Vietnamese Text-To-Speech Shared Task VLSP 2020: Remaining problems with state-of-the-art techniques

NGUYEN Thi Thu Trang[1,2], NGUYEN Hoang Ky[2], PHAM Quang Minh[2] and VU Duy Manh[2]

[1]*School of Information and Communication Technology*
Hanoi University of Science and Technology
Hanoi, Vietnam
trangntt@soict.hust.edu.vn, trangntt@vbee.vn

[2]*R&D Lab*
Vbee Services and Data Processing Solution Jsc.
Hanoi, Vietnam
kynh@vbee.vn, minhpq@vbee.vn, manhvd@vbee.vn

***Abstract*— The VLSP 2020 is the seventh annual international workshop whose campaign was organized at the Hanoi University of Science and Technology (HUST). This was the third time we organized the Text-To-Speech shared task. In order to better understand different speech synthesis techniques on a common Vietnamese dataset, we conducted a challenge that helps us better compare research techniques in building corpus-based speech synthesizers. Participants were provided with a single training dataset including utterances and their corresponding texts. There are 7,770 utterances of a female Southwest professional speaker (about 9.5 hours). There is a total of 59 teams registered to participate in this shared task, and finally, 7 participants were evaluated online with perceptual tests. The best synthetic voice with Tacotron 2 and Hifigan vocoder with Waveglow denoiser achieved 89.3% compared to the human voice in terms of naturalness, i.e. 3.77 over 4.22 points on a 5-point MOS scale). Some reasons for a quite-big gap between the best synthetic voice with state-of-the-art synthetic techniques and the human voice were: (i) improper prosodic phrasing for long sentences and (ii) wrong/bad pronunciation for loan words.**

Keywords—VLSP Campaign 2020, TTS shared task, speech synthesis, text-to-speech, evaluation, perception test, Vietnamese

I. Introduction

VLSP stands for Vietnamese Language and Speech Processing Consortium. It is an initiative to establish a community working on speech and text processing for the Vietnamese language [2]. The VLSP 2020 was the sixth annual international workshop. The Text-To-Speech (TTS) shared task was a challenge in the VLSP Campaign 2020, which was organized at Hanoi University of Science and Technology. This was the third time we organized the challenge in speech synthesis.

To the best of our knowledge, Vietnamese TTS systems can be divided into three main types:(i) Hidden Markov Model (HMM) based systems, (ii) Deep Neural Network (DNN) based systems, and (iii) state-of-the-art end-to-end systems. HMM-based TTS systems [6][10] and DNN-based TTS systems [4][9] need to provide pause position and loanword pronunciation in the text pre-processing step. Some end-to-end TTS systems, such as Tacotron [3][11], could use a massive amount of text and audio data pairs to learn prosody and loanword modeling directly from the TTS training process. Nevertheless, corpora do not always design to support that purpose.

This shared task has been designed for understanding and figuring out remaining problems in Vietnamese TTS with state-of-the-art speech synthesis techniques on the same dataset. Based on some subjective feedback from listeners of the last year's TTS shared task, three main problems have been raising for this year: prosodic phrasing (mainly focusing on pause detection) [5], text normalization (mainly focusing on loanwords) [6] [8], and removing noise for Internet datasets.

Participants took the released speech dataset, build a synthetic voice from the data and submit the TTS system. We then synthesized a prescribed set of test sentences using each submitted TTS system. The synthesized utterances were then imported to an online evaluation system. Some perception tests were carried out to rank the synthesizers focusing on evaluating the intelligibility and the naturalness of participants' synthetic utterances.

The rest of this paper is organized as follows. Section II presents the common dataset and its preparation. Section III introduces participants and a complete process of the TTS shared task in VLSP Campaign 2020. We then show the evaluation design and experimental results in Section IV. We finally conclude the task and give some possible ideas for the next challenge in Section V.

II. Common Dataset

The topic of this shared task is to address remaining problems of TTS systems using state-of-the-art synthesis techniques. Based on some analyses on the previous task results, aforementioned, we raised the following issues for this shared task: (i) prosodic phrasing (focusing on pause detection for long input sentences), (ii) text normalization (focusing on expanding loanwords), and (iii) removing background noises (of Internet audios).

Due to the topic of this year's task, we decided to collect audiobooks from the Internet. Vbee Jsc supported to build the dataset for this task. The corpus was taken from a novel called "Bell to Whom the Soul" by Hemingway, a famous American novelist. Audio stories were downloaded manually, divided into 28 long audio files, each had 30 to 60 minutes in length. These files were then automatically split into smaller audio files that are less than 10 seconds in length (using Praat scripting tool). After this process, the number of sound files was up to nearly 20,000 sound files with different lengths.

However, approximately 10,000 sound files that were too short in length (i.e. less than 750 ms) were discarded. Next, we used the ASR API of Vais Jsc to convert the remaining 10,000 audio files into text. These data were checked by the teams participating in the contest. Each team only had to check xxx files for participation. Finally, 7,770 best quality utterances and their corresponding texts were selected as the final dataset. Even though the speaker's voice was professional and pretty, the voice still contained some background noise due to the recording device's low quality.

III. PARTICIPANTS

For TTS shared task this year, participants had to follow a complete process (Fig. 1), which was managed in the website of the TTS shared task of VLSP Campain 2020 (https://tts.vlsp.org.vn).

First, each team registered to participate in the challenge. They were then provided with accounts to log into. On this site, all teams were asked to check the audio files to see if they match the corresponding text and edit if necessary. If they found that the text was exactly the content of the audio, they voted for that transcription. Each audio file needs to be checked by at least 3 teams. Audio files that had no vote after the validation process, we had to check them manually. The participants who completed the required task were asked to send their user license agreement with valid signatures. They were then able to download the training dataset. The dataset includes utterances and their corresponding texts in a text file.

Participants were asked to build only one synthetic voice from the released database. All teams had 20 days for training and optimizing their voices. Each team then submitted the result with a TTS API following the announced specification requirement. We also supported teams that could not deploy their TTS systems to a public server by accepting their docker images that contain the TTS API.

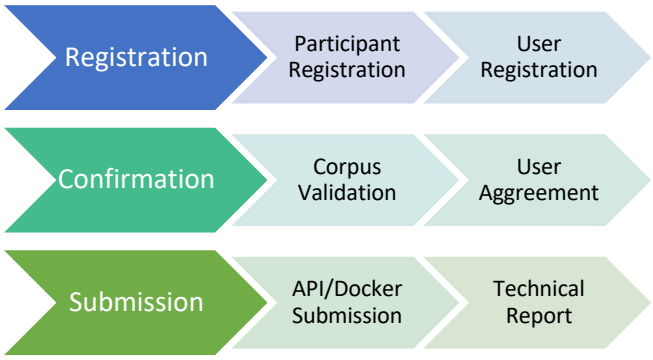

Fig. 1. A complete process for participating TTS shared task VLSP 2020.

We then synthesized audio files from the text files in the test dataset using teams' TTS API. Synthesized files will be evaluated. After receiving evaluation results, the teams proceed to write and submit technical reports.

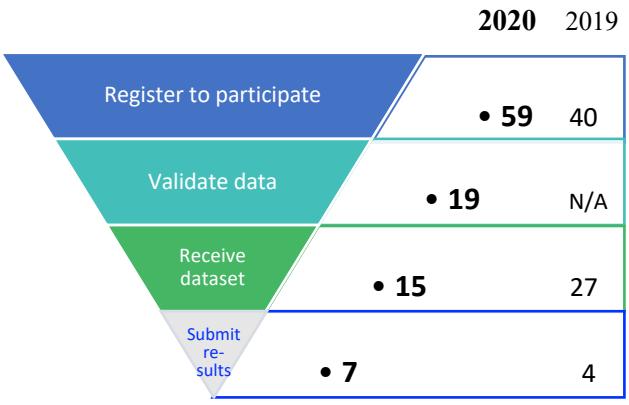

Fig. 2. Participants in VLSP TTS 2020 and 2019.

Fig. 2 compares the number of participants of last year to this year. Fifty-nine teams registered for this year's challenge. Unlike last year, participants were asked to validate the provided dataset, and 19 joined the data validation process, and 15 teams obtained the data after sending the signed user agreement. Finally, nine teams, compared to four in 2019, submitted their TTS system. We synthesized testing audio through the TTS API of each team. Unfortunately, we could not use the TTS API of the two teams due to problems with their TTS system or their server. Table I gives the list of participants that had final submissions to the VLSP TTS shared task 2020.

TABLE I. LIST OF TEAMS PARTICIPATING IN VSLP TTS 2020

No	Team ID	Affiliation	Submission
1	Team1	Unaffiliated	API (error)
2	Team2	Smartcall	API
3	Team3	Unaffiliated	Docker image (error)
4	Team4	Viettel Telecom	API
5	Team5	IC	IC
6	Team6	VAIS	API
7	Team7	Falcon	API
8	Team8	Sun Asterisk Inc.	Docker Image
9	Team9	UET	Docker Image

IV. EVALUATION

Perceptual testing was chosen for evaluating synthetic voices. First, an intelligibility test was conducted to measure the understandability, then the MOS test, which allowed us to score and compare the global quality of TTS systems with respect to natural speech references. All subjects conducted the online evaluation via a web application. This online evaluation system was built by the School of Information and Communication Technology, Hanoi University of Science and Technology, and Vbee Jsc. This system was integrated into https://tts.vlsp.or.vn.

They first registered on the website with necessary information including their hometowns, ages, genders, occupations. They were trained on how to use the website and how to conduct a good test. They were strictly asked to do the test in a controlled listening condition (i.e. headphones and in a quiet distraction-free environment). To ensure that the subjects focused on the test, we designed several sub-tests for each test due to a big number of testing voices (i.e. 8 voices including natural speech). As a result, each sub-test lasts from 25 to 30 minutes.

On completion of any sub-test, or after logging in again, a progress page showed listeners how much they had completed. Detailed instructions for each sub-test were only shown on the page with the first part of each sub-test; subsequent parts had briefer instructions in order to achieve a simple layout and a focussed presentation of the task.

In order to address the issue of duplicate contents of stimuli, we adopted the Latin square (nxn) [1] for all sub-tests, where n is a number of voices in the sub-test. To be more specific, each subject listened to one n^{th} of the utterances per voice, without any duplicate content. With the Latin square design, the number of subjects should be at least twice more than the ones with the normal design.

Stimuli were randomly and separately presented only once to subjects. Each stimulus was an output speech of a TTS system or a natural speech for a sentence. Details of the two tests are described in the following subsections.

A. *Intelligibility Test*

In the intelligibility test, subjects were asked to write down the text of the audio they heard (Fig. 3). The subjects might listen again a second time if they do not hear clearly or have long sentences. They only listened to the utterances the third time when the subjects were distracting, or the sentence were very long.

TABLE II. DESIGN FOR INTELLIGIBILITY SUB-TESTS

Sub-test 1	**Sub-test 2**	**Sub-test 3**
IntelligibilityTest-1	IntelligibilityTest-2	IntelligibilityTest-3
Team 7 Team 8 Team 9	NATURAL Team 5 Team 6	NATURAL Team 2 Team 4

There are three sub-tests in the intelligibility test, following the Latin Square design aforementioned. In each sub-test, there were 3 voices of 3 different teams with or without the natural speech reference (NATURAL). Details for each sub-test is presented in Table II. Each sub-test included voices of two (sub-test 2 and sub-test 3) or three teams (sub-test 1). The natural speech was put in both sub-test 2 and sub-test 3 for more reference. As a result, each sub-test had a total of 3 voices.

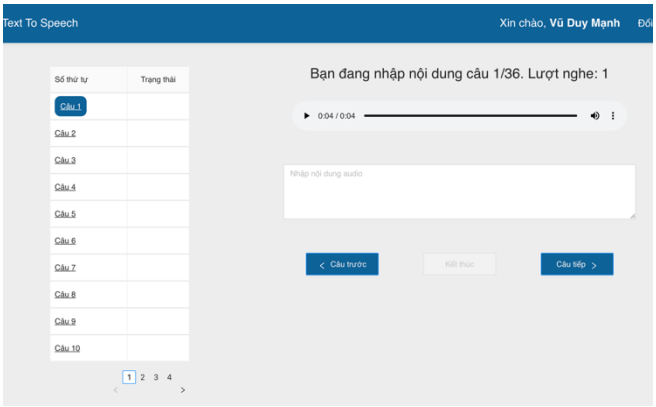

Fig. 3. Online Tool for Intelligibility Test.

Twenty-seven subjects participated in this test. There were two main types of subjects who participated in the test: (i) 19 students (19-22 years-old, 10 females) from Hanoi University of Science and Technology, VNU University of Science; (ii) 8 speech experts (23-38 years-old, 4 female).

The testing dataset included 36 sentences. Each subject needs to participate in at least two of the three sub-tests.

B. *MOS Test*

Subjects (i.e. listeners) were asked to assess by giving scores to the speech they had heard (Fig. 4). When taking this test, subjects listen to the voice once, unless they do not hear it clearly, then listen for a second time.

Subjects randomly listened to utterances and then gave their scores for the naturalness of the utterances. The question presented to subjects was "How do you rate the naturalness of the sound you have just heard?". Subjects could choose one of the following five options (5-scale):

- 5: Excellent, very natural (human speech)
- 4: Good, natural
- 3: Fair, rather natural
- 2: Poor, rather unnatural (rather robotic)
- 1: Bad, very unnatural (robotic).

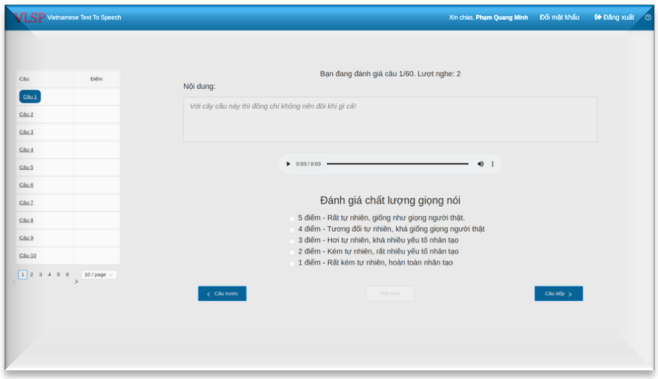

Fig. 4. Online Tool for MOS Test.

Testing text set includes 60 sentences. There are two sub-tests, including 60 random utterances each (taken from 480 utterances). Table III illustrates the design for the two MOS sub-tests. We put the natural speech (NATURAL) as a reference in both sub-tests. Due to an odd number of final participated teams, sub-test 1 included 3 teams (Team 2,4,5) while sub-test 2 had voices from the remaining 4 teams (Team 6,7,8,9).

TABLE III. DESIGN FOR MOS TEST SUB-TESTS

Sub-test 1	**Sub-test 2**
MOS Test 1	MOS Test 2
NATURAL Team 2 Team 4 Team 5	NATURAL Team 6 Team 7 Team 8 Team 9

Subjects participated in two sub-tests for voices built from the common dataset. Due to a rather big number of voices in each sub-test (i.e. 5 including the natural reference), we let the subjects to heard randomly half of the utterances for each voice. The number of subjects who listened to each sub-test was 48 (20 females). Each subject needs to participate in all two sub-tests, estimated at 25 to 30 minutes.

V. EVALUATION RESULTS

A. *Intelligibility Score*

Due to a large number of loanwords in the test set, the intelligibility results were not good, at about 68-89% at both word and syllable levels, even with natural speech. The subjects might do not know how to write these loanwords or present different orthography from the original text. We should have a special design and more analyses for this type of test in the future.

B. *MOS Score*

The perceptual evaluation of the general naturalness was carried out on different voices of participants and a natural speech reference (NATURAL) of the same speaker as the

training corpus. Fig. 5 and Table IV show the final MOS test results. Only three teams submitted technical reports, i.e. Team2, Team6, and Team7.

We can see that Team2 was the best team (i.e. 3.769) – about 89.3% compared to the natural speech (i.e. 4.220/5). This team adopted Tacotron-2 as the acoustic model, and HiFi-GAN as a real-time vocoder, and Waveglow as a denoiser. Team7 was the second place with a 3.698 score (only less than the first place 0.07 point). This team used FastSpeech and PostNet, which could be considered as a faster acoustic model, compared to Tacotron-2 or only FastSpeech. Team6 was the fifth place with a 3.313 score. Their acoustic model was Tacotron2, and their vocoder was Waveglow.

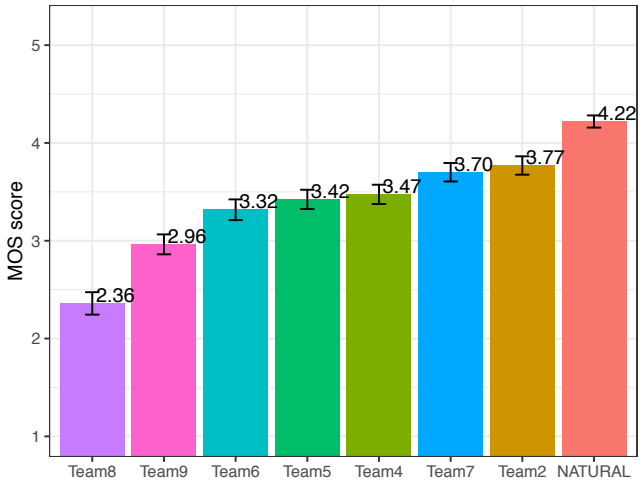

Fig. 5. MOS Test Final Results.

TABLE IV. MOS TEST RESULTS WITH SYNTHESIS TECHNIQUES

Testing voice	MOS Score (5-scale)	Synthesis Techniques
NATURAL	4.220	
Team2	**3.769**	• Acoustic model: Tacotron 2; Vocoder: HiFi-GAN; • Denoiser: Waveglow
Team7	3.698	• Acoustic model: FastSpeech + PostNet; • Vocoder: Waveglow
Team6	3.313	• Acoustic model: Tacotron 2; • Vocoder: Waveglow

Although using state-of-the-art synthesis techniques that lead to a high-quality synthetic voice, there were still some remaining problems in the results of participants. Some reasons were found for a quite-big gap between the best synthetic voice with state-of-the-art synthetic techniques and the human voice: (i) improper prosodic phrasing for long sentences and (ii) wrong/bad pronunciation for loan words.

C. Analysis and Discussion

Several two-factorial ANOVAs were run on the MOS results, illustrated in Table V. The two factors were the TTS system (8 levels) and the Sentence (60 levels) or the Subject (48 levels). All factors and their interactions in both ANOVAs had significant effect ($p<0.0001$).

The TTS system factor alone explained an important part of the variance over levels of both Sentence (29%) and Subject factors (30%). The Sentence factor explained only about 8% of the variance (partial $\eta^2 = 0.08$) while the Subject did 19% (partial $\eta^2 = 0.19$). The interaction between the System and Sentence or Subject explained a quite important part of the variance, i.e. 21% and 14% respectively.

TABLE V. ANOVA RESULTS OF MOS TEST

Factor	df	df error	F	p	η^2
System	7	5,688	335.38	0.0000	0.29
Sentence	59	5,688	8.71	0.0000	0.08
System:Sentence	412	5,688	3.57	0.0000	0.21
System	7	5,798	353.37	0.0000	0.30
Subject	47	5,798	29.29	0.0000	0.19
System:Subject	314	5,798	2.89	0.0000	0.14

We did observe the sentences with bad scores and found that they were long sentences or had a number of loanwords. Synthetic utterances having consecutive loanwords are extremely bad intelligible. These problems led to bad scores for both Intelligibility and MOS Test.

VI. CONCLUSIONS

We did some valuable experiments on TTS systems from different participants using a common dataset in the TTS shared task in the VLSP Campaign 2020. Participants had to validate a piece of training data before receiving the common dataset. There are 7,770 utterances of a female Southwest professional speaker (about 9.5 hours) in the released training dataset. Although using state-of-the-art synthesis techniques that lead to a high-quality synthetic voice, there were still some remaining problems in the results of participants. The best synthetic voice with Tacotron 2 and Hifigan vocoder with Waveglow denoiser achieved 89.3% compared to the human voice, i.e. 3.77 over 4.22 point on a 5-point MOS scale). Some reasons were found for a quite-big gap between the best synthetic voice with state-of-the-art synthetic techniques and the human voice: (i) improper prosodic phrasing for long sentences and (ii) wrong/bad pronunciation for loan words. For the next speech synthesis task of the VLSP Campaign in 2021, we may have more advanced topics for Vietnamese speech synthesis, such as speaker adaptation or expressive speech synthesis.

ACKNOWLEDGMENT

The VLSP 2020 TTS shared task was mainly supported by the R&D Lab, Vbee Services and Data Processing Solution Jsc, and School of Information and Communication Technology. They supported this shared task in developing, deploying, and conducting the online evaluation, based on perception tests as well as building the dataset for the challenge. This task was funded by the Vingroup Innovation Foundation (VINIF) under the project code DA116_14062019 / year 2019. We would like to thank Vais Jsc. for their ASR in building the dataset, and last but not least, the subjects who gave time and effort for the experiments.

REFERENCES

[1] Cochran William G. and Cox Gertrude M. "*Experimental Designs, 2nd Edition*". Wiley, 2 edition, April 1992. ISBN 0471545678.

[2] Luong Chi Mai. "*Special Issue in VLSP 2018*". Journal of Computer Science and Cybernetics, V.34, N.4 (2018).

[3] Shen, J., Pang, R., Weiss, R.J., et al. 2017. "*Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*". 2018 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

[4] H. Ze, A. Senior, and M. Schuster, *"Statistical Parametric Speech Synthesis using Deep Neural Networks"* on 2013 IEEE International Conference on Acoustics, Speech and Signal Processing 2013, pp. 7962–7966. IEEE.

[5] Nguyen Thi Thu Trang, Albert Rilliard, Tran Do Dat, and Christophe d'Alessandro, *"Prosodic Phrasing Modeling for Vietnamese TTS using Syntactic Information"* in 15th Annual Conference of the International Speech Communication Association. Singapore. 2014.

[6] Nguyen Thi Thu Trang, Alessandro Christophe, Rilliard Albert, and Tran Do Dat. *"HMM-based TTS for Hanoi Vietnamese: Issues in Design and Evaluation*" in 14th Annual Conference of the International Speech Communication Association (Interspeech 2013), pages 2311-2315. Lyon, France, August 2013b. ISCA.

[7] Nguyen Thi Thu Trang, Pham Thi Thanh, and Tran Do Dat. *"A method for Vietnamese Text Normalization to Improve the Quality of Speech Synthesis"* in Proceedings of the 2010 Symposium on Information and Communication (SoICT 2010), Hanoi, Vietnam. 2010.

[8] Nguyen Thi Thu Trang, Dang Xuan Bach, and Nguyen Xuan Tung. *"A Hybrid Method for Vietnamese Text Normalization"* in Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval (NLPIR 2019). Japan. 2019.

[9] Nguyen Van Thinh, Nguyen Quoc Bao, Phan Huy Kinh, Do Van Hai, *Development of Vietnamese Speech Synthesis System using Deep Neural Networks*, Journal of Computer Science and Cybernetics, V.34, N.4 (2018), 349-363.

[10] Vu Thang Tat, Luong Mai Chi, and Nakamura S. *"An HMM-based Vietnamese Speech Synthesis System"* in Proceedings of the Oriental COCOSDA International Conference on Speech Database and Assessments, pages 116–121, Beijing, China, 2009.

[11] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, YonghuiWu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, YingXiao, Zhifeng Chen, Samy Bengio, et al. *"Tacotron:Towards End-to-end Speech Synthesis"* in 18th Annual Conference of the International Speech Communication Association. Sweden. 2017.