a simple yes/no classification.

Figure 2 shows a histogram of predicted labels on the NRS data, with a separate set of models, using 12-fold cross-validation and 5 training epochs (to avoid evaluating on training data).

In [26], we used a decision threshold and found prosodic boundaries as peaks in the output above a certain height. However, here we can instead reinterpret the raw predicted scores as a continuous scale that measures the strength of the phrase boundaries between words. The process of creating these labels will be described in Section 3.2.

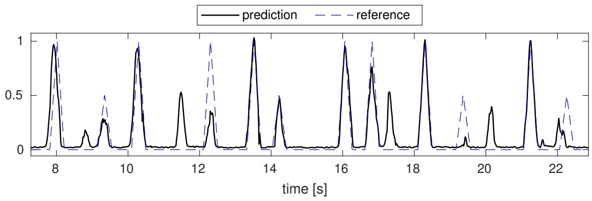

Figure 1: *Example of the reference labels and predictions of the phrase boundary detection model on NRS data.*

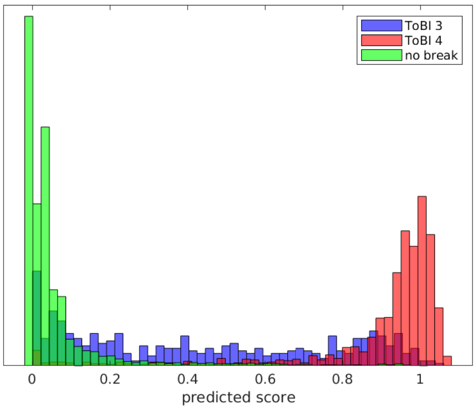

Figure 2: *Histogram of predicted scores on the NRS data, categorized by ground-truth labels and normalized as probability distributions (independently of each other).*

3. Speech synthesis experiment

In this section, we describe a speech synthesis experiment in which we use the phrase detector introduced in Section 2 to detect phrase boundaries in source speech datasets for speech synthesis and to enrich their phonetic representation accordingly. The aim of this experiment was to test whether the enriched phonetic representation can improve the quality of synthetic speech.

3.1. Speech datasets

To train speech synthesis models, we used two large corpora of Czech news-reading speech recorded by a professional male (**Speaker M**) and female (**Speaker F**) speaker. The corpora were primarily designed for the use with *unit-selection* speech synthesis [36], but Vít et al. [37] showed that the corpora are also suitable for neural speech synthesis. They contain paired text-audio data with approximately 14 hours of audio (including pauses) distributed over 12,240 (Speaker M) and 12,708 (Speaker F) utterances. For our purposes, the audio has been downsampled to 24 kHz, carefully annotated, and the resulting text has been normalized to expand out numbers, dates, ordinals, monetary amounts, etc. Finally, the text of each audio was transcribed into a sequence of phones using a set of carefully designed Czech phonetic rules and a pronunciation dictionary with words that do not obey Czech pronunciation rules [17]. Since Matoušek & Tihelka [12] showed that it is advantageous to explicitly include pauses and punctuation marks in the phonetic representation when training a synthesizer, each phonetic transcript was supplemented by pauses using an external speech

[2] Our code is available at `https://github.com/mkunes/w2v2_audioFrameClassification`

[3] **C**zech **l**anguage **TR**ransformer from **U**nlabeled **S**peech, available from `https://huggingface.co/fav-kky/wav2vec2-base-cs-80k-ClTRUS`

Table 1: *Mapping of the wav2vec2 model outputs to the phrase break indices.*

raw value	< 0.1	[0.1,0.2)	[0.2,0.3)	[0.3,0.4)	[0.4,0.5)	[0.5,0.6)	[0.6,0.7)	[0.7,0.8)	[0.8,0.9)	≥ 0.9
label - P10	0	1	2	3	4	5	6	7	8	9
label - P4	0	0	1	1	1	2	2	2	3	3

Table 2: *Example sentence from the listening test, as transcribed for the TTS models using International Phonetic Alphabet (IPA). The symbol # denotes pauses. English translation: "The elections were not interrupted; rather, due to high interest from voters, they were extended by two hours."*

Text	*Volby nebyly přerušeny, ale naopak kvůli vysokému zájmu voličů byly prodlouženy o dvě hodiny.*
P0	# vɔlbɪ nɛbɪlɪ pr̝ɛrʊʃɛnɪ, # alɛ naɔpak kvuːlɪ vɪsɔkɛːmʊ zaːjmʊ vɔlɪt͡ʃuː bɪlɪ prodlo͡ʊʒɛnɪ ɔ dvjɛ ɦɔɟɪnɪ. #
P4	# vɔlbɪ0 nɛbɪlɪ0 pr̝ɛrʊʃɛnɪ,3 # alɛ0 naɔpak2 kvuːlɪ0 vɪsɔkɛːmʊ0 zaːjmʊ0 vɔlɪt͡ʃuː3 bɪlɪ0 prodlo͡ʊʒɛnɪ2 ɔ0 dvjɛ0 ɦɔɟɪnɪ.3 #
P10	# vɔlbɪ1 nɛbɪlɪ0 pr̝ɛrʊʃɛnɪ,9 # alɛ0 naɔpak5 kvuːlɪ0 vɪsɔkɛːmʊ0 zaːjmʊ0 vɔlɪt͡ʃuː8 bɪlɪ0 prodlo͡ʊʒɛnɪ6 ɔ0 dvjɛ0 ɦɔɟɪnɪ.9 #

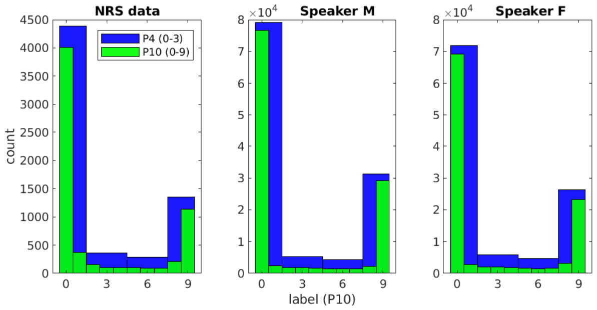

Figure 3: *Histogram of the P4 and P10 labels for the NRS data and for the training data of the two voices.*

segmentation tool [38].

3.2. Phrase boundary labels

To obtain labels indicating the strength of potential phrase boundaries in the speech corpora, we first used the model described in Section 2 to obtain frame-level predictions, as in Figure 1. For this, the signal was downsampled to 16 kHz, as required by wav2vec2. Then we mapped the predictions to individual words in the following manner:

First, we detected all peaks (i.e. all local maxima) in the output of the wav2vec2 model and assigned each peak to the nearest spoken word within 100 ms. This was based on the *end time* of the words, as given by the segmentation tool mentioned in Section 3.1. Then we labeled each word based on the highest peak assigned to it, using the mapping in Table 1 to convert the continuous values to digits 0–9 (P10) or 0–3 (P4), where 0 means "no break" and 9 (resp. 3) means "strongest break".

The reason why we use ten categories in one of the options is that the phrase breaks need to be represented by a single character in the transcripts used by the TTS system. Thus, digits 0–9 seem like a natural choice. However, they may be too many – so we also explore a second option with fewer labels.

The number of labels and the mapping for the second option P4 were selected based on the distribution of individual labels in the training data, especially in relation to the ground truth labeling in the NRS corpus – as illustrated in Figures 2 and 3.

Finally, in addition to the two phrasing schemes P4 and P10, we also have a baseline without any phrase markers, which we will refer to here as P0. Table 2 shows an example sentence, transcribed using all three options, P0, P4, and P10.

3.3. Speech synthesis model

To evaluate the effect of enriching the phonetic representation on the quality of synthetic speech, we trained VITS, a neural speech synthesis model using a conditional variational autoencoder with adversarial learning [7]. VITS could be viewed as a full *end-to-end model* in that it directly converts graphemes/phonemes into waveform. In our case, the input to the model was a sequence of phonemes (including pauses) supplemented by punctuation marks and phrase boundary labels as described in Section 3.2 and shown in Table 2.

In our experiments, VITS models were trained using the AdamW optimizer [39] with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and weight decay $\lambda = 0.01$. The learning rate decay was scheduled by a $0.999^{1/8}$ factor in every epoch with an initial learning rate of 2×10^{-4}. The batch size was set to 32 and the models were trained up to approximately 1.3M steps (720 hours) using mixed precision training on a single GeForce GTX 1080 Ti GPU using the Coqui-TTS framework[4].

4. Results and discussion

4.1. Listening test

Two *preference listening tests* (also known as AB tests), one for the male and one for the female synthetic voice, were conducted for a direct comparison of the investigated phenomena. Each listening test contained the same 16 sentences that were synthesized by each of the three synthesis models P0, P4, and P10 for the two voices.

To ensure accurate labeling of the phrase breaks, we chose the test sentences from other audio recordings we have available. These recordings were labeled in the same way as the training data for speech synthesis.

We selected 8 declarative sentences in a news-reading style and 8 questions from audiobooks, in such a way as to have a relatively balanced distribution of labels "1" to "8" (labels "0" and "9" are naturally much more common than others). The length of the sentences was 5-20 words.

For each voice, the listeners listened to two versions of the same sentence synthesized by different models (P0, P4, P10). In each test, 48 comparisons were made. 20 listeners completed the

[4] https://github.com/coqui-ai/TTS

	Preference: P0 vs. P10				Preference: P0 vs. P4				Preference: P10 vs. P4			
model	P0	same	P10	score	P0	same	P4	score	P10	same	P4	score
Speaker M - questions	29.4	27.5	**43.1**	0.14	37.5	21.3	**41.3**	0.04	35.6	**36.9**	27.5	-0.08
Speaker M - declarative	23.1	36.9	**40.0**	0.17	**39.4**	26.9	33.8	-0.06	**40.0**	38.1	21.9	-0.18
Speaker M - all	26.3	32.2	**41.6**	0.15	**38.4**	24.1	37.5	-0.01	**37.8**	37.5	24.7	-0.13
Speaker F - questions	29.4	27.9	**42.6**	0.13	25.7	28.7	**45.6**	0.20	28.7	27.9	**43.4**	0.15
Speaker F - declarative	**39.7**	35.3	25.0	-0.15	20.6	31.6	**47.8**	0.27	25.7	31.6	**42.6**	0.17
Speaker F - all	**34.6**	31.6	33.8	-0.01	23.2	30.1	**46.7**	0.24	27.2	29.8	**43.0**	0.16
Both - questions	29.4	27.7	**42.9**	0.13	37.4	21.3	**41.3**	0.12	35.6	**36.8**	27.6	0.03
Both - declarative	31.4	**36.1**	32.5	0.01	30.0	29.2	**40.8**	0.11	32.9	**34.9**	32.3	-0.01
Overall preference	30.1	31.9	**38.0**	0.08	31.4	26.9	**41.7**	0.10	32.9	**34.0**	33.1	0.00

Table 3: *Results of the preference listening tests, as preference [%] and as a score on a [-1,+1] scale with positive values preferring the proposed systems over baseline (and P4 over P10).*

first test (Speaker M) and 17 completed the second test (Speaker F). They were instructed to evaluate each pair of synthesized sentences on a three-point scale (better/same/worse) concerning the quality and naturalness of prosody and intonation in the synthesized speech. The synthetic sentences were presented in the same order to all listeners. All the listeners were native Czech speakers, some of whom had no prior experience with speech synthesis, and had no hearing problems.

To ensure that only the phenomena under examination will be evaluated, we initially synthesized a larger number of "candidate" sentences and filtered out those containing artifacts not related to the research question (around 10% of synthesized utterances from each model were discarded).

Examples of the sentences in the listening test, as well as some of those that were discarded due to unrelated artifacts, can be found on our demo page[5].

4.2. Discussion

Table 3 presents the overall results of the listening tests, as well as a more detailed breakdown separated by the type of sentence (declarative or question) and the speaker. We also analyzed the results of each listening test using the Wilcoxon signed-rank test, with the Holm-Bonferroni correction for multiple comparisons.

The results suggest that the enriched representation (either P4 or P10) outperforms the baseline (P0) but listener preferences depended on the voice. For the female voice, P4 was strongly preferred over the other two schemes. On the other hand, P10 was the preferred scheme for the male voice. All of these preferences are statistically significant at the 5% level.

Finding out why the phrasing schemes P4 and P10 behave differently for the voices studied (e.g., whether this can be due to differences between male and female voices and/or different ways of the prosodic style of speaking) remains our future work.

From the detailed breakdown of the results in Table 3, it also seems that questions were perceived differently than declarative sentences. The greatest difference is for Speaker F, in the comparison between P0 and P10, though the sample size is too small to make any definite conclusions.

The results of the listening tests show that enriching phonetic representation with phrase boundary labels can improve the resulting synthesized speech. However, the choice of a particular phrasing scheme has yet to be explored.

5. Conclusions

In the paper, we trained a phrase boundary detector from labeled audio data and applied it to detect phrase boundaries in two speech datasets (of a male and female voice) for speech synthesis. Phonetic representation enriched with the detected phrase boundary labels was then utilized to build a neural speech synthesis model. We tested the representation with two different levels of resolution for the phrase breaks – either ten (P10) or four (P4) distinct levels. In both cases, for both voices, listening tests show a statistically significant preference for either P10 or P4 over the baseline speech synthesis without the enriched representation (P0). The choice of the appropriate phrasing scheme for a particular voice remains our future work.

Although the model for phrase detection was trained using only 42 minutes of labeled data, it can achieve very good results [26]. In our future work, we plan to examine whether more labeled data will result in even more accurate phrase detection performance and, consequently, in further improvements in the quality of synthetic speech.

The aim of the experiment described in this paper was to test whether the enriched phrase boundaries have the potential to improve the quality of speech synthesized by a neural speech synthesis model. For this proof of concept, the phrase boundary labels were obtained from existing speech recordings. Since the enriched phrase boundaries indeed helped in improving the quality of synthetic speech, the next step will be to predict the phrase boundary labels solely from the textual and/or phonetic representation at the input of a TTS system. Some work towards this has already been done in [40].

6. Acknowledgements

This research was supported by the Czech Science Foundation (GA CR), project No. GA21-14758S. Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

7. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *Speech Synthesis Workshop*, Sunnyvale, USA, 2016.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgian-

[5] `https://artic-tts-experiments.github.io/demo_Interspeech2023`

nakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 4006–4010.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Calgary, Canada, 2018, pp. 4779–4783.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with Transformer network," in *AAAI Conference on Artificial Intelligence*, Honolulu, USA, 2019, pp. 6706–6713.

[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[6] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Toronto, Canada, 2021.

[7] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, 2021, pp. 5530–5540.

[8] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Brighton, United Kingdom, may 2019, pp. 3617–3621.

[9] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *International Conference on Machine Learning*, 2021.

[10] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-End Adversarial Text-to-Speech," in *International Conference on Learning Representations*, 2021.

[11] J. Fong, J. Taylor, K. Richmond, and S. King, "A comparison of letters and phones as input to sequence-to-sequence models for speech synthesis," in *Speech Synthesis Workshop*, Vienna, Austria, 2019, pp. 223–227.

[12] J. Matoušek and D. Tihelka, "On comparison of phonetic representations for Czech neural speech synthesis," in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence. Cham: Springer, 2022, vol. LNAI 13502, pp. 410–422.

[13] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase break prediction for long-form reading TTS: Exploiting text structure information," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1064–1068.

[14] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Computer Speech and Language*, vol. 12, no. 2, pp. 99–117, 1998.

[15] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *INTERSPEECH*, Singapore, 2014, pp. 2268–2272.

[16] A. Rosenberg, R. Fernandez, and B. Ramabhadran, "Modeling phrasing and prominence using deep recurrent learning," in *INTERSPEECH*, Dresden, Germany, 2015, pp. 3066–3077.

[17] M. Řezáčková, J. Švec, and D. Tihelka, "T5G2P: Using text-to-text transfer transformer for grapheme-to-phoneme conversion," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 6–10.

[18] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in Japanese text-to-speech synthesis," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 3126–3130.

[19] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: a standard for labelling English prosody," in *International Conference on Spoken Language Processing*, Banff, Canada, 1992, pp. 867–870.

[20] A. K. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech Communication*, vol. 33, no. 1-2, pp. 135–151, 2001.

[21] A. Rosenberg, "AuToBI - A tool for automatic ToBI annotation," in *INTERSPEECH*, Makuhari, Japan, 2010, pp. 146–149.

[22] Y. Zou, S. Liu, X. Yin, H. Lin, C. Wang, H. Zhang, and Z. Ma, "Fine-grained prosody modeling in neural speech synthesis using ToBI representation," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 3146–3150.

[23] A. Suni, J. Simko, and M. Vainio, "Boundary detection using continuous wavelet analysis," in *Speech Prosody*, 2016, pp. 267–271.

[24] B. Schuppler and B. Ludusan, "An analysis of prosodic boundary detection in German and Austrian German read speech," in *Speech Prosody*, 2020, pp. 990–994.

[25] B. Lin, L. Wang, X. Feng, and J. Zhang, "Joint detection of sentence stress and phrase boundary for prosody," in *INTERSPEECH*, Shanghai, China, 2020, pp. 4392–4396.

[26] M. Kunešová and M. Řezáčková, "Detection of prosodic boundaries in speech using wav2vec 2.0," in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence. Cham: Springer, 2022, vol. LNAI 13502, pp. 377–388.

[27] G. Christodoulides, M. Avanzi, and A. C. Simon, "Automatic labelling of prosodic prominence, phrasing and disfluencies in French speech by simulating the perception of naïve and expert listeners," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3936–3940.

[28] F. Gallwitz, H. Niemann, E. Nöth, and V. Warnke, "Integrated recognition of words and prosodic phrase boundaries," *Speech Communication*, vol. 36, no. 1, pp. 81–95, 2002.

[29] D. Kocharov, T. Kachkovskaia, and P. Skrelin, "Prosodic boundary detection using syntactic and acoustic information," *Computer Speech & Language*, vol. 53, pp. 231–241, 2019.

[30] J. Volín, "The size of prosodic phrases in native and foreign-accented read-out monologues," *Acta Universitatis Carolinae – Philologica*, no. 2, pp. 145–158, 2019.

[31] M. E. Beckman and G. Ayers Elam, *Guidelines for ToBI Labelling, version 3*, The Ohio State University Research Foundation, Ohio State University, 1997.

[32] 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.

[33] S.-w. Yang *et al.*, "SUPERB: Speech processing Universal PERformance Benchmark," in *INTERSPEECH*, Brno, Czechia, 2021, pp. 1194–1198.

[34] M. Kunešová and Z. Zajíc, "Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0," in *Proc. ICASSP 2023*, 2023.

[35] J. Lehečka, J. Švec, A. Pražák, and J. Psutka, "Exploring capabilities of monolingual audio transformers using large datasets in automatic speech recognition of Czech," in *INTERSPEECH*, Incheon, Korea, 2022, pp. 1831–1835.

[36] J. Matoušek, D. Tihelka, and J. Romportl, "Building of a speech corpus optimised for unit selection TTS synthesis," in *Language Resources and Evaluation Conference*, Marrakech, Morocco, 2008, pp. 1296–1299.

[37] J. Vít, Z. Hanzlíček, and J. Matoušek, "On the analysis of training data for Wavenet-based speech synthesis," in *IEEE International Conference on Acoustics Speech and Signal Processing*, Calgary, Canada, 2018, pp. 5684–5688.

[38] Z. Hanzlíček and J. Vít, "LSTM-based speech segmentation trained on different foreign languages," in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, P. Sojka, I. Kopeček, K. Pala, and A. Horák, Eds. Springer, 2020, vol. 12284, pp. 456–464.

[39] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, New Orleans, USA, 2019.

[40] M. Řezáčková and J. Matoušek, "Text-to-text transfer transformer phrasing model using enriched text input," in *Text, Speech and Dialogue*, ser. Lecture Notes in Artificial Intelligence. Cham: Springer, 2022, vol. LNAI 13502, pp. 389–400.