Can prosody embeddings from FastSpeech2 identify word and syllable prominence in native and non-native speech?: A preliminary study

Anonymous submission to Interspeech 2024

Abstract

The synthesized speech from TTS becomes natural by incorporating prosody. FastSpeech2 is one such state-of-the-art TTS. To understand the quality of prosody embeddings obtained from FastSpeech2, in this work, a study is conducted to identify word and syllable prominence in native and non-native speech. We consider the prominence related embeddings that include duration and energy, which were proposed to generate prominence at different word and syllable locations for a given input. Using these, the following is analysed: can text-only suffice to incorporate prominence at different locations or both speech plus text is needed? The embeddings for the first case are obtained during inference setup and the latter during training setup. The outcome shows that the embeddings can identify the prominence in native and non-native speech when speech plus text is considered. However, when text-only is considered, the embeddings are not effective in non-native compared to native speech.

Index Terms: FastSpeech2, prosody embeddings, prominence, native speech, non-native speech

1. Introduction

Speech serves as a multi-layered signal, carrying not only the essence of lexical and grammatical content (the "what") but also the nuances of prosody (the "how") and distinctive features tied to individual speakers (such as identity and emotional nuances). Prosody, which extends beyond individual phonetic units, covers elements related to syllables, words, phrases, sentences, and more extended utterances, collectively known as supra-segmental features [1, 2]. Prominence (also referred to as stress), within this context, manifests as the heightened emphasis on the syllable or words [3]. The acoustic correlation of the emphasis can be observed through the manifestation of duration and energy.

Native speakers acquire their language naturally from early childhood whereas non-native speakers, often influenced by their native language phonetics, may introduce distinctive stress patterns when speaking a new language. Investigating stress across different linguistic levels is pivotal for unraveling its intricate implications, drawing substantial attention in the realm of research. A multitude of studies has scrutinized stress, probing into its manifestations at word level[4, 5, 6] as well as syllable level [7, 8, 9, 10]. An unsupervised approach for automatic word prominence detection is presented in [4]. The algorithm scores prominence by combining various acoustic feature sets, incorporating both spectral and temporal features extracted from the correlation envelope of the word signal along with part of speech (POS) as a linguistic correlate for word prominence. Sonority, which represents the carrying power of sounds in words or longer utterances [11], plays a key role in the method proposed by [9] for automatic syllable stress detection. Their method computes a feature contour is computed by combining sonority-motivated cues with sub-band short-time energy contours.

Recent advancements in Text-to-Speech (TTS) research have prominently focused on incorporating expressiveness, prosody, and achieving more natural speech synthesis. Tacotron is a sequence-to-sequence architecture for speech synthesis that directly generates mel spectrogram frames from text [12]. An improved version of Tacotron, Tacotron2 [13] utilizes a combination of a sequence-to-sequence model and a modified Griffin-Lim algorithm [14] for high-quality speech synthesis. FastSpeech [15] is a text-to-speech (TTS) model that employs a non-autoregressive approach, enabling faster and parallelized generation of mel spectrograms from input text. FastSpeech 2 [16] improves upon its predecessor, FastSpeech, addressing issues like a complex teacher-student distillation process. Additionally, it tackles the one-to-many mapping problem in Text-to-Speech (TTS) more effectively by incorporating a unique variance adaptor that takes into account various speech variations such as pitch, energy, and more accurate duration as conditional inputs.

In this study, the aim is to explore how FastSpeech2 captures the prosody embeddings from text-only during the inference unlike speech plus text usage in training and heuristic-based feature computation in the context of both native and non-native speakers. Additionally, our analysis examines syllable and word-level stress patterns. We consider the energy and duration embeddings from Fastspeech2 for our analysis using native and non-native speech corpus (Tatoeba and ISLE). For the analysis, a selected subset of 3000 audios from these two datasets were annotated with word-level prominence. We perform the comparative analysis in the following three stages: 1) with Principal Component Analysis (PCA), 2) computing similarity and dissimilarity measures between features of stressed and unstressed groups and 3) classification based analysis in supervised and unsupervised manner. The study reveals that speech plus text provides better embeddings than text-only case in both native and non-native speech. For non-native speech, it is found that the embeddings provide superior performance for German speech as compared to Italian speech because of semantic and phonological similarities between English and German languages.

2. Dataset

Tatoeba[1] stands as an expansive and openly accessible compilation of English sentences accompanied by high-quality translations spanning over 300 languages. It serves as a comprehen-

[1] https://tatoeba.org/en

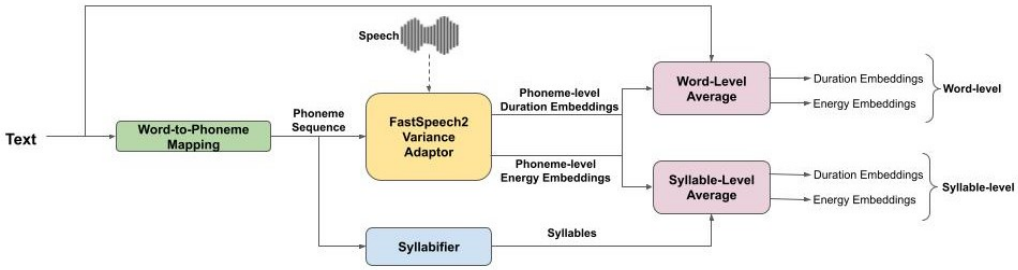

Figure 1: *Block Diagram showing the prosposed approach for obtaining embeddings from Fastspeech2 variance adaptor*

sive repository facilitating linguistic exploration and multilingual understanding [17, 18]. This dataset is continuously expanding through voluntary contributions from numerous members. We diligently curated a subset of the Tatoeba Dataset, specifically focusing on English audio files accompanied by their transcripts. Each audio file within this subset underwent a thorough manual annotation process to accurately identify and label prominence. This subset encompasses approximately 7122 instances of English words, excluding silences and noise. Adopting a word-centric unit of analysis, we successfully transcribed the prominence markers to the word-level, ensuring precise identification, and consolidated the prominence labels into a binary classification: 1 for prominent words and 0 for non-prominent words. This annotation resulted in 1083 prominent words and 6039 non-prominent words.

The **ISLE Corpus** [19] comprises 7,834 speech utterances from 46 non-native English learners (23 are German (GER), and 23 are Italian (ITA)). NIST [20] syllabification software is used to derive syllable transcriptions, and aligned syllable boundaries are obtained from phone transcriptions. It had 48868 syllables as stressed and 16693 syllables as unstressed. A combined subset of 2000 utterances, with 1000 from both German and Italian speakers, underwent word-level annotation, mirroring the methodology used in the earlier dataset. We specifically opted for a dataset from non-native speakers to thoroughly analyze the variability of stress patterns in both syllables and words.

3. Proposed Methodology

Figure 1 shows a block diagram describing the steps involved in embedding extraction, which is developed based on the variance adaptor in FastSpeech2. The block diagram has the following steps: 1) mapping from word sequence in text to phoneme sequence, 2) the variance adaptor in FastSpeech2, 3) syllabifier for converting phoneme sequence to syllable sequence for obtaining syllable-level embeddings, and 4) word and syllable level averaging for deducing the embeddings for the respective words and syllables.

3.1. FastSpeech2 Variance Adaptor

The variance adaptor is a neural network framework specifically to extract embeddings, which correspond to the following three acoustic components associated with the prosody: duration, energy and pitch. The variance adaptor was designed such that during the training stage, it learns the association between the text and embeddings using the three acoustic components extracted from the speech. Thus the training requires parallel speech and text data, referred to as speech plus text case. During the inference stage, with the learned association, the variance adaptor generates embeddings representing the three types of acoustic components from using only text, referred to as text-only case. In both stages, the embeddings are computed for each phoneme in the phoneme sequence of the text, which is obtained from an inbuilt word-to-phoneme mapping tool obtained based on pronunciation lexicon, grapheme-to-phoneme (G2P) conversion. A user is facilitated to modify the entries in the lexicon according to the requirement.

Architecture details: The variance adaptor in Fastspeech2 consists of a 2-layer 1D-convolutional network with ReLU activation, each followed by the layer normalization and the dropout layer, and an extra linear layer to project the hidden states into the output sequence. The 1D-convolution kernel sizes are configured to be 3, with the output sequence size set to 256. In the training phase, actual duration, pitch, and energy values extracted from the audio serve as input for the hidden sequence. Additionally, the text is transformed into a phoneme embedding sequence using an encoder with 4 feed-forward Transformer (FFT) blocks, resulting in a phoneme hidden sequence. Concurrently, true duration, pitch, and energy values are employed as targets to train the corresponding predictors.

Speech plus text case vs text-only case: The overall process of extracting embeddings remains consistent for both "text-only" and "speech plus text" cases, with the exception of the dotted arrow depicted in the block diagram. This dotted arrow is exclusive to the embedding extraction process during the "speech plus text" case, where speech input is supplied to the FastSpeech2 variance adaptor. This is because the speech plus text case needs to operate under the training mode of variance adaptor, which requires speech. However, for the text-only case, the variance adaptor can generate the embeddings in the inference model solely from the text without speech. In this work, it is proposed to analyse the embeddings from both cases as the prosody in the speech can vary for a given text. Thus, in the "text-only" case, one-to-many possibilities exist. So it is unsure about the relation between the prosody embeddings generated from the text-only case and the prosody that exists in the speech associated with the text used in the text-only case. Unlike the text-only case, in the speech plus text case, the obtained prosody embeddings could capture prosody in the speech due to the usage of the speech in the training process. Therefore, the anticipation is that the "speech plus text" approach offers more comprehensive prosody embeddings compared to the "text-only" case, which solely relies on textual cues. This motivated to analyse the embeddings extracted in these two distinct modes.

3.2. Word & Syllable level embeddings extraction

The variance adaptor provides the embeddings for each phoneme in the phoneme sequence obtained from the text after the word-to-phoneme mapping step. Typically, the prosody is conveyed through the acoustic variations in the syllable segments and the above i,e. suprasegmental level. In this work, the analysis is proposed to perform at the word and syllable level.

Text-only case: The word and syllable level embeddings are obtained from the phoneme level embeddings for each

phoneme associated with the text. For this, phoneme level embeddings are averaged across all the phonemes within each syllable or a word associated with the text, respectively to obtain syllable and word level embeddings. The embeddings derived from variance adaptor exhibit dimensions of (number of phonemes $\times 256$) and are subsequently transformed to dimensions of (number of words $\times 256$) and (number of syllables $\times 256$), respectively, for word and syllable level embeddings. The phonemes within each word are mapped using inbuilt word-to-phoneme mapping in the variance adaptor. Whereas, the phonemes association with syllables are obtained with a syllabifier, which maps one or more phonemes to a syllable thereby phoneme sequence associated with the text is converted to a respective syllable sequence.

Speech plus text case: Variance adaptor uses the phoneme sequence obtained from the inbuilt lexicon, which is based on native English speech. Typically, the lexicon contains multiple pronunciations and one of these is selected with the Montreal force-aligner used in the variance adaptor during the training. This process may result in processing errors in picking the correct pronunciation that exists in the speech. These errors would be more in the non-native speech i.e., speech from German and Italian in the ISLE corpus. To overcome these errors, the manual phonetic transcriptions available in the corpus are utilized. We modify the pronunciation lexicon by incorporating phoneme sequences for each word based on the manual transcriptions. This modification is performed dynamically for each speech and text pair while extracting the embeddings. The variance adaptor, while running in the training mode, refers to this speech and text pair specific modified lexicon during the embedding extraction. The training mode is run for only one epoch by fine-tuning the inbuilt pre trained LJSpeech [21] model. The choice of the one epoch is due to limited computational resources. Also, the analysis performed with embeddings extracted with the one epoch can be generalizable for those extracted with multiple epochs as with increasing epochs the model become better in learning the embeddings.

4. Experiments and Results

4.1. Experimental Setup

In our approach, we obtain both word-level and syllable-level embeddings for duration and energy using ISLE Corpus, by leveraging two distinct scenarios: one involving speech plus text case and the other is the text-only case whereas for Tatoeba corpus, we extract only word-level embeddings for both the cases. The prominence analysis at both the word and syllable levels is performed in three stages:

1. Using Principal Component Analysis (PCA) as it allows us to visualize linear projected nuanced variations in embeddings at both the word and syllable levels
2. Computing the distances between stressed and unstressed words or syllables.
3. Considering supervised and unsupervised categorization approaches.

The comparisons across all the three stages are performed with embeddings as well as heuristics-based features. The following sub-sections describe the feature computation and experimental setup for detecting word and syllable-level prominence.

4.1.1. Word Prominence

Prosodic information such as area under the F0 curve, Voiced-to-unvoiced ratio, F0 peak/valley amplitude and location etc and lexico-syntactic information like part-of-speech tags, word type (content word or a function word) were used to obtain **heuristics-based features** [5] from the word-level annotated subset of audios in the ISLE and Tatoeba datasets. These features were used to train a deep neural network (DNN) model having 6 dense layers along with Batch Normalization [22] and Dropout [23] in its architecture. After the first Dense layer (64 units) and the third Dense layer (32 units), Rectified Linear Unit (ReLU) activation function [24] is applied, followed by Batch Normalization and Dropout (0.3). This DNN model is trained for 50 epochs with Adam optimizer [25] and binary cross-entropy loss [26].

4.1.2. Syllable Prominence

We utilize the state-of-the-art 19-dimensional acoustic-based features, accompanied by 19-dimensional binary features capturing context dependencies (hereafter referred to as **heuristics-based features**), in line with [8]. Following the approach in [10], a DNN classifier, comprising 8 hidden layers with ReLU activation, Adam optimizer, and binary cross-entropy loss function, underwent training for 200 epochs to perform syllable-level stress detection. The analysis covered the entire ISLE corpus, which includes syllable-level annotations for all audio samples.

The DNN classifiers and K-mean clusters are modelled considering extracted embeddings and heuristic-based features. The recorded accuracies for both supervised and unsupervised methods for the heuristic-based features serve as a baseline to assess the effectiveness of the extracted embeddings.

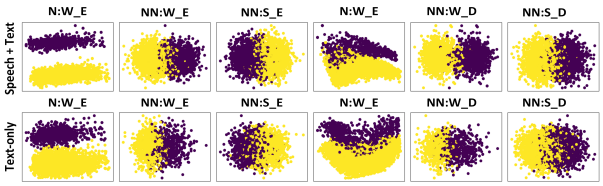

Figure 2: *Scatterplots comparing two principal components under Speech+Text and Text-only cases.('N' denotes native, 'NN' denotes Non-native, 'W': word level, 'S': Syllable level, 'E': Energy and 'D': Duration)*

4.2. PCA Based Analysis

Figure 2 shows the scatterplots of the two principal components obtained from the energy and duration embeddings under speech plus text and text-only cases for word and syllable level prominence in native and non-native speech conditions. The observations from the figure indicate that there is less overlap between components corresponding to energy embeddings as compared to duration emneddings, for stressed (purple dots) and unstressed (yellow dots) elements across speech plus text and text-only under word and syllable-levels. This indicates that the energy embeddings offer superior separation compared to duration embeddings. Notably, the embeddings exhibit less overlap when considering speech-plus-text compared to text-only scenarios. The findings suggest that incorporating speech data enhances the quality of the embeddings and this aligns with the claim mentioned in section 3.1 that prominence is particularly present in speech and to some extent in text.

4.3. Distance Metrics

Proximity measures hold considerable importance in determining the degree of similarity or dissimilarity between different embeddings. Drawing inspiration from [27] that distinctly delves into proximity measures tailored for numeric data attributes, the seven metrics were computed as listed below:

1. **Similarity Measures:**
   (a) Cosine Similarity $= \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$
   (b) Jaccard Similarity $= \frac{|X \cap Y|}{|X \cup Y|}$
2. **Dissimilarity Measures:**
   (a) Manhattan Distance $= \sum_{i=1}^{n} |x_i - y_i|$
   (b) Euclidean Distance $= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
   (c) Chebyshev Distance $= \max_i(|x_i - y_i|)$
   (d) Canberra Distance $= \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$
   (e) Mahalanobis Distance $= \sqrt{(x_i - y_i)^T \cdot S^{-1} \cdot (x_i - y_i)}$ where S is the covariance matrix.

In the above measures $x_i \in X$, $y_i \in Y$, where X and Y denote the subset of features that belong to stressed and unstressed categories, respectively. In this work, the features are taken from the following three types of feature sets: 1) energy embeddings, 2) duration embeddings and 3) heuristics-based features.

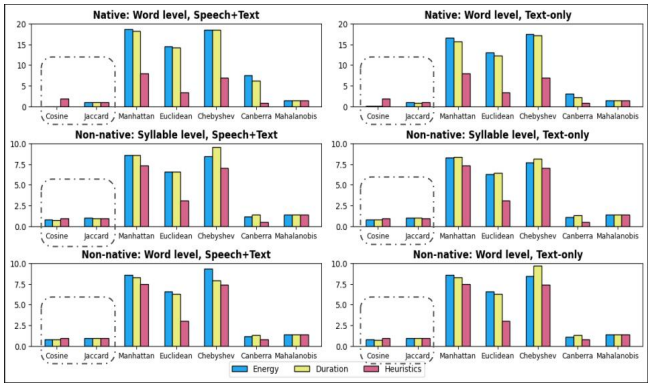

Figure 3: *Comparison of Distance metrics with three feature sets for native and non-native speech under text-only and Speech + Text cases (the box with dotted lines indicates dissimilarity measures)*

Out of the seven metrics, Cosine and Jaccard similarities capture the resemblance between stressed and unstressed features whereas the remaining capture the dissimilarity. Thus, it is expected to have higher and lower values respectively with similarity and dissimilarity metrics for better discrimination between stressed and unstressed categories for a given feature set.The seven metrics are computed between stressed and unstressed words and syllables, employing three distinct feature sets: energy and duration embeddings for both speech + text and text-only, and heuristics-based features. For all scenarios, the metrics for three feature sets (energy and duration embeddings and heuristics) are shown with bar graph in Figure 3.

From the figure, it is observed that the energy and duration embeddings exhibit greater values compared to the heuristics features when all dissimilarity metrics are used. This observation is further supported by the lower values of the similarity measures, indicating a higher degree of dissimilarity between the corresponding representations. The distinct patterns in the dissimilarity metrics highlight the unique characteristics and separability of energy and duration embeddings extracted from the FastSpeech2 between stressed and unstressed words and syllables.

4.4. Classification Analysis

A comparison with the baseline discussed in Section 4.1.1 and 4.1.2 was done by training the DNN classifiers, described in the aforementioned section, with the energy and duration embeddings. Additionally, as part of the unsupervised approach, K-Means clustering was applied on these embeddings to further analyze the stress patterns across words and syllables.The first half of the Table 1 shows the accuracies for native and non-native cases. The accuracies indicate that the embeddings yield higher accuracy in native scenarios compared to heuristics. Notably, DNN and K-Means gave 100% accuracy with energy embeddings in native speech+text case because of the complete separation of stressed and unstressed components in the PCA plot, indicating no overlap.

Table 1: *Comparison of Accuracies for Word and Syllable-Level Prominence Detection in Supervised and Unsupervised manner (K-M denotes K-Means Clustering, E,D,H denote Energy embeddings, Duration embeddings and Heuristics-based features respectively)*

		Word Prominence				**Syllable Prominence**			
		Speech+Text		**Text-only**		**Speech+Text**		**Text-only**	
		K-M	**DNN**	**K-M**	**DNN**	**K-M**	**DNN**	**K-M**	**DNN**
Native	E	**100**	**100**	75	**99**	–	–	–	–
	D	84.9	92.5	70	89	–	–	–	–
	H	66	86.5	66	86.5	–	–	–	–
Non - Native	E	84.2	**89.5**	81	**87.9**	82.6	**87.4**	80.3	**87.7**
	D	80.1	86.2	77.2	84.1	81.7	84.9	70.7	84.5
	H	75.8	78.5	75.8	78.5	75.1	78.5	75.1	78.5
GER	E	92.2	**92.7**	70	**89.4**	88.6	**90.1**	80.5	**89.3**
	D	88	90	81	88.4	80.2	87.5	78.5	88.3
	H	67	83	67	83	62.6	88.6	62.6	88.6
ITA	E	90	**92.3**	74	75.4	82.2	81.7	80.3	78.2
	D	81	87.7	71	73.7	78.8	77.9	75.19	73.9
	H	74	79	74	**79**	57.7	**88.2**	57.7	**88.2**

A separate analysis of the performance of the DNN classifier and K-Means clustering was done across different non-native scenatios namely Italian (ITA) and German (GER) in the second half. In non-native cases, GER consistently performs well in both speech + text and text-only, while ITA excels mainly in speech + text for word prominence and a similar trend is observed for syllable prominence between GER and ITA. This suggests that embeddings exhibit better representation in the presence of speech + text for non-native cases also, emphasizing sensitivity to speaker variability and nuances in syllabic emphasis. Notably, the linguistic proximity between English and German, both Germanic languages, contributes to their higher similarity compared to English and Romance languages (French, Italian, Spanish), which originate from Latin [28]. Therefore, the FastSpeech2 model, initially trained on the native LJSpeech Dataset, demonstrates better performance on non-native English spoken by German speakers than Italian speakers.

5. Conclusion

The study conducted in this work infers that the energy embeddings, particularly those obtained during native speech+text case, consistently outperform other scenarios. They also exhibit superior performance compared to duration embeddings in the native as well as non-native text-only scenario. Distances computed to measure the separation between stressed and unstressed words and syllables indicate minimal similarity and a notable degree of dissimilarity. This observation holds true in clustering and classification experiments across various scenarios, except for specifically Italian speakers in the non-native scenario. In this case, embeddings do not surpass baseline accuracies. This deviation can be attributed to the TTS pretraining on a native English dataset, with German exhibiting phonological similarities to English, contributing to enhanced performance.

6. References

[1] S. Werner and E. Keller, *Prosodic aspects of speech*. GBR: John Wiley and Sons Ltd., 1995, p. 23–40.

[2] I. Lehiste, *Suprasegmentals*. Cambridge: MIT Press, 1970.

[3] A. Cutler, "Lexical stress," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford: Blackwell, 2005, pp. 264–289.

[4] D. Wang and S. Narayanan, "An unsupervised quantitative measure for word prominence in spontaneous speech," in *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1, 2005, pp. I/377–I/380 Vol. 1.

[5] T. Mishra, V. K. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Interspeech*, 2012.

[6] D. Wang and S. Narayanan, "An acoustic measure for word prominence in spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 690–701, 2007.

[7] J. Tepperman and S. Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners," *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, vol. 1, 01 2005.

[8] C. Yarra, M. K. Ramanathi, and P. K. Ghosh, "Comparison of automatic syllable stress detection quality with time-aligned boundaries and context dependencies," in *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, 2019, pp. 79–83.

[9] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5845–5849.

[10] J. Mallela, P. S. Boyina, and C. Yarra, "A comparison of learned representations with jointly optimized vae and dnn for syllable stress detection," in *Speech and Computer: 25th International Conference, SPECOM 2023, Dharwad, India, November 29 – December 2, 2023, Proceedings, Part II*, 2023, p. 322–334.

[11] A. Cruttenden, *Gimson's Pronunciation of English*, 8th ed. Routledge, 2014.

[12] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. A. J. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017.

[13] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[14] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, 1983, pp. 804–807.

[15] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, *FastSpeech: fast, robust and controllable text to speech*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[16] 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," 2022.

[17] A. Mikel and S. Holger, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *arXiv:1812.10464v2*, 2018.

[18] J. Tiedemann, "Parallel data, tools and interfaces in opus," in *International Conference on Language Resources and Evaluation*, 2012.

[19] E. Atwell, P. Howarth, and D. Souter, "The isle corpus: Italian and german spoken learner's english," *ICAME Journal: International Computer Archive of Modern and Medieval English Journal*, vol. 27, pp. 5 – 18, April 2003.

[20] B. Fisher, "tsylb2-1.1 syllabification software, national institute of standards and technology," 1996.

[21] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 448–456.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 06 2014.

[24] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[26] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, 10 2020.

[27] V. Mehta, S. Bawa, and J. Singh, "Analytical review of clustering techniques and proximity measures," *Artificial Intelligence Review*, vol. 53, pp. 1–29, 12 2020.

[28] L. K. Şenel, V. Yücesoy, A. Koç, and T. Çukur, "Measuring cross-lingual semantic similarity across european languages," in *2017 40th International Conference on Telecommunications and Signal Processing (TSP)*, 2017, pp. 359–363.