Text-To-Speech Shared Task in VLSP Campaign 2019: Evaluating Vietnamese speech synthesis on common datasets

NGUYEN Thi Thu Trang[1,2] and NGUYEN Xuan Tung[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
tungnx@vbee.vn

***Abstract*—The VLSP 2019 is the sixth annual international workshop in conjunction with the 2019 Conference of the Pacific Association for Computational Linguistics (PACLING 2019). Its campaign was organized at the VNU University of Science, with support from the other members of the VLSP 2019 committee. This was the second 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 two training datasets; each included utterances and their corresponding texts in a text file: (i) A small training dataset: 821 utterances of a female South professional speaker (about 45 minutes), (ii) A big training dataset: 13,462 utterances of a female North non-professional speaker (about 22 hours).**

***Keywords*—VLSP Campaign, speech synthesis, text-to-speech, evaluation, perception test**

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 Vietnamese language [1]. The VLSP 2019 was the sixth annual international workshop in conjunction with the 2019 Conference of the Pacific Association for Computational Linguistics (PACLING 2019).

The Text-To-Speech (TTS) shared task was a challenge in the VLSP Campaign 2019, which was organized at the VNU University of Science, with support from the other members of the VLSP 2019 committee. This was the second time we organized the challenge in speech synthesis. This challenge has been designed for understanding and comparing research techniques in building Vietnamese corpus-based speech synthesizers on the same data. Participants take the released speech database, build a synthetic voice from the data and synthesize a prescribed set of test sentences. An online evaluation the submitted synthesised utterances focused on naturalness, based on perception tests, was then carried out to try to rank the synthesizers and help identify the effectiveness of the techniques.

In this paper, we summarise TTS shared task in VLSP Campaign 2019: Participants (Section II), Common datasets (Section III), Evaluation design (Section IV) and Results (Section V) and consider possible designs for the next challenge.

II. Participants

There are 40 teams that registered for this year's challenge and 27 ones that obtained the data after sending the signed user agreement. Finally, only 4 submitted entries as listed in Table 1 alongside human speech.

The three teams from Zalo, Sun and VNGGRD chose the state-of-the-art synthesis technique, i.e. Tacotron 2 [2] while VTCC adopted the DNN model [3] as their participation in the previous campaign [4].

TABLE I. The participating teams and their short names

Short name	Detail	Method
NATURAL	Natural speech from the same speaker as the corpus	Human
ZALO	Zalo Group VNG Corporation	Tacotron 2
VTCC	Viettel CyberSpace Center	DNN
SUN	R&D Lab, Sun* Inc.	Tacotron 2
VNGGRD	IoT Department VNG Corporation	Tacotron 2

III. Common Datasets

The data for voice building was provided by Vbee Jsc (small corpus) and InfoRe Jsc (big corpus). Participants who had signed a user agreement were able to download these two datasets. Each dataset includes utterances and their corresponding texts in a text file. Participants were asked to build two synthetic voices from the database.

A. Big corpus

The final big training dataset that sent to participators included about 22 hours with 13,462 utterances of a female North non-professional speaker. The purpose of this corpus is to allow participators experiment with state-of-the-art synthesis techniques such as Tacotron.

After looking at the dataset from InfoRe (about 25 hours), we found that this dataset was not recorded in a studio, but in different environments. Besides, the speaker's voice was not a professional and pretty one. We had to remove unacceptable audios with the aid of an ASR and some processing and analyses on Vietnamese phonetic and phonology.

B. Small corpus

The small training dataset composed of 821 utterances of a female South professional speaker (about 45 minutes), which was provided by Vbee Jsc. This dataset was recorded in a professional studio and by a South TV broadcaster. The quality of the dataset is pretty good in terms of voice, environment recording and phonetically rich corpus. The purpose of this corpus is to allow participators can build compact TTS systems on devices with traditional synthesis techniques.

IV. Evaluation

The evaluation was conducted online, based on perceptual testing. The MOS test was chosen for the evaluation, which allowed us to score and compare the global quality of TTS systems with respect to natural speech references. Subjects (i.e. listeners) were asked to assess by giving scores to the speech they had heard.

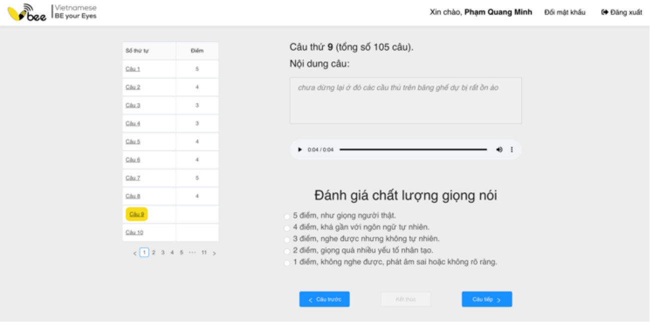
A. *Testing datasets*

82 test sentences were chosen from two sources and kept out from the training big corpus:

- 44 testing sentences were chosen from the big corpus so that the natural references are recorded by the same speaker and environment. This condition made the test scenarios not really various although we tried to find out all the richest phonetic ones.
- 38 testing sentences from Vbee Jsc with various contexts and rich phonetical features.

B. *Subjects*

Two main types of subjects were used: (i) 18 students (19-22 years-old, 8 female) from Hanoi University of Science and Technology, VNU University of Science and University of Science and Technology of Hanoi; (ii) 10 speech experts (23-34 years-old, 4 female). All subjects conducted the online evaluation via a web application [1] (Fig. 1). They first registered in the website with necessary information including their hometowns, ages, genders, occupation. They were trained to 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 quiet distraction-free environment).



Số thứ tự	Điểm
Câu 1	5
Câu 2	4
Câu 3	3
Câu 4	3
Câu 5	4
Câu 6	4
Câu 7	5
Câu 8	4
Câu 9	
Câu 10	

Câu thứ **9** (tổng số 105 câu).

Nội dung câu:

chưa dừng lại ở đó các cầu thủ trên băng ghế dự bị rất ồn ào

0:04 / 0:04

Đánh giá chất lượng giọng nói

○ 5 điểm, như giọng người thật.
○ 4 điểm, khá gần với ngôn ngữ tự nhiên.
○ 3 điểm, nghe được nhưng không tự nhiên.
○ 2 điểm, giọng quá nhiều yếu tố nhân tạo.
○ 1 điểm, không nghe được, phát âm sai hoặc không rõ ràng.

Fig. 1. Main screen of the online evaluation system.

C. *Listening task*

Subjects randomly listened to utterances and then gived 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 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).

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. The numbers of syllables of sentences ranged from 4 to 30, hence the speech samples lasted between 3 and 24 seconds.

Subjects participated 2 sections for voices built from Big training corpus and the ones built from Small training corpus. Due to a rather big number of voices in each section (i.e. 5 including the natural reference), we let the subjects to heard randomly a half of utterances for each voices. The number of subjects listened to each section was 24 (about 10 female).

Listeners were encouraged to do the evaluation in a single session, estimated at 25 minutes, but the evaluation could be done in multiple sessions if desired. On completion of any section, or after logging in again, a progress page showed listeners how much they had completed. Detailed instructions for each section were only shown on the page with the first part of each section; subsequent parts had briefer instructions in order to achieve a simple layout and a focussed presentation of the task.

V. Evaluation Results

A. *MOS Score*

The perceptual evaluations of the general naturalness were carried out on different voices of participants and a natural speech reference of the same speaker as the training corpus. Table II shows the MOS test results of 4 teams for both voices that built from Big corpus and Small corpus.

TABLE II. MOS TEST RESULTS FOR THE NATURALNESS

System	MOS Score (5-scale)		
	Small corpus	*Big corpus*	*Final score*
NATURAL	4.58	4.30	4.44
SUN	**4.13**	3.47	3.80
ZALO	3.77	**4.10**	**3.94**
VNGGRD	3.99	3.70	3.85
VTCC	3.37	2.43	2.90

We can see that SUN was the best team (i.e. 4.13) with the small training corpus while ZALO had a highest MOS score with the big training corpus. In average, ZALO was the first place with 3.94 point (natural speech was 4.44) on a 5-point MOS scale. VNGGRD was the second place with 3.85 score (nearly less than the first place 0.1 point) while SUN was the third place with 3.80.

B. *Analysis*

In this year, statistics were presented for two conditions: "strict" (using responses only from listeners who completed the whole listening test) and "lax" (using responses from all listeners, but discarding partially-completed sections). A two-factorial ANOVA was run on the results.

For the Big corpus test, as the Table III, the two factors were the TTS system (5 levels) and the Sentence (44 levels). The TTS system and the interaction between System and Sentence had significant effect ($p<0.0001$) while the Sentence factor is

[1] Available on https://ptest.vbee.vn

not significant (p>0.05). The TTS system factor alone explained an important part of the variance, about 36% (partial η^2 = 0.36), while the Sentence factor and their interaction explained about 23% and 11% correspondingly.

TABLE III. ANOVA RESULTS OF MOS TEST FOR THE BIG CORPUS

Factor	df	df error	F	p	η^2
System	4	1,131	335.32	0.0000	0.36
Sentence	43	49	1.34	0.0691	0.23
System:Sentence	172	247	1.71	0.0000	0.11

For the Small corpus test, as the Table IV, the two factors were the TTS system (5 levels) and the Sentence (38 levels). All factors had significant effect (p<0.0001) while their interactions are likely due to chance (p>0.1). The TTS system factor alone explained an important part of the variance, about 18% (partial η^2 = 0.18), while the Sentence factor and their interaction explained only about 4% to 7%.

TABLE IV. ANOVA RESULTS OF MOS TEST FOR THE SMALL CORPUS

Factor	df	df error	F	p	η^2
System	4	365	117.45	0.0000	0.18
Sentence	37	74	2.56	0.0000	0.04
System:Sentence	148	123	1.07	0.2722	0.07

C. Discussion

The natural speech of the big corpus was evaluated not good, only 4.30 point on a 5-point MOS scale. The reason was found that the natural speech had many noises and was recorded by a non-professional speaker in a non-professional place. Whereas, the small corpus was recorded in a professional studio by a TV broadcaster with a professional voice. It was given 4.58 score.

There were some objective feedbacks for synthetic voices of partipants from our primilary tests as well as from listeners after the evaluation. The subjects seemed classified audios to some groups corresponding to participants although these groups were hidden to them:

- ZALO voice: A clear and good voice from the big training corpus in terms of removing noises. The voice from the small training corpus was still not really natural, especially with loanwords or uncommon phrases;
- SUN voice: A very sweet and smooth voice from the small training corpus. The voice from the big corpus still had some background noises that made the listeners feel quite uncomfortable;
- VNGGRD voice: The synthetic voices had mechanical noises that made the listeners feel unforcomforatble;
- VTCC voice: The synthe voices was intelligible but not really smooth and natural in general.

VI. CONCLUSIONS

The Text-To-Speech (TTS) shared task in the VLSP Campaign 2019 has been a valuable exercise in building voices from a common dataset and has brought together different teams looking at a common goal. We plan to have somem further specific research on Speech Synthesis such as text normalization, speaker adaptation for the next VLSP Campaign in 2020.

ACKNOWLEDGMENT

This challenge was mainly supported by the R&D Lab, Vbee Services and Data Processing Solution Jsc. Vbee supported us in developing, deploying and conducting the online evaluation, based on perception tests as well as providing the small corpus for the challenge. We would like to thank InfoRe Jsc. for their big corpus. Last but not least, the subjects who gave time and effort for the experiments.

REFERENCES

[1] Luong Chi Mai. Special issue in VLSP 2018. Journal of Computer Science and Cybernetics, V.34, N.4 (2018).

[2] 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.

[3] H. Ze, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” In 2013 ieee international conference on acoustics, speech and signal processing 2013, pp. 7962–7966. IEEE.

[4] Nguyen Van Thinh, Nguyen Quocs 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.