การใช้คุณสมบัติทางสเปคตรัลกับคุณลักษณะซ่อนสำหรับการประเมินระบบสังเคราะห์เสียงพูด

นายธนัญชัย คงถาวร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตร์มหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

SPECTRAL AND LATENT REPRESENTATION DISTORTION FOR TTS EVALUATION

Mr. Thananchai Kongthaworn

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering in Computer Engineering

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2021

Thesis Title	SPECTRAL AND LATENT REPRESENTATION DISTORTION FOR TTS EVALUATION
By	Mr. Thananchai Kongthaworn
Field of Study	Computer Engineering
Thesis Advisor	Ekapol Chuangsuwanich, Ph.D.
Thesis Co-advisor	Assoc. Prof. Atiwong Suchato, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master of Engineering

.... Dean of the FACULTY OF ENGINEERING

(Prof. Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

.... Chairman

(Assoc. Prof. Proadpran Punyabukkana, Ph.D.)

.... Thesis Advisor

(Ekapol Chuangsuwanich, Ph.D.)

.... Thesis Co-advisor

(Assoc. Prof. Atiwong Suchato, Ph.D.)

.... External Examiner

(Kwanchiva Thangthai, Ph.D.)

ธนัญชัย คงถาวร: การใช้คุณสมบัติทางสเปคตรัลกับคุณลักษณะซ่อนสำหรับการประเมินระบบสังเคราะห์เสียงพูด. (SPECTRAL AND LATENT REPRESENTATION DISTORTION FOR TTS EVALUATION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ดร. เอกพล ช่วงสุวนิช, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : รศ.ดร.อติวงศ์ สุชาโต 42 หน้า.

ปัญหาหลักประการหนึ่งในการพัฒนาระบบแปลงข้อความเป็นเสียงพูด (TTS) คือการวัดนิยมใช้การวัดเชิงอัตวิสัยอย่าง Mean Opinion Score (MOS) ซึ่งต้องใช้คนจำนวนมากในการให้คะแนนเสียงพูดแต่ละเสียงเพื่อความน่าเชื่อถือ ทำให้กระบวนการพัฒนาช้าและมีค่าใช้จ่ายสูง งานวิจัยล่าสุดเกี่ยวกับการประเมินคุณภาพเสียงพูดมีแนวโน้มที่จะมุ่งเน้นไปที่ใช้สร้างโมเดลมาประเมิน MOS ซึ่งต้องใช้ข้อมูลการฝึกฝนจำนวนมาก ซึ่งข้อมูลเหล่านี้อาจไม่มีให้ใช้งานในภาษาที่ใช้ทรัพยากรต่ำ งานนี้จึงนำเสนอการประเมินเชิงวัตถุประสงค์แบบ DTW โดยใช้สเปกโตรแกรมและคุณสมบัติระดับสูงจากแบบจำลองการรู้จำเสียงอัตโนมัติ (ASR) เพื่อให้ครอบคลุมทั้งข้อมูลเสียงและภาษาศาสตร์ การทดลองบนชุดข้อมูล Thai TTS และ Blizzard Challenge แสดงให้เห็นว่าวิธีการที่นำเสนอมีประสิทธิภาพเหนือกว่าวิธีการวัดอื่นๆ ที่นำมาเป็นบรรทัดฐาน ทั้งในระดับประโยคและระดับระบบในแง่ของค่าสัมประสิทธิ์สหสัมพันธ์ เมตริกของเรายังทำได้ดีกว่าบรรทัดฐานที่ดีที่สุด 9.58% เมื่อใช้ในการเปรียบเทียบระดับประโยคแบบตัวต่อตัว จากการศึกษาเพิ่มเติมแนะนำว่าชั้นกลางของแบบจำลอง ASR เหมาะสมที่สุดสำหรับการประเมิน TTS เมื่อใช้ร่วมกับคุณลักษณะสเปกตรัม

ภาควิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อนิสิต	....
สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่อ อ.ที่ปรึกษาหลัก	....
ปีการศึกษา	2564	ลายมือชื่อ อ.ที่ปรึกษาร่วม	....

6370120621: MAJOR COMPUTER ENGINEERING

KEYWORDS: TTS / EVALUATION / MEASUREMENT

THANANCHAI KONGTHAWORN : SPECTRAL AND LATENT REPRESENTATION DISTORTION FOR TTS EVALUATION. ADVISOR : Ekapol Chuangsuwanich, Ph.D., THESIS COADVISOR : Assoc. Prof. Atiwong Suchato, Ph.D., 42 pp.

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.

Department	:	Computer Engineering	Student's Signature	....
Field of Study	:	Computer Engineering	Advisor's Signature	....
Academic Year	:	2021	Co-advisor's signature	....

Acknowledgements

I would like to thank Professor Ekapol Chuangsuwanich, my advisor, for driving me into the speech processing field and accepting me to enter graduate school. He always helps and gives his guidance on how to do research in a broad range of fields with novelty. I also would like to thank Burin Naowarat, my lab mate, who help me to write a conference paper and be a good friend during my study at graduate school.

I 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. I 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.

CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Chapter I

INTRODUCTION

Nowadays, Text-to-Speech (TTS) systems have advanced to synthesize speech close to human-level based on deep learning approaches. However, there is still a problem in indicating the model's performance during the development process when the objective loss is not reliable on model selection. A common approach is to evaluate TTS using the Mean Opinion Score (MOS) result depending on human perception. MOS is rating scores from listeners in the range of 1 to 5 for quality and naturalness of speech audio. MOS is simple but not suitable for TTS development due to expensive and time-consuming. According to these problems, many objective quality assessments had been developed for TTS evaluation.

Objective quality assessments 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 (Soni and Patil, 2016) and Neural Networks (Qiang Fu et al., 2000; Fu et al., 2018; Tang and Zhu, 2019; Avila et al., 2019; Mittag and Möller, 2020). 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 (Kubichek, 1993; Wang et al., 1992; Rao et al., 2015; Sailor and Patil, 2014). Lately, hidden representations from Automatic Speech Recognition (ASR) models have been used instead of traditional features (Bińkowski et al., 2020).

The ASR embeddings from end-to-end models have been studied in various works in order to understand their relationships to human concepts (Belinkov and Glass, 2017; Li et al., 2020; Belinkov et al., 2019). Linguistic properties (Belinkov and Glass, 2017; Belinkov et al., 2019), such as phonetics information, and the speaker characteristics (Li et al., 2020), 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 (Bińkowski et al., 2020).

Bińkowski et al. (2020) has 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 (Bińkowski et al., 2020). 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) (Sakoe and Chiba, 1978; Sakoe and Chiba, 1971) was often used as an utterance-level distance function for comparing two audios (Kubichek, 1993; Wang et al., 1992; Rao et al., 2015; Sailor and Patil, 2014). 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 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 (King and Karaiskos, 2012, 2014; Wu et al., 2019) show that SLSRD outperforms the baselines in terms of the correlation coefficient. SLSRD also has a higher agreement rate with human raters in head-to-head evaluation scenarios than other baselines. Many parts of this thesis was reported in InterSpeech 2021 (Kongthaworn et al., 2021).

1.1 Objective

The goal of this thesis is to evaluate TTS systems by simulating human perception by utilizing spectral and latent data from the ASR model. We uses speech synthesis from various TTS systems such as unit selection, Hidden Markov Model, Hybrid, and Deep Learning to test the efficacy of proposed method.

1.2 Scope of Work

The area of this thesis is to assess the performance of TTS systems with the proposed method by analyzing the listening results corpus from four datasets, which are our in-house dataset as well as two datasets from the Blizzard Challenge 2012-2013 and 2019.

Chapter II

BACKGROUND

2.1 Speech assessment of TTS

Most speech audios are assessed using a subjective assessment that was human-engaged. Subjective assessment assesses speech audios in many attributes depending on The objective usage of systems and the user's needs. Most attributes usually assessed are comprehensibility, intelligibility, and quality. Comprehensibility assesses the listeners' understanding of the content of the speech audio by asking questions that consider how well a listener has understood. Whereas intelligibility assesses the content of speech audio is consistent with the given text by asking the listener to transcribe. Nowadays, TTS systems have become more highly intelligible causing less need for intelligibility measurements. Quality is the most used attribute to assess nowadays and is highly subjective by measuring many attributes of speech audios (e.g. intelligibility and naturalness) in one score. A listener was asked to score how the goodness of speech audio in various ranges. However, subjective quality assessment is difficult to be reliable and accurate caused of the different opinions of listeners. Many TTS studies have reported results that cannot be compared because the results are subjective. Other disadvantages include the fact that it is time-consuming and costly. Thus, objective quality assessments for TTS evaluation were developed to solve this problem.

An objective assessment is more exact and neutral than a subjective assessment because it is based on perception theory and signal processing. Objective assessment can be classified into intrusive and non-intrusive assessments. Intrusive assessments are metrics that need a reference for measuring goodness. It's usually done by calculating the similarity between synthesized and reference speech audio using spectral features such as MFCCs, and spectrogram. However, the intrusive assessment that exists nowadays does not match well with human opinion. For example, the listeners felt that the sounds were totally worse, but they just gave the score of ”good.” Because of this constraint, the intrusive assessment was mostly employed in tuning and we used subjective assessment in the final experiment (Wagner et al., 2019). On the other hand, Non-intrusive assessments eliminate the need for reference speech by evaluating synthesized speech using machine learning model. The correlation of predicted scores and human scores between system-level assessments might be passable, but utterance-level correlations are low (Wagner et al., 2019). Moreover, it needs a lot of training data to be generalized that might be unavailable for low-resource languages. As a result, objective assessment with a high correlation to subjective assessment is essential in the TTS development process and helps fine-tune modeling and model selection prior to subjective testing.

2.2 Dynamic Time Warping

Most traditional objective speech quality uses DTW to compare the reference speech to synthesized speech. DTW is an algorithm for comparing two time series that differ in time and speed. For example, with voice authentication, a word or a brief sentence from the same user may be faster or slower than another, even if they are the same utterance. As a result, it is impossible to compare it to the linear method. DTW is a non-linear method for finding the optimal match between two time series with limitations. The optimal match is the one that constrains all of the conditions while also being the lowest. The equation of DTW can be written as

Equation of DTW can be written as

$$DTW(X,Y) = d(N,M) \tag{2.1}$$

$$d(i,j) = ||x_i - y_j|| + \min \begin{cases} d(i,j-1) \\ d(i-1,j) \\ d(i-1,j-1) \end{cases} \tag{2.2}$$

where X and Y are time series data, N and M are data point numbers of X and Y, and d is a recursive function that calculates the distance between data points at i and j.

Chapter III

RELATED WORKS

In this chapter, we will describe embedding-based assessment in Sec. 3.1. Then, in Sec. 3.2, we'll examine into ASR Feature study, followed by other speech quality objective measures that are still used. It has been divided into intrusive and non-intrusive assessments in Sec. 3.3 and 3.4, respectively.

3.1 Embedding-Based Metric

The use of discriminative model features to evaluate generative models has attracted researchers' interest. A text-generating machine translation example is BERTScore (Zhang* et al., 2020), which takes contextual embeddings from the BERT model and computes weighted matching using cosine similarity between the generated and reference sentences. In terms of correlation, BERTScore outperforms common metrics and is useful for model selection. Fréchet Inception Distance (FID) (Heusel et al., 2017), is another embedding-based metric used in computer vision to evaluate generated images from GANs. FID takes Inception model features and computes a distance between reference and generated images using Fréchet Distance. For speech synthesis, Fréchet DeepSpeech Distance (FDSD) (Bińkowski et al., 2020) was inspired by FID to calculate the similarity between synthesized and reference speech feature distributions using Fréchet distance as same as FID. The characteristics were retrieved from the DeepSpeech2 (Amodei et al., 2016a) speech recognition model's penultimate layer (before the softmax layer). This research has demonstrated that ASR high-level features can be utilized to evaluate TTS models. However, The objective assessment is not the main finding of their study. As a result, they did not test the influence that each layer of the model might have, and the method's disadvantage is that it can’t be utilized for utterance-level evaluation. .These studies demonstrate that the discriminative model’s features are meaningful and adequate for evaluating a generative model.

3.2 Research on ASR Features

There have been numerous studies that have looked into the meaning of ASR Features. Belinkov and Glass (2017); Li et al. (2020) investigated layer-wise quality features in terms of several classification tasks and found that the features in different layers capture different levels of speech information. The top layers more accurately represent character than phonetic information, while the intermediate layer more accurately represents phonetic information. Li et al. (2020) investigated the reduction of speaker characteristics through the layers by synthesizing speech from ASR features. These studies show that ASR features are a good representation of speech, and the quality of features in each layer is different.

3.3 Intrusive assessments

Mel Cepstral Distortion (MCD) (Kubichek, 1993): is a measure of the difference between two sequences of MFCC. Typically differences in timing are allowed for; this is enabled by either DTW, or by synthesizing test utterances with the ”gold” durations from the original speech. The formula of MCD calculataion is

$$MCD(X,Y) = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t}\sqrt{\sum_{i}^{20} 2||C_x(t,i) - C_y(t,i)||^2} \tag{3.1}$$

where C_x and C_y are cepstral coefficients of reference, X, and synthesized, Y, and T is number of frame.

Mel spectral Distortion (MSD) (Weiss et al., 2021): is very similar to MCD. The only difference is that the 80-channels log-mel magnitude spectrogram is used instead of the MFCC.

Perceptual Evaluation of Speech Quality (PESQ) (Rix et al., 2001): is defined by the ITU-T recommendation P.862. PESQ is a speech quality metric used in telecommunications (Cerňak and Rusko, 2005). It was also used to assess TTS. The speech signals are aligned before being transformed into a Bark Spectrogram. The absolute difference between two speech signals is used to calculate their similarity. The PESQ score can range from –0.5 to 4.5. In most circumstances, the output range will be the same as that of MOS.

Virtual Speech Quality Objective Listener (ViSQOL) (Chinen et al., 2020): measures the similarity between spectrograms with patch alignment. The reference and synthesized speech are segmented into patches. The similarity between the reference and a synthesized patch is calculated frame by frame using the Neurogram Similarity Index Measure (Hines and Harte, 2012) (NSIM) and converted to MOS.

The equation of NSIM is defined as

$$Q(X,Y) = l(X,Y) \cdot s(X,Y) = \frac{2u_x u_y + C_1}{u_x^2 + u_y^2 + C_1} \cdot \frac{\sigma_{xy} + C_2}{\sigma_x \sigma_y + C_2} \tag{3.2}$$

where C_1 and C_2 are constant values that value $0.01L$ and $(0.03L)^2$, and L is the intensity range.

Fréchet DeepSpeech Distance (FDSD) (Bińkowski et al., 2020): is the Fréchet distance between synthesized and reference speech feature distributions as described in Sec. 3.1

3.4 Non-Intrusive assessments

MOSNet (Lo et al., 2019): is a MOS-predictive model trained using the Voice Conversion Challenge (VCC) 2018 Lorenzo-Trueba et al. (2018) dataset. This work has shown success in using deep learning to estimate the quality of converted speech audio. The predicted MOS from MOSNet has a high correlation to humans' MOS at the system level and the utterance level.

NISQA-TTS (Mittag and Möller, 2020): is a MOS-predictive model based on deep learning architecture trained on listening score data from various years of the Blizzard Challenge and the VCC2018 to estimate the quality of synthesized speech.

Chapter IV

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.

4.1 Preprocessing

The raw speech audios were preprocessed to remove aspects that are unrelated to quality assessment. Silences were removed from both ends of the recording by using a simple energy-based method. The inner silences were kept as is to preserve prosody. The loudness of each synthesized utterance was also normalized to be the same as the reference recording.

4.2 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, $\mathbf{y} \in \mathbb{R}^{T_y}$, Spectrograms are computed, $S_x \in \mathbb{R}^{N_x \times L}$ 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.

4.3 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 is rich enough for speech synthesis (Li et al., 2020). By measuring the difference between the features of the synthesized and reference audios, The human-likeness and pronunciation correctness of the synthesized speeches can roughly be estimated.

The high-level features are computed, $h_x \in \mathbb{R}^{P_x \times K}$ 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.

4.4 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. Euclidean distance was used 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: (4.1).

$$\text{SLSRD }(x, y) = \frac{1}{T\sqrt{C}} DTW(S_x \oplus h_x, S_y \oplus h_y) \tag{4.1}$$

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, the lower the better.

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

$$\text{LSRD }(x, y) = \frac{1}{T\sqrt{K}} DTW(h_x, h_y) \tag{4.2}$$

Chapter V

EXPERIMENTAL RESULTS

5.1 Datasets

Blizzard Challenge dataset

We used the test results from the Blizzard Challenge 2012, 2013, and 2019 (King and Karaiskos, 2012, 2014; Wu et al., 2019) that contain audios synthesized using HMM-based, diphone, unit-selection, hybrid TTS, and deep learning models. We used the subset of EH1 test results from the 2012 and 2013 competitions, and all available test results from the 2019 in our experiments and denoted them as BLZ2012, BLZ2013, and BLZ2019.

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 (Shen et al., 2018) and the non-autoregressive FastSpeech (Ren et al., 2019) 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 (Prenger et al., 2019) was used for generating the waveform values from the predicted Mel frequency values.

Thirty-four new sentences were used for two evaluation tasks, MOS and head-to-head comparison. The synthesized audios were scored and compared by 138 Thai native speakers volunteers with the criteria of naturalness and sound quality. To control the quality of the score, we tracked action on the play button and found four people score the speech audios without clicking the play button then we removed them out. Finally, we have evaluation results from 134 Thai native speakers.

For head-to-head comparison, the human raters were presented with 15 random audio pairs from 136 pairs of the same Thai sentence but generated from different models and asked to select the better one or indicate a tie.

For MOS evaluation, we used a straightforward 1–5-point scale. The highest MOS of 5 was given to the goods, and the lowest MOS of 1 was given to the bads. The 30 random synthesized audios were presented to a human rater for scoring. The MOS scores from each model were summarized in Table 5.1.

Table 5.1: 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

5.2 Experimental setup

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 (Kuchaiev et al., 2018). For English, we used the pre-trained model provided in the OpenSeq2Seq[1] library. For Thai, we

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

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 Naowarat et al. (2021). The model achieved 6.76% Character Error Rate (CER). We also trained the model from scratch for Chinese Mandarin with AISHELL-1 (Hui Bu, 2017) dataset that achieved 12.87% Character Error Rate (CER).

As for the DTW scoring, we used FastDTW (Salvador and Chan, 2007) 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.

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 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.

Table 5.2: Correlation values between different objective assessment methods and MOS at utterance-level. The dagger symbol (†) indicates negative correlation coefficients of the MOSNet model.

Method	TH		BLZ2012		BLZ2013		BLZ2019	
	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$
PESQ	0.07	0.12	0.20	0.17	0.02	0.06	0.05	0.07
MCD	0.41	0.25	0.31	0.2	0.17	0.12	0.36	0.20
MSD	0.41	0.25	0.28	0.18	0.24	0.17	0.26	0.10
VISQOL	0.38	0.27	0.49	0.34	0.30	0.21	0.27	0.20
FDSD	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
MOSNet	0.1	0.08	†	†	0.49	0.35	†	†
NISQA-TTS	0.06	0.0	0.62	0.43	*	*	*	*
LSRD (ours)	0.64	0.43	0.53	0.36	0.58	0.40	**0.61**	**0.35**
SLSRD (ours)	**0.64**	**0.44**	**0.67**	**0.48**	**0.58**	**0.40**	0.60	0.34

Table 5.3: Correlation values between different objective assessment methods and MOS at system-level. The dagger symbol (†) indicates negative correlation coefficients of the MOSNet model.

Method	TH		BLZ2012		BLZ2013		BLZ2019	
	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$
PESQ	0.40	0.0	0.37	0.33	0.19	0.18	0.08	0.05
MCD	0.80	0.67	0.37	0.24	0.37	0.17	0.37	0.10
MSD	0.89	0.67	0.40	0.33	0.44	0.33	0.59	0.39
VISQOL	0.77	0.67	0.69	0.56	0.64	0.50	0.52	0.45
FDSD	0.77	**1.0**	0.55	0.53	0.67	0.61	0.65	0.35
MOSNet	†	†	†	†	0.66	0.56	†	†
NISQA-TTS	0.77	0.67	**0.89**	**0.73**	*	*	*	*
LSRD (ours)	**0.91**	**1.0**	0.66	0.64	0.78	0.61	**0.78**	**0.46**
SLSRD (ours)	0.90	**1.0**	0.88	**0.73**	**0.8**	**0.67**	0.76	0.42

5.3 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, BLZ2013, and BLZ2019 datasets. The results are summarized in Table 5.2 and 5.3 for utterance- and system-level, respectively. 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 proposed SLSRD and LSRD had superior utterance-level performance than other metrics by a large margin in all four datasets. Adding the spectral features improves 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 5.2. FDSD is not applicable for utterance-level evaluation as it computes the distance between distributions.

Note that Table 5.2 and 5.3 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 did not generalize well. SLSRD has less correlation with MOS than LSRD in BLZ2019 while gaining a lot improve in BLZ2012. We have investigated more ASR models to compare between LSRD and SLSRD in BLZ2012 and BLZ2019 as shown in Table 5.4 and Table 5.5.

Most ASR models used in this experiment choose from available pre-trained models from several open source to investigate. For English, we use pretrained from Nvidia NeMo (Kuchaiev et al., 2019) for QuartzNet15x5 (Kriman et al., 2020), CitriNet (Majumdar et al., 2021) and Conformer Large (Gulati et al., 2020), pretrained from SpeechBrain (Ravanelli et al., 2021) for Transformer, and pretrained from github[2] for Deepspeech2. For Chinese Mandarin, we use pretrained from library as same as English except for DeepSpeech2 (Amodei et al., 2016b) which is trained from scratch . We also investigated self-supervised for ASR like Wav2Vec2 (Baevski et al., 2020).

As a results in Table 5.4 and 5.5, LSRD is quite better than SLSRD in several models in BLZ2019 and shows that adding spectral features has less effective

[2] https://github.com/SeanNaren/deepspeech.pytorch

Table 5.4: Compare correlation coefficients of different ASR models at utterance-level between SLSRD and LSRD in BLZ2012 and BLZ2019.

Model	BLZ2012				BLZ2019			
	LSRD		SLSRD		LSRD		SLSRD	
	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$
QuartzNet15x5	0.54	0.38	**0.61**	**0.43**	**0.62**	**0.36**	0.58	0.32
CitriNet	0.62	0.45	**0.70**	**0.52**	**0.73**	**0.49**	0.71	0.46
Conformer Large	0.57	0.41	**0.71**	**0.52**	**0.69**	**0.42**	0.68	0.41
Transformer	0.57	0.40	**0.67**	**0.49**	**0.61**	**0.36**	0.59	0.33
DeepSpeech2	**0.25**	0.20	0.23	**0.23**	0.26	0.17	**0.27**	**0.20**
Wav2Vec2	0.57	0.39	**0.57**	**0.39**	**0.64**	**0.40**	0.61	0.38

Table 5.5: Compare correlation coefficients of different ASR models at system-level between SLSRD and LSRD in BLZ2012 and BLZ2019.

Model	BLZ2012				BLZ2019			
	LSRD		SLSRD		LSRD		SLSRD	
	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$	$\|r\|$	$\|\tau\|$
QuartzNet15x5	0.72	0.64	**0.93**	**0.73**	**0.62**	**0.36**	0.58	0.32
CitriNet	0.88	0.69	**0.96**	**0.78**	**0.81**	**0.55**	0.80	0.52
Conformer Large	0.77	0.64	**0.92**	**0.78**	**0.81**	**0.56**	0.81	0.43
Transformer	0.71	0.56	**0.85**	**0.64**	**0.61**	**0.36**	0.59	0.33
DeepSpeech2	**0.28**	0.24	0.28	**0.29**	0.26	0.17	**0.27**	**0.20**
Wav2Vec2	**0.76**	0.51	0.74	**0.63**	**0.85**	**0.60**	0.81	0.54

when evaluating high quality of synthesized speech audios. BLZ2019 dataset contains many Deep Learning-based models that generate speech quality higher than HMM used in BLZ2012 as we can see in MOS density in Figure 5.1. MOS density of BLZ2012 is right-skewed distribution and BLZ2019 is a left-skewed distribution which means most synthesis speech audio in BLZ2019 has MOS higher than BLZ2012. SLSRD demonstrates the simple yet effective use of spectral and latent ASR features for evaluating TTS models and is more effective to artifacts than LSRD. ASR is usually trained using data augmentation strategies for model generalization which leads to using latent ASR features only less effective to evaluate low quality speech audios. 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.

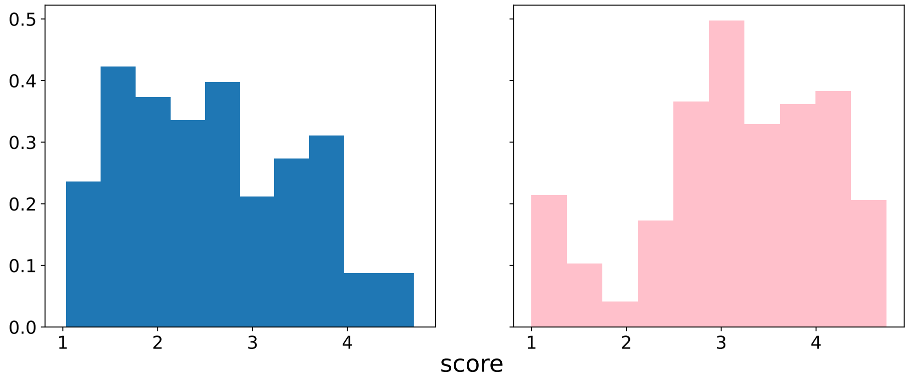

Figure 5.1: Density of MOS on BLZ2012 (left) and BLZ2019 (right)

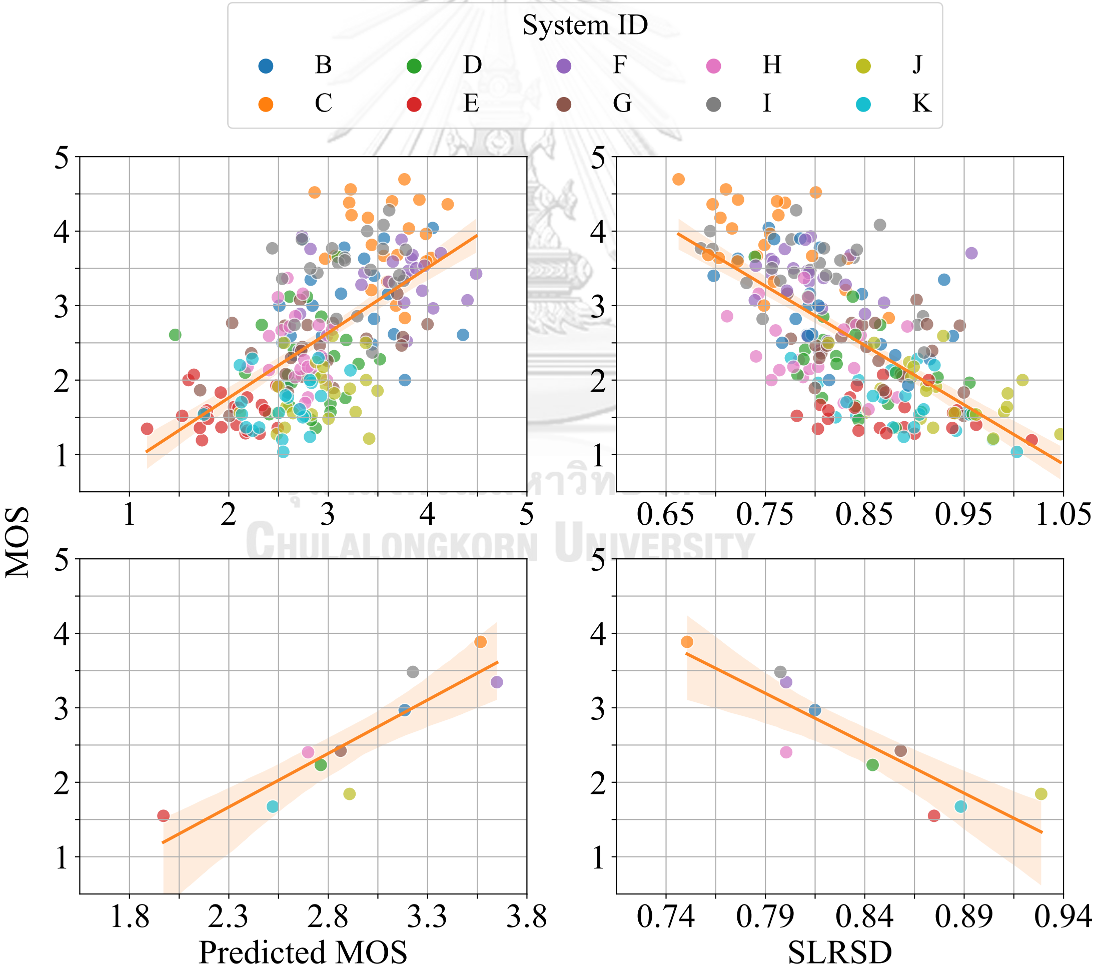

Figure 5.2: Regression analysis for utterance-level (top) and system-level (bottom) of NIQSA-TTS (left) and SLSRD (right) on BLZ2012. Translucent bands represent the 95% confidence interval for the regression estimate.

5.4 Ablation study on the layer location to extract the latent feature

In this section, we investigated how changing the choice of the hidden layer from an ASR model affects the correlation coefficient in the Thai dataset. As pointed out by previous works (Belinkov and Glass, 2017; Belinkov et al., 2019; Li et al., 2020), 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 5.6 shows that outputs from ReLU_9, a layer in the middle of the entire network, yield the highest correlation coefficient, contrary to Bińkowski et al. (2020) which uses the penultimate layer. A similar trend was also found in other ASR models in Table 5.4 and 5.5. All of the highest correlation coefficient was yielded from an intermediate layer. According to Belinkov and Glass (2017) and Belinkov et al. (2019), the intermediate layers better capture the phonetic properties than the top layers. Li et al. (2020) 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.

5.5 Head-to-head comparison between synthesized audios

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

Table 5.6: 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

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

SLSRD outperformed the best baseline by 9.58% absolute as illustrated in Table 5.7. Figure 5.3 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.

Table 5.7: Agreement rates between objective measures and human on the head-to-head comparison between synthesized audios

Method	Human agreement rate (%)
MCD	61.03
MSD	65.95
LSRD	67.02
SLSRD	**75.53**

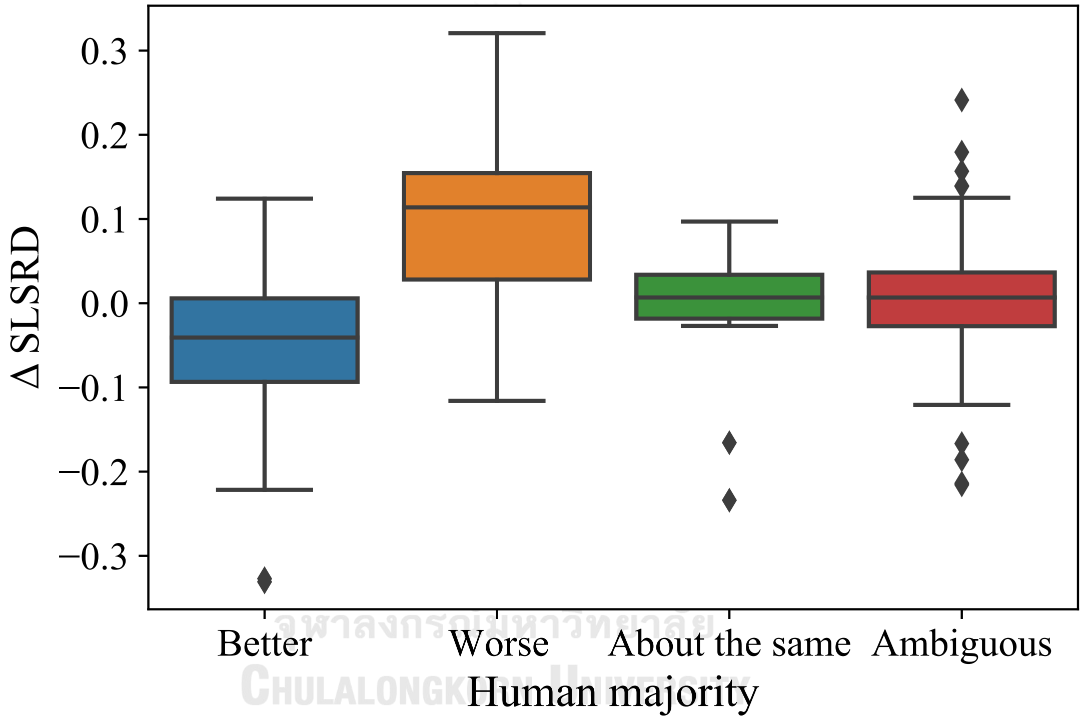

Figure 5.3: Boxplot of the difference between SLSRD scores for each kind of audio pair in the head-to-head experiment.

Chapter VI

CONCLUSION

We proposed intrusive quality assessments, SLSRD and LSRD, for TTS evaluation. LSRD is the DTW distance between the reference and the synthesized audios using ASR features while SLSRD uses spectrogram tandem. Experiments on the Thai and Blizzard Challenge datasets showed that SLSRD and LSRD outperformed other objective measures 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.

REFERENCES

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L. V., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. 2016a. Deep speech 2: End-to-end speech recognition in english and mandarin. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, p. 173–182. : JMLR.org.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L. V., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng, A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. 2016b. Deep speech 2: End-to-end speech recognition in english and mandarin. In Proceedings of the

33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16, p. 173–182. : JMLR.org.

Avila, A. R., Gamper, H., Reddy, C., Cutler, R., Tashev, I., and Gehrke, J. 2019. Non-intrusive speech quality assessment using neural networks. In ICASSP 2019, pp. 631–635. :

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations [Online]. Available from: https://arxiv.org/abs/2006.11477 [2020,].

Belinkov, Y. and Glass, J. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. In Advances in Neural Information Processing Systems (NIPS). :

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

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

Cerňak, M. and Rusko, M. 2005. An evaluation of a synthetic speech using the PESQ measure. In Proceedings of Forum Acusticum 2005. :

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

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

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In Proc. Interspeech 2020, pp. 5036–5040. :

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (ed.), Advances in Neural Information Processing Systems, volume 30. : Curran Associates, Inc.

Hines, A. and Harte, N. 2012. Speech intelligibility prediction using a Neurogram Similarity Index Measure. Speech Communication 54 (02 2012): 306–320.

Hui Bu, X. N. B. W. H. Z., Jiayu Du. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In Oriental COCOSDA 2017, p. Submitted. :

King, S. and Karaiskos, V. 2012. The Blizzard Challenge 2012. :

King, S. and Karaiskos, V. 2014. The Blizzard Challenge 2013. :

Kongthaworn, T., Naowarat, B., and Chuangsuwanich, E. 2021. Spectral and Latent Speech Representation Distortion for TTS Evaluation. In Proc. Interspeech 2021, pp. 2741–2745. :

Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., and Zhang, Y. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In ICASSP 2020

- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6124–6128. :

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. In Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, volume 1, pp. 125–128 vol.1. :

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

Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., et al. 2019. Nemo: a toolkit for building ai applications using neural modules. arXiv preprint arXiv:1909.09577 (2019):

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

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

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., and Ling, Z. 2018. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. In Proc. Odyssey 2018 The Speaker and Language Recognition Workshop, pp. 195–202. :

Majumdar, S., Balam, J., Hrinchuk, O., Lavrukhin, V., Noroozi, V., and Ginsburg, B. 2021. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition [Online]. Available from: https://arxiv.org/abs/2104.01721 [2021,].

Mittag, G. and Möller, S. 2020. Deep Learning Based Assessment of Synthetic Speech Naturalness. In Interspeech 2020, pp. 1748–1752. :

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

Prenger, R., Valle, R., and Catanzaro, B. 2019. Waveglow: A Flow-based Generative Network for Speech Synthesis. In ICASSP 2019, pp. 3617–3621. :

Qiang Fu, Kechu Yi, and Mingui Sun. 2000. Speech quality objective assessment using neural network. In ICASSP 2000, volume 3, pp. 1511–1514 vol.3. :

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

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. 2021. SpeechBrain: A general-purpose speech toolkit.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. 2019. FastSpeech: Fast, Robust and Controllable Text to Speech. In Wallach,

H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (ed.), Annual Conference on Neural Information Processing Systems 2019, pp. 3165–3174. :

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

Sailor, H. B. and Patil, H. A. 2014. 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 (2014): 521–525.

Sakoe, H. and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26.1 (1978): 43–49.

Sakoe, H. and Chiba, S. 1971. A dynamic programming approach to continuous speech recognition. In Proceedings of the Seventh International Congress on Acoustics, Budapest, volume 3, pp. 65–69. Budapest: Akadémiai Kiadó.

Salvador, S. and Chan, P. 2007. Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis 11.5 (2007): 561–580.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In ICASSP 2018, pp. 4779–4783. :

Soni, M. H. and Patil, H. A. 2016. Non-intrusive quality assessment of synthesized speech using spectral features and support vector regression. In 9th ISCA Speech Synthesis Workshop, pp. 127–133. :

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

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Henter, G. E., Maguer, S. L., Malisz, Z., Székely, É., Tånnander, C., and Voße, J. 2019. Speech synthesis evaluation — state-of-the-art assessment and suggestion for a novel research program. 10th ISCA Workshop on Speech Synthesis (SSW 10) (2019):

Wang, S., Sekey, A., and Gersho, A. 1992. An objective measure for predicting subjective quality of speech coders. IEEE Journal on Selected Areas in Communications 10.5 (1992): 819–829.

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

Wu, Z., Xie, Z., and King, S. 2019. The blizzard challenge 2019. :

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. 2020. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations. :

Biography

Thananchai Kongthaworn was born in Bangkok on November, 1994. He graduated from Yannawate Wittayakom school and then went to King Mongkut's University of Technology North Bangkok where he received B.Eng in computer engineering. His field of interest includes various topics in speech processing and natural language processing.