INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Spectral and Latent Speech Representation Distortion for TTS Evaluation

Thananchai Kongthaworn, Burin Naowarat, Ekapol Chuangsuwanich

Department of Computer Engineering, Chulalongkorn University, Thailand
{6370120621, 6270145221}@student.chula.ac.th, ekapolc@cp.eng.chula.ac.th

Abstract

One of the main problems in the development of text-to-speech (TTS) systems is its reliance on subjective measures, typically the Mean Opinion Score (MOS). MOS requires a large number of people to reliably rate each utterance, making the development process slow and expensive. Recent research on speech quality assessment tends to focus on training models to estimate MOS, which requires a large number of training data, something that might not be available in low-resource languages. We propose an objective assessment metric based on the DTW distance using the spectrogram and the high-level features from an Automatic Speech Recognition (ASR) model to cover both acoustic and linguistic information. Experiments on Thai TTS and the Blizzard Challenge datasets show that our method outperformed other baselines in both utterance- and system-level by a large margin in terms of correlation coefficients. Our metric also outperformed the best baseline by 9.58% when used in head-to-head utterance-level comparisons. Ablation studies suggest that the middle layers of the ASR model are most suitable for TTS evaluation when used in conjunction with spectral features.

Index Terms: TTS evaluation, speech synthesis, speech recognition

1. Introduction

Text-to-Speech (TTS) systems are commonly evaluated by Mean Opinion Score (MOS) since the audio quality strongly depends on human perception. MOS is simply preference scores from listeners with the range of 1 to 5. The MOS is intuitive, but it can only be obtained solely through human participation which is expensive and time-consuming. In order to make TTS evaluation less expensive and more robust, many objective quality assessments had been explored as alternatives.

Objective measures are mainly classified into two categories: intrusive and non-intrusive. Intrusive methods rely on the existence of human reference recordings which are used to compare against the synthesized speeches. Due to this limitation, there is an increasing focus on non-intrusive assessments. Any synthesized speech can be scored according to the predictions from machine learning models such as Support Vector Machines [1] and Neural Networks [2, 3, 4, 5, 6]. Predictive models are easy to use, but they require a large amount of costly labeled data for training. This usually requires different kinds of TTS models to be trained and evaluated, which can be hard for low resourced-languages. On the other hand, intrusive assessments, needing no expensive training data, use the distance between synthesized and reference audios as a proxy for quality estimates. In this work, we will focus on this kind of assessment method.

Traditionally, intrusive approaches compare the similarity between spectral features of two audios [7, 8, 9, 10]. Lately, hidden representations from Automatic Speech Recognition (ASR) models have been used instead of traditional features [11].

The ASR embeddings from end-to-end models have been studied in various works in order to understand their relationships to human concepts [12, 13, 14]. Linguistic properties [12, 14], such as phonetics information, and the speaker characteristics [13], such as speaker identity, are found in the different layers of the model and are proven to be effective for estimating the goodness of synthetic audios [11]. Bińkowski et al. have shown that the Fréchet distance between the distributions of high-level features from synthesized and reference speech is highly correlated to MOS and can be used as an objective evaluation of TTS models [11]. However, they did not analyze the different effects each layer of the model might have. ASR features might also fail to discern noise or artifacts due to the fact that ASR models are often trained to ignore such distinctions. Moreover, their metric, which is based on distributions, cannot be used to estimate errors and artifacts on the utterance level.

To score on the utterance level, dynamic time warping (DTW) [15, 16] was often used as an utterance-level distance function for comparing two audios [7, 8, 9, 10]. Using low-level signal representations, such as spectral features and MFCCs, this approach better captures the fine artifacts in synthesized speech. However, nowadays, TTS models have become very high quality and often indistinguishable by using only the low-level signal representations.

In this work, we present a new intrusive assessment method, Spectral and Latent Speech Representation Distortion (SLSRD), for TTS evaluation. SLSRD uses both low-level signal and high-level linguistic information for measuring both the naturalness and the correctness of the synthesized audios. The spectrogram is used to represent the signal information and high-level linguistic representations are extracted from the hidden units of an ASR model. Extensive experiments on the Thai dataset and the Blizzard Challenge [17, 18] show that SLSRD outperforms the baselines in terms of correlation coefficient. SLSRD also has a higher agreement rate with human raters in head-to-head evaluation scenarios than other baselines.

2. Proposed Method

The main idea of our proposed method, SLSRD, is to evaluate the sound quality using both acoustic and phonetic properties in the objective evaluation, mimicking the opinion scores provided by humans. This is done by using traditional spectral features in tandem with latent features extracted from an ASR model to compute the distance between the reference and the synthesized speech using DTW. Our implementation is publicly available at `https://github.com/SLSCU/SLSRD`.

2.1. Preprocessing

The raw speech audios were preprocessed to remove aspects that are unrelated to quality assessment. Silence were removed from on both ends of the recording by using a simple energy-

http://dx.doi.org/10.21437/Interspeech.2021-2258

based method. The inner silences were kept as is to preserve prosody. We also normalized the loudness of each synthesized utterance to be the same as the reference recording.

2.2. Spectral features

The spectrogram is used as the low-level signal representation to capture the fine details in the speech.

Given the synthesized waveform, $\mathbf{x} \in \mathbb{R}^{T_x}$, and the reference waveform, $\mathbf{y} \in \mathbb{R}^{T_y}$, we compute the spectrograms, $S_x \in \mathbb{R}^{N_x \times L}$ and $S_y \in \mathbb{R}^{N_y \times L}$, respectively, where T_x and T_y are the number of synthesized and reference samples, L is the number of frequency bins, N_x and N_y are the number of synthesized and reference frames. These features are standardized using utterance-level statistics.

2.3. Latent features

The latent features extracted from the ASR model are used for capturing phonetic properties from the input signal. It has been shown that the phonetic information in the features are rich enough for speech synthesis [13]. By measuring the difference between the features of the synthesized and reference audios, we can roughly estimate the human-likeness and pronunciation correctness of the synthesized speeches.

We compute the high-level features, $h_x \in \mathbb{R}^{P_x \times K}$ and $h_y \in \mathbb{R}^{P_y \times K}$, from the synthesized and reference speeches by using the synthesized and reference audios as the input to the ASR model. P_x and P_y, are the number of frames for the synthetic and reference audios, respectively. Again, the features, h_x and h_y , are standardized.

2.4. Scoring with DTW

Since the reference and the synthesized audio can have different lengths, DTW is a great choice for computing the distance which can then be used as a quality measure. The SLSRD score is the DTW distance between two audios using the concatenated spectral and latent features. We use the Euclidean distance for the frame-level distance. Since the two features are based on different frame rates, the high-level ASR features are upsampled before the concatenation to prevent the time-resolution mismatch. SLSRD is calculated according to the following equation: (1).

$$\text{SLSRD}\ (x, y) = \frac{1}{T\sqrt{C}} DTW(S_x \oplus h_x, S_y \oplus h_y) \qquad (1)$$

where T is number of points used in the DTW backtraced, $\oplus$ denotes the concatenate operation, $C = L + K$ is the size of the concatenated features, and D is the distance score, the lower the better.

A version without the spectrogram features, LSRD, can also be computed as shown:

$$\text{LSRD}\ (x, y) = \frac{1}{T\sqrt{K}} DTW(h_x, h_y) \qquad (2)$$

3. Experimental setup

In this section, we provide details on the hyperparameters, training conditions, evaluation criteria, and baselines.

3.1. Features

We used 200 bins for the spectrogram features. It was computed using the Hann window with 20ms window size, and 10ms hop size. The sampling rate were kept at 16 kHz for all experiments.

The ASR model used to extract the latent feature was Wav2Letter+, an ASR model with 17 convolutional layers [19]. For English, we used the pre-trained model provided in the OpenSeq2Seq[1] library. For Thai, we trained the model from scratch using the official NVIDIA implementation without any modification except for the alphabets. Manually transcribed 636-hour Thai utterances taken from YouTube videos were used for training. For more details regarding the dataset see [20]. The model achieved 6.76% Character Error Rate (CER).

As for the DTW scoring, we used FastDTW [21] to approximate the DTW distance instead of the standard algorithm. We only found minimal difference between the two, but the approximation method had superior time and space performance.

3.2. Quality Assessment datasets

3.2.1. Blizzard Challenge dataset

We used the test results from the Blizzard Challenge 2012 and 2013 [17, 18] that contains audios synthesized using HMM-based, diphone, unit-selection, and hybrid TTS models. We used the subset of EH1 test results from the 2012 and 2013 competitions in our experiments and denoted them as BLZ2012 and BLZ2013. In both subsets, we discarded audios that did not have references. There remained 220 and 387 data samples with 10 and 9 TTS models for the EH1 test set of BLZ2012 and BLZ2013, respectively.

3.2.2. Thai dataset

We also created our Thai dataset (TH) for evaluating the proposed objective metric. The synthesized audios were created from two variations of end-to-end TTS models, the autoregressive Tacotron 2 [22] and the non-autoregressive FastSpeech [23] which focuses more on inference speed. Both models were trained on the grapheme level using a 29-hour (20k utterances) single-speaker Thai dataset with the same training hyperparameters as in the original papers. The grapheme duration information for training FastSpeech was extracted from Tacotron 2. The WaveGlow vocoder [24] was used for generating the waveform values from the predicted Mel frequency values.

For evaluation, thirty-four new sentences were used for comparison. The synthesized audios were then scored by 134 Thai native speakers with the criteria of naturalness and sound quality. The highest MOS of 5 was given to the goods, and the lowest MOS of 1 was given to the bads. The MOS scores from each model were summarized in Table 1.

Table 1: *MOS values for the Thai TTS dataset. We report the average with 95% confidence interval.*

Model	Checkpoint	MOS
Real speech	n/a	4.17 ± 0.07
Tacotron 2	90k	3.46 ± 0.07
Tacotron 2	96k	3.28 ± 0.08
Tacotron 2	100k	3.33 ± 0.08
FastSpeech	1.15M	3.18 ± 0.09

3.3. Evaluation criteria

To evaluate the performance of an objective metric, we compute the correlation between the objective score and the MOS value. We report the Pearson correlation coefficient, r, that measures

[1] https://github.com/NVIDIA/OpenSeq2Seq

the linear relationship and the Kendall Tau Rank Correlation, τ, that measures the correspondence between two rankings.

The evaluation can be measured on two different granularities: utterance-and system-level scoring. The utterance-level score compares the objective score and the average of humans' subjective scores of a single utterance. As a result, there is one score for one utterance. The system-level score compares the average of objective scores and the average of subjective scores of the same system. As a result, there is one score per system.

3.4. Baselines

We compare SLSRD with 7 other existing and frequently used speech quality objective metrics. Detailed descriptions of the 7 baselines are as follow:

1. Mel cepstral distortion (MCD) [7]: this method only uses the MFCC instead of both spectral and latent features. We used 20 channels, 50ms window, and 12.5ms hop size for the MFCC calculation.
2. Mel spectral distortion (MSD) [25]: this metric is very similar to MCD. The only difference is that the 80-channels log-mel magnitude spectrogram is used instead of the MFCC.
3. Perceptual Evaluation of Speech Quality (PESQ) [26]: defined by the ITU-T recommendation P.862, PESQ is designed to measure the speech quality in telecommunication. It was also used for TTS evaluation [27].
4. Virtual Speech Quality Objective Listener (ViSQOL) [28] : this method measures the similarity between spectrograms with two kinds of alignment, global and local patches, using Neurogram Similarity Index Measure [29] (NSIM) as the distance function.
5. Fréchet DeepSpeech Distance (FDSD) [11]: a Fréchet distance between features distributions of synthesized and reference audios. Instead of DeepSpeech2's penultimate layer (before the softmax layer) used in the original work, we used Wav2Letter+ instead.
6. MOSNet [30]: a MOS-predictive model trained using the Voice Conversion Challenge (VCC) 2018 [31] dataset. We used the pre-trained model provided by speechmetrics[2].
7. NISQA-TTS [6]: a MOS-predictive model trained on listening score data from various years of the Blizzard Challenge and the VCC2018 to estimate the quality of synthesized speech.

4. Results and Discussion

4.1. Correlation with MOS

In this experiment, we compared the Pearson (r) and Kendall (τ) correlation coefficients of the proposed metrics, LSRD and SLSRD, to other baselines using TH, BLZ2012, and BLZ2013 datasets. The results are summarized in Table 2.

The proposed SLSRD had superior utterance-level performance than other metrics by a large margin in all three datasets. Adding the spectral features improve the performance over just using ASR features especially on BLZ2012. As for the system-level evaluation, SLSRD ranks the top three models better than NIQSA-TTS as shown in Figure 1. FDSD is not applicable for utterance-level evaluation as it computes the distance between distributions.

[2]https://github.com/aliutkus/speechmetrics

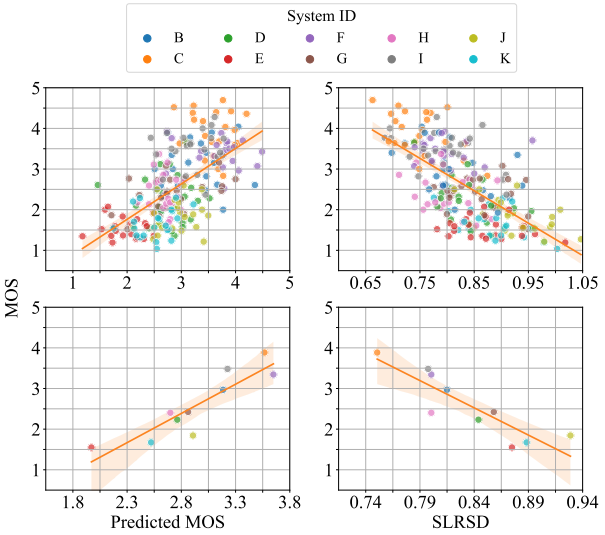

Figure 1: *Regression analysis for utterance-level (top) and system-level (bottom) of NIQSA-TTS (left) and SLSRD (right) on BLZ2012. Translucent bands represent the 95% confidence interval for the regression estimate.*

Note that Table 2 also highlights another downside for the use of predictive models. MOSNet which was trained on VCC2018 only has a low or even negative correlation with the MOS on the Thai dataset and BLZ2012 (denoted with †), which shows that predictive models can be language dependent, and hard to generalize. SLSRD and LSRD demonstrate the simple yet effective use of spectral and latent ASR features for evaluating TTS models. This can be helpful in TTS development on low-resource languages which might not have access to large-scale MOS data from multiple TTS systems.

4.2. Ablation study on the layer location to extract the latent feature

In this section, we investigated how changing the choice of the hidden layer affects the correlation coefficient in the Thai dataset. As pointed out by previous works [12, 14, 13], different layers of the ASR network play different roles and capture different kinds of information. The purpose of this experiment is to find the layer that is most suitable for TTS evaluation.

Table 3 shows that outputs from ReLU_9, a layer in the middle of the entire network, yields the highest correlation coefficient, contrary to [11] which uses the penultimate layer. According to [12, 14], the intermediate layers better capture the phonetic properties than the top layers. [13] also reported that the speaker characteristics were gradually removed in the deeper layers. This result suggests that the sound quality might rely on both the phonetic and speaker information.

LSRD, which has no spectral features, has the best Kendall correlation by using latent features of the shallowest layer. However, the deeper layer is better for SLSRD, since lower layers and spectral features provide highly correlated information. Note that the best τ values for SLSRD and LSRD are tied at 0.444. This might be due to the fact that our simple feature concatenation method cannot fully utilize the additional information. This is a venue for further investigation.

Table 2: *Correlation values between different objective assessment methods and MOS. For easier comparison, the absolute values are shown (higher is always better). The asterisk symbol (*) denotes that the predictive model uses the dataset as the training data, thus, cannot be used for evaluation. The dagger symbol (†) indicates negative correlation coefficients of the MOSNet model. Actual values are -0.643 and 0.000 in the TH dataset, and -0.268, -0.219, -0.390 and -0.289 in BLZ2012.*

Method	TH				BLZ2012				BLZ2013			
	utterance		system		utterance		system		utterance		system	
	$\lvert r\rvert$	$\lvert\tau\rvert$	$\lvert r\rvert$	$\lvert\tau\rvert$	$\lvert r\rvert$	$\lvert\tau\rvert$	$\lvert r\rvert$	$\lvert\tau\rvert$	$\lvert r\rvert$	$\lvert\tau\rvert$	$\lvert r\rvert$	$\lvert\tau\rvert$
PESQ [26]	0.073	0.122	0.397	0.0	0.195	0.174	0.370	0.333	0.023	0.057	0.186	0.167
MCD [7]	0.409	0.251	0.795	0.667	0.309	0.199	0.371	0.244	0.173	0.120	0.374	0.167
MSD [25]	0.414	0.251	0.890	0.667	0.279	0.182	0.403	0.333	0.243	0.170	0.437	0.333
VISQOL [28]	0.382	0.273	0.767	0.667	0.488	0.337	0.692	0.556	0.301	0.208	0.636	0.5
FDSD [11]	n/a	n/a	0.768	**1.0**	n/a	n/a	0.545	0.527	n/a	n/a	0.672	0.611
MOSNet [30]	0.098	0.075	†	†	†	†	†	†	0.490	0.352	0.662	0.556
NISQA-TTS [6]	0.058	0.004	0.774	0.667	0.615	0.426	**0.891**	**0.733**	*	*	*	*
LSRD (ours)	0.638	0.426	**0.907**	**1.0**	0.528	0.362	0.659	0.644	0.575	0.398	0.777	0.611
SLSRD (ours)	**0.640**	**0.444**	0.903	**1.0**	**0.667**	**0.475**	0.879	**0.733**	**0.583**	**0.399**	**0.799**	**0.667**

Table 3: *The correlation coefficients when different layers were used to extract the latent features.*

Layer	LSRD		SLSRD	
	$\lvert r\rvert$	$\lvert\tau\rvert$	$\lvert r\rvert$	$\lvert\tau\rvert$
ReLU_3	0.585	**0.444**	0.555	0.391
ReLU_6	0.627	0.439	0.6	0.420
ReLU_9	**0.638**	0.426	**0.640**	**0.444**
ReLU_{12}	0.598	0.421	0.611	0.416
ReLU_{15}	0.584	0.421	0.575	0.388
ReLU_{17}	0.578	0.389	0.578	0.38

4.3. Head-to-head comparison between synthesized audios

In this experiment, we studied the agreement between SLSRD and human annotations in head-to-head quality comparison.

The human raters were presented with audio pairs of the same Thai sentence but generated from different models and asked to select the better one or indicate a tie. The same pair of audios were also scored by our approach and other top performers in order to study whether objective scoring methods match human judgement in a head-to-head scenario. There were a total of 204 different pairings. The same set of raters from the MOS experiment were asked to judge 10 randomly selected pairs.

To remove ambiguous pairs in the analysis, we only considered pairs that resulted in a majority. The majority option must have at least three more votes than the runner-up[3]. This leaves 110 pairs of which 17 of them were voted to be of equal quality.

SLSRD outperformed the best baseline by 9.58% absolute as illustrated in Table 4. Figure 2 shows the boxplot of the difference between SLSRD scores for each kind of audio pair. The score difference shows a clear separation between each type. For the pairs that were judged to be equal, the SLSRD score difference is very close to zero.

5. Conclusion

We proposed SLSRD for TTS evaluation. SLSRD is the DTW distance between the reference and the synthesized audios using spectrogram and ASR features. Experiments on the Thai and Blizzard Challenge datasets showed that SLSRD outperformed other objective measures by a in terms of correlation with the MOS and agreement rate with human raters in head-to-head comparisons. Our metric can be used to guide TTS development in any language requiring just reference speeches and an ASR model, which can be acquired more easily than predictive methods that require an assessment dataset.

[3]We have also tried smaller majority cutoffs and observed similar trends.

Table 4: *Agreement rates between objective measures and human on the head-to-head comparison between synthesized audios*

Method	Human agreement rate (%)
MCD	57.45
MSD	65.95
LSRD	67.02
SLSRD	**75.53**

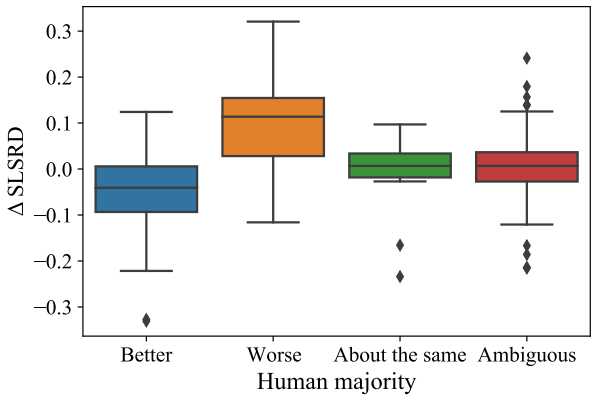

Figure 2: *Boxplot of the difference between SLSRD scores for each kind of audio pair in the head-to-head experiment.*

6. Acknowledgements

The authors would like to thank the members of the Thai Natural Language Processing Facebook group and other volunteers who participated in the assessment of the Thai dataset. We also would like to thank the organizers of the Blizzard challenges for providing the quality assessment datasets. Lastly, we wish to thank New Era AI Robotic Inc. for providing the Thai datasets and infrastructures used in this work.

7. References

[1] M. H. Soni and H. A. Patil, "Non-intrusive quality assessment of synthesized speech using spectral features and support vector regression," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 127–133. [Online]. Available: http://dx.doi.org/10.21437/SSW.2016-21

[2] Qiang Fu, Kechu Yi, and Mingui Sun, "Speech quality objective assessment using neural network," in *ICASSP 2000*, vol. 3, 2000, pp. 1511–1514 vol.3.

[3] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Interspeech*, 2018.

[4] M. Tang and J. Zhu, "Text-to-speech quality evaluation based on lstm recurrent neural networks," in *2019 International Conference on Computing, Networking and Communications (ICNC)*, 2019, pp. 260–264.

[5] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019*, 2019, pp. 631–635.

[6] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Interspeech 2020*, 2020, pp. 1748–1752. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2382

[7] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128 vol.1.

[8] S. Wang, A. Sekey, and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 819–829, 1992.

[9] S. Rao, C. Mahima, S. Vishnu, S. Adithya, A. Sricharan, and V. Ramasubramanian, "TTS evaluation: Double-ended objective quality measures," in *2015 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2015, pp. 1–6.

[10] H. B. Sailor and H. A. Patil, "Fusion of magnitude and phase-based features for objective evaluation of TTS voice," *Proceedings of the 9th International Symposium on Chinese Spoken Language Processing, ISCSLP 2014*, pp. 521–525, 2014.

[11] M. Bińkowski, J. Donahue, S. Dieleman, A. Clark, E. Elsen, N. Casagrande, L. C. Cobo, and K. Simonyan, "High fidelity speech synthesis with adversarial networks," in *International Conference on Learning Representations*, 2020.

[12] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Advances in Neural Information Processing Systems (NIPS)*, December 2017.

[13] C. Li, P. Yuan, and H. Lee, "What does a network layer hear? Analyzing hidden representations of end-to-end ASR through speech synthesis," in *ICASSP 2020*, 2020, pp. 6434–6438.

[14] Y. Belinkov, A. Ali, and J. Glass, "Analyzing phonetic and graphemic representations in end-to-end automatic speech recognition," in *Advances in Neural Information Processing Systems (NIPS)*, 09 2019, pp. 81–85.

[15] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[16] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proceedings of the Seventh International Congress on Acoustics, Budapest*, vol. 3. Budapest: Akadémiai Kiadó, 1971, pp. 65–69.

[17] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proceedings of the Blizzard Challenge workshop 2012*, 2012.

[18] ——, "The Blizzard Challenge 2013," in *Proceedings of the Blizzard Challenge workshop 2013*, 2013.

[19] O. Kuchaiev, B. Ginsburg, I. Gitman, V. Lavrukhin, J. Li, H. Nguyen, C. Case, and P. Micikevicius, "Mixed-Precision Training for NLP and Speech Recognition with OpenSeq2Seq," 2018.

[20] B. Naowarat, T. Kongthaworn, K. Karunratanakul, S. H. Wu, and E. Chuangsuwanich, "Reducing Spelling Inconsistencies in Code-Switching ASR using Contextualized CTC Loss," in *ICASSP 2021*, 2021.

[21] S. Salvador and P. Chan, "FastDTW: Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[22] 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 *ICASSP 2018*, 2018, pp. 4779–4783.

[23] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "FastSpeech: Fast, Robust and Controllable Text to Speech," in *Annual Conference on Neural Information Processing Systems 2019*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 3165–3174.

[24] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in *ICASSP 2019*, 2019, pp. 3617–3621.

[25] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *ICASSP 2021*, 2021.

[26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP 2001*, vol. 2, 2001, pp. 749–752 vol.2.

[27] M. Cerňak and M. Rusko, "An evaluation of a synthetic speech using the PESQ measure," in *Proceedings of Forum Acusticum 2005*, 2005.

[28] M. Chinen, F. S. C. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[29] A. Hines and N. Harte, "Speech intelligibility prediction using a Neurogram Similarity Index Measure," *Speech Communication*, vol. 54, pp. 306–320, 02 2012.

[30] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019.

[31] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 195–202. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2018-28