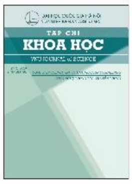
VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce



Original Article

VLSP 2021 - TTS Challenge: Vietnamese Spontaneous Speech Synthesis

Nguyen Thi Thu Trang[1,2,*], Nguyen Hoang Ky[2]

[1]Hanoi University of Science and Technology, 1 Dai Co Viet, Hai Ba Trung, Hanoi, Vietnam
[2]Vbee Services and Data Processing Solution Joint Stock Company,
160 Ton Duc Thang, Dong Da, Hanoi, Vietnam

Received 27 December 2021
Revised 5 April 2022; Accepted 5 May 2022

Abstract: Text-To-Speech (TTS) was one of nine shared tasks in the eighth annual international VLSP 2021 workshop. All three previous TTS shared tasks were conducted on reading datasets. However, the synthetic voices were not natural enough for spoken dialog systems where the computer must talk to the human in a conversation. Speech datasets recorded in a spontaneous environment help a TTS system to produce more natural voices in speaking style, speaking rate, intonation... Therefore, in this shared task, participants were asked to build a TTS system from a spontaneous speech dataset. This 7.5-hour dataset was collected from a channel of a famous youtuber "Giang ơi..."and then pre-processed to build utterances and their corresponding texts. Main challenges at this task this year were: (i) inconsistency in speaking rate, intensity, stress and prosody across the dataset, (ii) background noises or mixed with other voices, and (iii) inaccurate transcripts. A total of 43 teams registered to participate in this shared task, and finally, 8 submissions were evaluated online with perceptual tests. Two types of perceptual tests were conducted: (i) MOS test for naturalness and (ii) SUS (Semantically Unpredictable Sentences) test for intelligibility. The best SUS intelligibility TTS system had a syllable error rate of 15%, while the best MOS score on dialog utterances was 3.98 over 4.54 points on a 5-point MOS scale. The prosody and speaking rate of synthetic voices were similar to the natural one. However, there were still some distorted segments and background noises in most of TTS systems, a half of which had a syllable error rate of at least 30%.

Keywords: VLSP Campaign 2021, TTS shared task, speech synthesis, text-to-speech, spontaneous speech, evaluation, perception test, Vietnamese.

1. Introduction

VLSP (Vietnamese Language and Speech Processing Consortium) is an initiative to establish a community working on speech and text processing for the Vietnamese language [1]. The Text-To-Speech (TTS) shared task was a

[*] Corresponding author.
E-mail address: trangntt@soict.hust.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.358

challenge in the VLSP 2021, the eighth annual international workshop. This was the fourth time we organized the challenge in speech synthesis (Table 1). DNN TTS model was the winner in the VLSP 2018 [1]. For the next 2 years, 2019 [2] and 2020 [3], the acoustic model Tacotron2 and Waveglow or HifiGAN vocoder showed strength and won in both these competitions.

This year's contest topic was inspired by spoken dialog systems. These systems are getting to be even more across-the-board. Nevertheless, interaction quality is not reaching its full potential, possibly due to problems with the voice [4]. Adapting read speech voices for synthesizing conversations is not direct [5] and it stands to reason that interactions might improve if dialogue systems were able to speak truly conversationally, rather than with voices based on written prompts read aloud.

This shared task has been designed for understanding and figuring out remaining problems in Vietnamese TTS with spontaneous speech dataset. The main challenge participants had to deal with were the development and use of appropriate TTS models to train spontaneous data. Spontaneous data with many different intonations, different speaking speeds, and different speaking environments will cause difficulties in the model training process. The participating teams also need to have reasonable audio data preprocessing strategies to make the training process easier.

Table 1. Previous Vietnamese TTS shared tasks in VLSP:

Year	**Topic**	**Common Datasets**	**Challenge**	**Winner Tech Stack**
2018	Freely TTS	No	TTS Techniques	DNN
2019	TTS on difference dialects	Big (Northern) common datasets	Non-professional & noisy voice	Tacotron 2 & WaveGlow
		Small (Southern) common datasets	Low-resource	
2020	TTS on collected reading datasets	Southern-West speaker ("For whom the bell tolls")	Prosodic phrasing Loanwords	Tacotron 2, Hifi-GAN WaveGlow Denoiser

Participants took the released speech dataset, built a synthetic voice from the data and submitted 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 2 presents the spontaneous common dataset and its preparation. Section 3 introduces participants and a complete process of the TTS shared task in VLSP Campaign 2021. We then show the evaluation design and experimental results in Section 4 and Section 5. We finally conclude the task and give some possible ideas for the next challenge in Section 6.

2. Spontaneous Common Dataset

The topic of this shared task is to address the main problems of TTS systems using spontaneous dataset to build natural speech. Due to the topic of this year's task, we decided to collect audio from the Internet, especially Youtube for more specific. Vbee Jsc supported to build the dataset for this task. The corpus was taken from a youtube channel named "Giang ơi". The youtube channel belongs to Tran Le Thu Giang - a Vietnamese content creator, vlogger and environmental activist - she is widely known through her YouTube channel for sharing videos about her life, work experience, study and inspiration. We automatically collect the audio of a total of 325 videos on the "Giang ơi"youtube channel. The voice activity detection system was used for splitting audio into smaller audio files that are less than 10 seconds in length. The

Speech-to-Text system then automatically converted these audio files to text transcripts. After this process, the number of sound files was up to 22,839 sound files (equivalent to 72 hours audio) with different lengths.

The collected data is spontaneous and was recorded in different environments so there are quite a few preprocessing steps to do to improve the quality of the dataset. Some of the main issues and challenges of this year's TTS shared task are (i) background noises, (ii) multiple voices mixed in one audio, (iii) differences in intensity, stress, and prosody across the dataset, and (iv) inexact transcripts. We help participants get rid of audio files that are corrupted or too short, and also remove audio files that have a lot of voices mixed together. The number of audio files of the dataset is reduced to 6,266 files (equivalent to about 11 hours). These data were checked by the teams participating in the contest. Each team only had to check 400 files for participation. Finally, 5,341 best quality utterances (approximately 7.5 hours) and their corresponding texts were selected as the final dataset. Although the dataset has been cleaned up, problems still exist, especially in terms of background noises and differences in reading styles. These problems are also the main challenges for the participating teams.

3. Participants

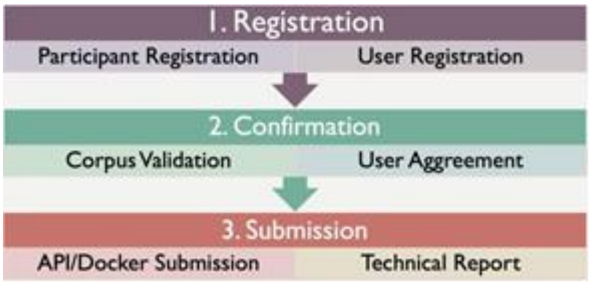


Figure 1. A complete process for participating TTS shared task VLSP 2021.

As the TTS shared task last year, participants had to follow a complete process (Figure 2), which was managed in the website of the TTS shared task of VLSP Campain 2021 (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 2 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.

Table 2. Teams with final submissions:

Team ID	Team Name	Affiliation
Team1	Navi	HUST
Team2	-	VCCorp
Team3	-	VinBDI
Team4	-	VinBrain
Team5	Smartcall	Smartcall
Team6	-	HCMUS
Team7	-	HUST
Team8	-	HUST
Team9	-	HUST
Team10	Thunder	HUST

Participants were asked to build only one synthetic voice from the released database. All teams had 24 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.

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.

Figure 2 compares the number of participants of last year to this year. Forty-three teams registered for this year's challenge. This is the second year in a row since last year that we have asked competing teams to validate the provided data. Unlike last year, nearly all teams participated in data validation (42/43 teams), and 18 teams obtained the data after sending the signed user agreement. Finally, ten teams, compared to eight in 2020, submitted their TTS system. We synthesized testing audio through the TTS API of each team. This year, we require all participating teams to submit a solution paper and this paper will be scored along with other tests. Table 2 gives the list of participants that had final submissions to the VLSP TTS shared task 2021 and their respective technique stack. There were 4 teams that did not submit the paper so they were eliminated from the final standings. Almost every team used FastSpeech2 as the Acoustics model and HifiGAN as the Vocoder, only Smartcall used a fully end-to-end VITS model.

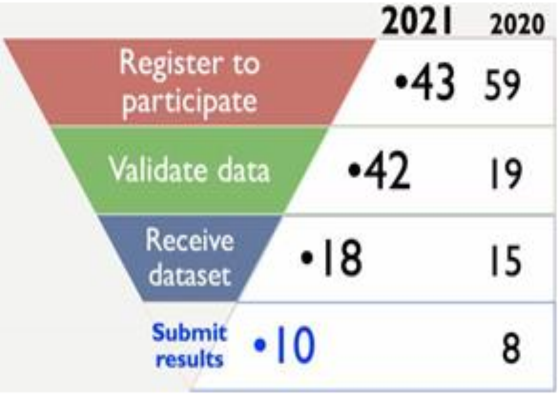


Figure 2. Participants in VLSP TTS 2021 and 2020.

4. Evaluation

We chose perceptual testing for evaluating synthetic voices. First, the in-domain MOS test and out-domain MOS test were performed for comparing the global quantity of the TTS system with respect to natural speech references. Testing transcripts were chosen from real conversations to evaluate the ability of the TTS system for applying in dialog scenario case. Second, an intelligibility test - Semantically Unpredictable Sentences (SUS) Test - was conducted to measure the intelligibility. 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).

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)[6] for all sub-tests, where n is a number of voices in the sub-test. To be more specific, each subject listened to one nth 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.

4.1. MOS Test

Subjects (i.e. listeners) were asked to assess by giving scores to the speech they had heard (Figure 3). 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)
- 4: Good, natural
- 3: Fair, rather natural
- 2: Poor, rather unnatural (rather robotic)
- 1: Bad, very unnatural (robotic).

Table 3. Final Perceptual Test Results:

Voice	Technique Stack	MOS	SUS SER	MOS
		(out-domain)	**Intelligibility**	**(in-domain)**
Ground Truth	-	-	-	4.54
Team1 (1st)	Fastspeech2 (External aligner) + HifiGAN + HifiGAN denoiser	3.56	0.20	3.73
Team2	Fastspeech2 + HifiGAN	2.81	0.25	3.27
Team3	-	3.0	0.20	3.53
Team4	-	2.52	0.38	2.75
Team5 (3rd)	VITS - Fully End2End	3.52	0.22	3.81
Team6	Fastspeech2 + HifiGAN	2.66	0.32	3.79
Team7 (3rd)	FastSpeech2 + HifiGAN + Waveglow Denoiser	3.37	0.16	3.98
Team8	-	2.27	0.38	3.1
Team9	-	2.81	0.30	3.88
Team10 (2nd)	Fastspeech2 (+ Postnet layer) + HifiGAN + Waveglow Denoiser	3.30	0.15	3.94

Table 4. Setup for MOS Test Out-domain, MOS Test In-domain and Intelligibility Test (SUS):

	MOS-Out	MOS-In	SUS
Text file #	30	24	36
Utterance #	300	240	360
Session #	2	2	3
Subject #	50 (30-30)	50 (24-24)	50 (28-28-28)

Table 4 describes information about the Intelligibility and MOS tests. There is a difference in the MOS test compared to previous years. There are two kinds of MOS test including in-domain MOS test and out-domain MOS test. The in-domain MOS test dataset contains 30 individual sentences. Text and natural voice set is taken from “Giang ơi” youtube channel which is excluded from the training data. The out-domain MOS test - an important test for the criteria of this year’s competition - includes 24 short multi-turn conversations, each containing several lines of dialogue and separated by short pauses (3 seconds). Natural voices are not used in this test. Each MOS test was processed in two sessions and the number of subjects who joined each session was 50 (33 males).

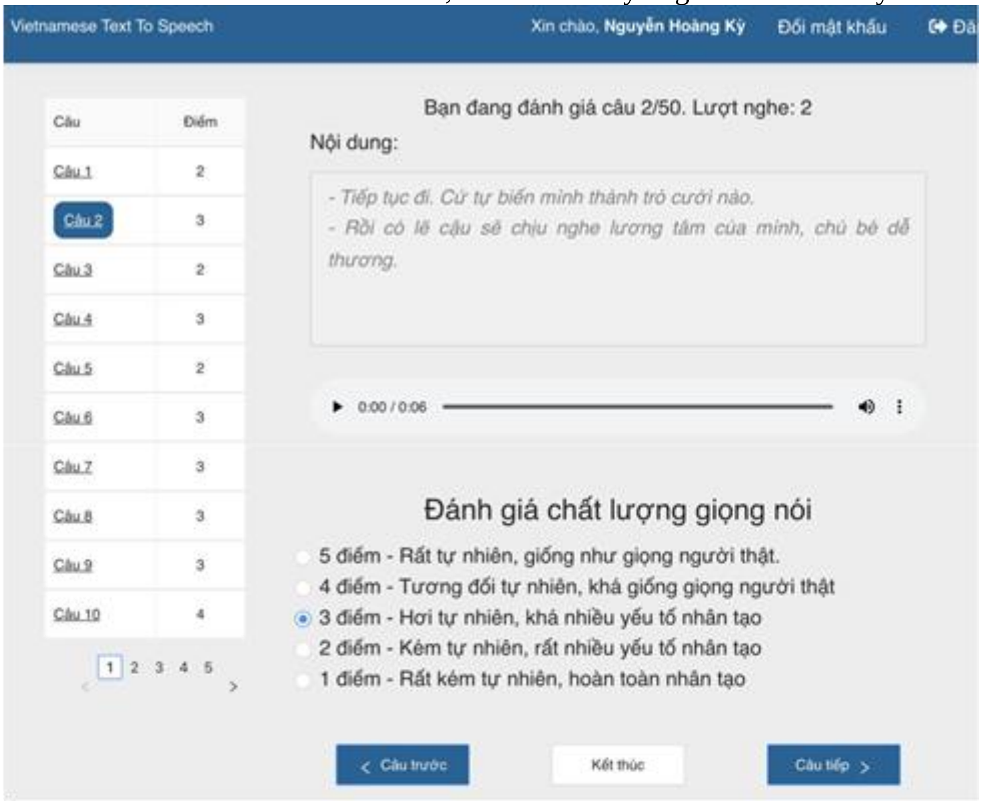
4.2. Intelligibility Test

One of the biggest changes in this year’s competition is the Intelligibility test designed based on the Semantically Unpredictable Sentences (SUS) method [7]. The idea behind this method is to avoid using simple and meaningful sentences. These sentences provide semantic and syntactic contextual cues whose effect on intelligibility scores cannot readily be quantified.

The sentence structures used to build the SUS test in [7] are the sentence structures used for English. As far as we know, there has not been any test dataset for Vietnamese built according to this method that has been public and used in TTS competitions. Therefore, we rely on [7] and make changes to build a set of sentences suitable for the Vietnamese language.

We build the dataset by a semi-automatic method. Vbee Jsc provides a text dataset with about 1 million dialogues and stories. We split the sentences and use VnCoreNLP [8] to assign POS labels to these sentences, each sentence has a corresponding POS pattern. For example, "anh/N khỏi/V lo/V" (you don’t have to worry) has a POS pattern of "N - V - V". We filter out the top 50 most frequently occurring patterns. For each pattern, we randomly fill in the pattern with the words in the corpus with the same POS label, thereby generating a new sentence with the same POS pattern. A total of 300 new sentences are generated (6 sentences per pattern). Finally, we observed and selected the best 36 sentences, these sentences need to satisfy the requirements of SUS and at the same time have the correct Vietnamese grammar structure. Example sentences in Intelligibility test data:

- "Chị/N, uống/V bao_nhiêu/P điếu/N rồi/C mới/R bay/V lên/R thế/P anh/N ơi/I" ("how many cigarettes do you drink before you fly into the air")
- "Trời_ơi/I bấy_nhiêu/P thời_gian/N giỏi/A chiều/N quá/R" ("oh gosh that much time is good at pampered")
- "Con_người/N nên/C làm/V tổ/N và/Cc đi/V ngủ/V rớt/V lên/V" ("humans should make nests and go to sleep falling up")
- "Vợ/N bà/N để/E bạn/N đi/V phía/N em/N đột_nhiên/R chạy/V nhiều/A" ("Your wife who let you go to me suddenly ran a lot").



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

Bạn đang đánh giá câu 2/50. Lượt nghe: 2

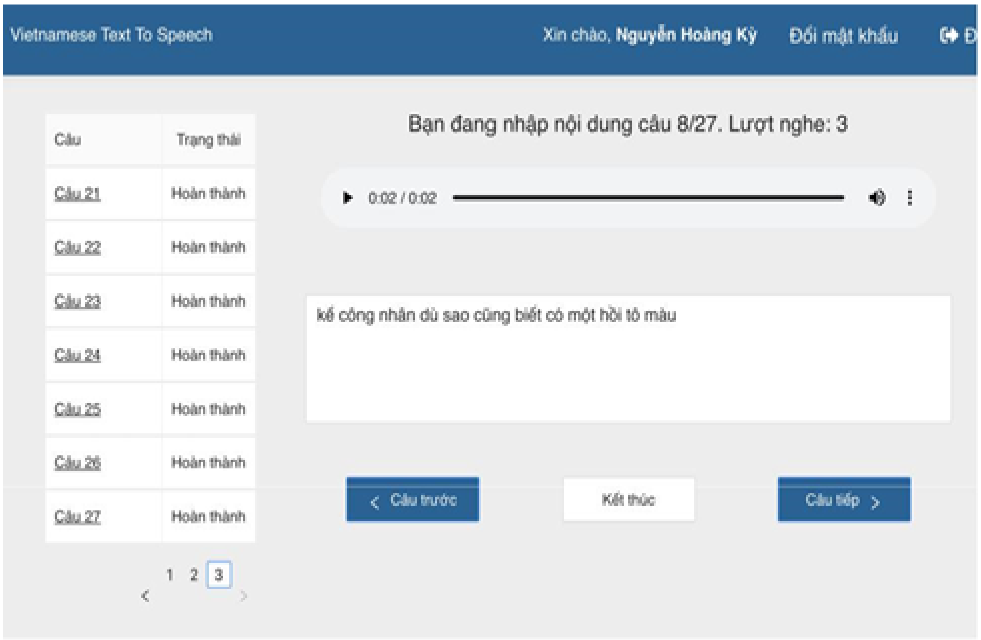
Nội dung:

- Tiếp tục đi. Cứ tự biến mình thành trò cười nào.
- Rồi có lẽ cậu sẽ chịu nghe lương tâm của mình, chú bé dễ thương.

Đánh giá chất lượng giọng nói

○ 5 điểm - Rất tự nhiên, giống như giọng người thật.
○ 4 điểm - Tương đối tự nhiên, khá giống giọng người thật
● 3 điểm - Hơi tự nhiên, khá nhiều yếu tố nhân tạo
○ 2 điểm - Kém tự nhiên, rất nhiều yếu tố nhân tạo
○ 1 điểm - Rất kém tự nhiên, hoàn toàn nhân tạo

Figure 3. Online Tool for MOS Test.



Câu	Trạng thái
Câu 21	Hoàn thành
Câu 22	Hoàn thành
Câu 23	Hoàn thành
Câu 24	Hoàn thành
Câu 25	Hoàn thành
Câu 26	Hoàn thành
Câu 27	Hoàn thành

Bạn đang nhập nội dung câu 8/27. Lượt nghe: 3

kế công nhân dù sao cũng biết có một hồi tô màu

Figure 4. Online Tool for Intelligibility Test (SUS).

Subjects were asked to write down the text of the audio they heard (Figure 4). 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 sentences were very long. There were three sessions performed with 50 subjects and 25-30 minutes estimated each.

5. Evaluation Results

5.1. 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. Table 3 show the final MOS test results. Six teams (more than half of the total number of teams) submitted technical reports, i.e. Team1, Team2, Team5, Team6, Team7 and Team10.

In the out-domain MOS test, we can see that Team1 was the best team (i.e. 3.56). This team adopted FastSpeech2 as the acoustic model with the external aligner replacement, and HiFi-GAN as a vocoder, and HiFi-GAN denoiser. One of the highlights of Team1's preprocessing is the audio filtering technique. This solved the problems regarding voices from other speakers and also removed recordings with abnormal speaking styles. They chose 5 audio samples that have the main speaker's voice and used [9] to filter out audio files having the different embedding. Noise reduction, audio normalization and punctuation prediction were also proceed. Team5 was the second place with a 3.52 score (0.04 point less than Team1). This stack compared to other teams, they used team used a completely different technology the VITS - a fully End2End model. In the in-domain MOS test, the outcome was completely different. The results show that Team7 was the best team (i.e. 3.98) – about 87.7% compared to the natural speech (i.e. 4.54/5). This team adopted FastSpeech2 as the acoustic model, and HiFi-GAN as a real-time vocoder, and Waveglow as a denoiser. Team10 was the second place with a 3.94 score (only less than the first place 0.04 point). This team used the same technical stack as Team7 with some Postnet layer modification. In this year's training dataset, there were many audios that are less than 2 seconds long. Therefore, Team10 implemented data enhancement by concatenating short utterances into longer utterances that lasted about 4 - 10 seconds. Team7 chose fully end2end model (i.e. VITS) that could produce natural utterances with good prosody, compared to other teams. However, some synthetic utterances were wrong pronounced or partly skipped, and still had some background noises. Those may lead to not best score for Team7.

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 in the synthetic voice, compared to the human voice: (i) background noises, (ii) not natural representation for the conversational style like in certain contexts and (iii) wrong/bad pronunciations or distorted words.

Table 5. ANOVA Results for MOS Test (in-domain):

Factor	df	df error	f	p	η2
System	10	2,040	71.7979	0.0000	0.26
Sentence	23	2,040	2.8769	0.0000	0.03
System:Sentence	230	2,040	1.5255	0.0000	0.15
System	10	2,016	104.2655	0.0000	0.34
Subject	47	2,016	23.3967	0.0000	0.35
System:Subject	230	2,016	1.9673	0.0000	0.18

Table 6. ANOVA Results for MOS Test (out-domain):

Factor	df	df error	f	p	η2
System	9	2,700	63.2035	0.0000	0.17
Sentence	29	2,700	9.7403	0.0000	0.10
System:Sentence	261	2,700	1.6802	0.0000	0.14
System	9	2,700	75.3682	0.0000	0.20
Subject	49	2,700	18.7290	0.0000	0.25
System:Subject	241	2,700	1.9159	0.0000	0.15

Table 7. ANOVA Results for Intelligibility Test (SUS):

Factor	df	df error	f	p	η2
System	9	2,160	66.1440	0.0000	0.22
Sentence	35	2,160	30.8373	0.0000	0.33
System:Sentence	315	2,160	3.1577	0.0000	0.32
System	9	2,240	42.7217	0.0000	0.15
Subject	49	2,240	5.5310	0.0000	0.11
System:Subject	221	2,240	1.0121	0.4412	0.09

5.2. Intelligibility Score

Because the sentences in the test set are quite semantically unreasonable, they have caused a certain difficulty for the listeners and affected the results. Listeners had difficulty guessing inaudible words, making the test results more accurate and meaningful. After normalization, the highest SUS score was 85.00 and the lowest score was 62.10, belonging to Team10 and Team8 respectively. Although Team5 has a good result in the MOS test, they have a pretty bad result in this test with only 78.2 points.

5.3. Analysis and Discussion

Several two-factorial ANOVAs were run on the MOS Test Out-domain, MOS Test In-domain and Intelligibility Test (SUS) results, illustrated in Table 5, 6, 7 correspondingly. The two factors were the TTS system (10 levels for MOS Test Out-domain and 9 levels for the rest) and the Sentence (23, 29, and 35 levels respectively) or the Subject (47, 49, and 49 levels). All factors and their interactions in both ANOVAs had significant effect ($p<0.0001$), except the interaction between System and Subject in the Intelligibility Test ($p=0.4412$).

In all tests, the TTS system factor alone explained an important part (15-34%) of the variance. The Sentence factor in the two MOS tests explained only 3-10% of the variance (partial η2 = 0.03 and 0.10 respectively) while it was 33% in the intelligibility test. The Subject factor did a great effect, i.e. 25-35%, in the two MOS tests, but only 11% in the intelligibility. The interaction between System and Sentence in the two MOS tests explained a quite important part (14-15%) of the variance but 33% in the intelligibility one (partial η2 = 0.33).

6. Conclusion

We did some valuable experiments on TTS systems from different participants using a spontaneous dataset in the TTS shared task in the VLSP Campaign 2021. Participants had to validate a piece of training data before receiving the common dataset. There are 5,341 utterances of a female Northern speaker (about 7.5 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 Fastspeech2 and Hifigan vocoder with HifiGAN denoiser got a high MOS test score (3.73 in in-domain MOS test and 3.56 in out-domain MOS test) and good intelligibility (0.2% SER). End2End model, e.g VITS, got good voice quality and prosody when training with the spontaneous dataset but it was not as good as other models in the SUS intelligibility test. Some reasons were found in the synthetic voice, compared to the human voice: (i) background noises, (ii) not natural representation for the conversational style like in certain contexts and (iii) wrong/bad pronunciations or distorted words, especially with end2end models. For the next speech synthesis task of the VLSP Challenge in 2022, we may have more advanced topics for Vietnamese speech synthesis, such as speaker adaptation or expressive speech synthesis.

Acknowledgments

The VLSP 2021 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, Hanoi University of Science and 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. Last but not least, we would like to thank the subjects from VNU University of Science and Hanoi University of Science and Technology, who gave time and effort for the experiments.

References

[1] L. C. Mai, Special issue in vlsp 2018, Computer Science and Cybernetics, Vol. 34, No. 4, 2018.

[2] N. T. T. Trang, N. X. Tung, Text-to-speech shared task in vlsp campaign 2019: evaluating vietnamese speech synthesis on common datasets, Vietnamese Language Signal Processing. VLSP.

[3] T. T. T. Nguyen, H. K. Nguyen, Q. M. Pham, D. M. Vu, Vietnamese text-to-speech shared task vlsp 2020: Remaining problems with state-of-the-art techniques, in: Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, 2020, pp. 35–39.

[4] L. Clark, N. Pantidi, O. Cooney, P. Doyle, D. Garaialde, J. Edwards, B. Spillane, E. Gilmartin, C. Murad, C. Munteanu, et al., What makes a good conversation? challenges in designing truly conversational agents, in:

Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–12.

[5] M. Wester, O. Watts, G. E. Henter, Evaluating comprehension of natural and synthetic conversational speech, in: Proc. speech prosody, Vol. 8, 2016, pp. 736–740.

[6] W. G. Cochran, Experimental designs 2nd ed., John wiley & sons, 1957.

[7] C. Benoˆıt, M. Grice, V. Hazan, The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences, Speech communication 18 (4) (1996) 381–392.

[8] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, M. Johnson, Vncorenlp: A vietnamese natural language processing toolkit, arXiv preprint arXiv:1801.01331.

[9] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, et al., Speechbrain: A general-purpose speech toolkit, arXiv preprint arXiv:2106.04624.