Direct speech-reply generation from text-dialogue context

Kenichi Fujita*,Yusuke Ijima*, Hiroaki Sugiyama*
* NTT Corporation, Japan
E-mail: kenichi.fujita.wv@hco.ntt.co.jp

Abstract—**Natural speech-dialogue generation has been achieved with cascade systems combining automatic speech recognition, text-dialogue, and text-to-speech models. However, it is still challenging to generate expressive speech-replies depending on context because text-replies could lead to information loss in estimating appropriate expressions for speech generation. One promising approach is generating speech without requiring text. Direct speech generation from a dialogue context has never been achieved because it is difficult to learn the semantically one-to-many relationship between context and reply. This paper proposes a direct speech-reply generation model from the text-dialogue context in the same manner as the text-dialogue model. We focus on two challenges: an insufficient number of training dialogue pairs of text-context and speech-reply, and the difference between continuous speech signals and discrete text sequences. For the former, we applied text-to-speech to a text-dialogue dataset to acquire huge-scale training pairs. For the latter, we introduced the vector quantization on acoustic features to convert them into discrete sequences. The results indicate that the proposed model can successfully generate speech-reply directly from text-dialogue contexts, although a quality gap still exists with the text-dialogue model.**

Index Terms: dialogue generation, open-domain chat, speech synthesis, vector quantization

I. Introduction

Highly natural text-dialogue generation in small talk has recently been achieved in the use of encoder-decoder networks such as Transformer [1], [2], [3]. Certain systems achieved speech-reply generation by connecting automatic speech recognition (ASR), text-dialogue, and text-to-speech (TTS) models [4], [5]. In such a system, ASR converts input spoken-dialogue context into text-dialogue context. The text-dialogue model generates text-reply from the text-dialogue context. Finally, the TTS model generates speech-reply from the generated text-reply. It is still difficult for such a system to generate expressive speech and natural speech, speech including filler and mumbles, appropriate for the given context since the system generates speech just from the text information. Yamanaka et al. proposed using auxiliary information such as an emotional label for speech generation [6]. However, the types of emotions are limited and insufficient to generate appropriate speech reflecting the context.

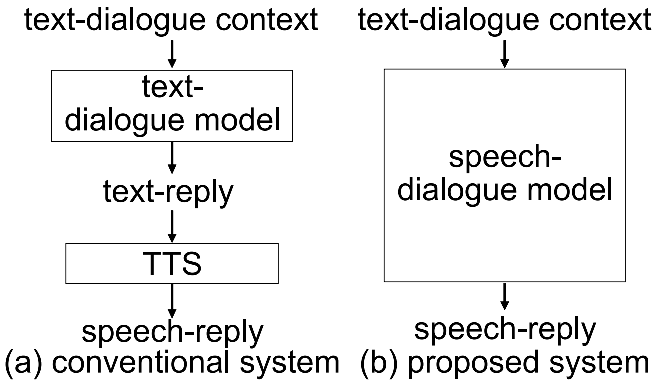
We propose a direct speech-reply generation model from a text-dialogue context, which includes richer information than a simple text. This model resemble conventional speech generation models, such as TTS [7], [8], image-to-speech [9], [10], [11], and speech-to-speech-translation [12], [13], [14]. These speech generation models only model semantically one-to-one relationships between input and output. However, our model requires modeling the semantically one-to-many relationship between dialogue context and reply. For example, TTS generates speech of input text. The research of image-to-speech generates speech almost uniquely determined from the detected objects in the given image. The research of speech-to-speech translation has generated translated speech almost uniquely determined from the input speech. In contrast, in small talk, a wide variety of semantically different replies are expected to be uttered in response to a single context.

Because of this challenge, we emphasize research questions over achieving expressive speech-replies generation. Our research questions are as follows. (1) Is it possible to generate natural speech from a model trained with a semantically one-to-many relationship? (2) If possible, does the generated speech have diversity as with a text-dialogue model? We trained our model, which generates speech-replies from a given text-dialogue context, to answer these questions. The key points of the proposed model are the use of TTS and vector quantization. For the first point, we applied TTS for text-pairs (the pairs of text-dialogue context and text-reply) to acquire enough speech-pairs (the pairs of text-dialogue context and speech-reply) for training the model. Text-pairs are easily obtained from social networking services such as Reddit and Twitter. For the second point, we introduced vector quantization on acoustic features to handle continuous speech signals the same as a discrete text sequence. We conducted objective and subjective experiments showing that the proposed model could successfully generate speech-reply directly from text-dialogue contexts. However, there was still a gap in quality with a conventional text-dialogue model.

II. Text-dialogue model

This section describes BlenderBot [2], which is the basis of the conventional text-dialogue model used in our study. BlenderBot is an open-domain chatbot that first achieved conversation by combining skills such as empathy, persona, and knowledge. BlenderBot has an encoder-decoder architecture based on Transformer. This model is trained with pre-training and fine-tuning. In pre-training, about 1.5 billion post-reply pairs from the Reddit dataset are used to learn the features of natural conversations. In fine-tuning, the model is trained with the domain-specific dataset. The model learns empathy, persona, and knowledge from the Blender Skill Talk dataset consisting of ConvAI2, Wizard of Wikipedia, and Empathetic

978-616-590-477-3 ©2022 APSIPA



(a) conventional system (b) proposed system

Fig. 1. Overview of conventional text-dialogue and speech-dialogue models.

Dialogues datasets.

III. Proposed speech-dialogue model

Conventionally, speech-dialogue systems generate speech-replies by applying TTS for the generated texts from a text-dialogue model (Fig. 1 (a)). The proposed model, however, focuses on direct speech-reply (acoustic features) generation from text-dialogue context without using a TTS model (Fig. 1 (b)) based on a Transformer-based encoder-decoder model. Two significant challenges need to be addressed to achieve this: (1)the lack of a dataset containing speech-pairs sufficient for training the proposed model and (2)the conventional text-dialogue models, which require discrete sequences for their output, do not accept the speech signals as their output because the signals are continuous.

Regarding Challenge 1, we generate speech-pairs from text-pairs by using TTS. The many generated speech-pairs are comparable with the training dataset for text-dialogue models such as Meena [1] and BlenderBot.

Regarding Challenge 2, we convert speech signals into a discrete sequence by vector quantization to handle speech signals in the same manner as a text sequence. A text-dialogue model, such as BlenderBot, converts text-dialogue context and text-replies into a sequence of indices by using SentencePiece [15]. However, it is difficult to apply the tokenization method to speech signals. Therefore, we introduced an LBG algorithm [16] for the acoustic features of the speech-replies. The algorithm determines C-cluster centroids for encoding the target speech into a sequence of cluster indices. The sequence of acoustic features is generally longer than that of a text sequence. Therefore, we concatenated L frames of acoustic features to easily train and make the sequence fit into the model's maximum output length. The following is the process of quantization. First, the vectors in a melspectrogram consisting of T frames ($\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_T}]$) are concatenated at every L frame ($\boldsymbol{Y} = [\boldsymbol{y_1}, \boldsymbol{y_2}, ..., \boldsymbol{y_{T/L}}]$), where $\boldsymbol{y_k} = (\boldsymbol{x_{k/L}}, \boldsymbol{x_{k/L+1}}, \boldsymbol{x_{k/L+2}}, \boldsymbol{x_{k/L+3}})(k = 0, L, 2L, ...)$. The sequence is then converted into a sequence of cluster indices $\boldsymbol{Z} = [z_1, z_2, ...z_{T/L}]$, where $z_t \in [C]$ is the C-class categorical variable (Fig. 2).

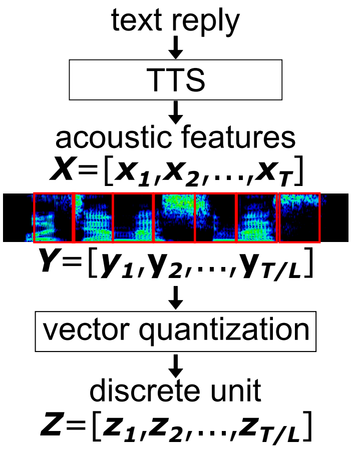


Fig. 2. Vector quantization

IV. Experiments

A. Dataset

1) Dataset for pre-training and fine-tuning: The dataset for training and evaluation was the text-pairs from the part of the dataset used for Japanese BlenderBot [17]. This dataset contains 2.1 billion pairs from Twitter for pre-training, and FavoriteThingsChat for fine-tuning.

Our dataset for pre-training consists of 400 million pairs from Twitter for training, 50,000 for validation, and 50,000 for testing. We retrieved the tweets from randomly sampled Japanese users from January 2016 to September 2016. After cleaning, we extracted the tweets in a reply relationship and paired them as dialogue context and replies.

The dataset for fine-tuning generated from the FavoriteThingsChat dataset consists of 101,587 pairs for training, 4,017 for validation, and 5,254 for testing. This dataset is from extensively collected text-chats between pairs of 80 participants talking with more than 60 other participants about their favorite things. The fine-tuning dataset uses up to four utterances as a dialogue context until the maximum character length reaches 128.

2) Preprocess for dataset: The text sequences were tokenized with the SentencePiece tokenizer [15] with a dictionary of 32,000 words implemented on the Official Github site. The tokenizer was trained with 20 million sentences sampled from the data of the Japanese question-answer community service “Oshiete goo!.” These data cover the period from 2001 to 2019 containing more recent topics than our pre-training dataset.

For generating speech-pairs from text-pairs, text-replies from text-pairs were converted to synthesized speech by using TTS. The TTS model was TransformerTTS [18] trained with about 10 hrs of speech data uttered by a single Japanese professional female narrator. The sampling frequency was 24 kHz. We used 80-dimensional melspectrograms as the acoustic feature. The frameshift and window size were 12.5 and 50 ms, respectively. The 80-dimensional acoustic features were concatenated every four frames into 320-dimensional vectors before converting generated continuous acoustic features into discrete symbols. They were then quantized via vector quantization into 4,096 discrete symbols. The number of discrete symbols was experimentally determined. The total duration of generated speech for pre-training was 338,321 hrs, and that for fine-tuning was 133 hrs.

B. Training conditions

We trained Transformer-based encoder-decoder models for the proposed model with the generated speech-pairs described in the previous section. We determined the model parameters on the basis of Japanese BlenderBot with 1.6 billion parameters [17]. The encoder has 2 layers, and the decoder has 24 layers. Each layer has 1,920 units and 32 multi-heads. The number of dimensions of the hidden layers was adjusted to avoid memory errors on the GPU (V100 16GB) available at AIST ABCI Cloud. The dropout of the feed-forward layer and attention was 0.1. To compare the naturalness and diversity of the generated replies, we also trained the conventional text-dialogue model on the basis of BlenderBot with the same dataset as the proposed model. The parameters of the models were the same except for the difference in the dimension of the output target sequence.

During pre-training, we set 1e-4, 3,000, and 2.1 million as the learning rate, warmup steps, and maximum number of tokens per step, respectively. The objective function was cross-entropy. The computational resources were 128 V100 16GB cards. The number of training steps was 200,000, almost equivalent to three epochs. Our encoders used the fixed parameters from the encoders of Japanese BlenderBot [17]. During fine-tuning, we respectively set 5e-5, 100, and 22k as the learning rate, warmup steps, and maximum number of tokens per step. The computational resource was a single A100 40GB card. The model parameters were not fixed during fine-tuning.

C. Decoding condition and waveform generation

We used Sample-and-Rank, a method used in Meena, to generate more diverse replies than beam-search decoding. This method selects a candidate with the lowest Perplexity as a final output from independently generated N candidates, where N was 20 in the experiment. We also introduced temperature T and top-p sampling, where T controls the output probability when calculating the softmax in generating tokens. The condition $T = 1$ is the standard sampling, and when T is lower, more likely safe and common words will be selected. From preliminary experiments, we set 0.95 for T in the proposed model and 0.99 for the conventional text-dialogue model, respectively. The top-p sampling limit the number of words sampled by the probability cumulative density. Thus, we set 0.9 for top-p.

Finally, the waveform was generated from the sequence of the quantized acoustic features. First, the acoustic feature was reconstructed from the sequence of the cluster indices by converting each cluster ID into the corresponded centroid. Next, waveform generation was conducted on the reconstructed acoustic features. We used HiFi-GAN [19] for waveform generation.

TABLE I
RESULTS OF OBJECTIVE EVALUATIONS. PRE(TEXT/SPEECH) DENOTES PRE-TRAINING MODEL(TEXT/SPEECH), AND FT(TEXT/SPEECH) DENOTES FINE-TUNED MODEL(TEXT/SPEECH). ORIGINAL WAS CALCULATED FROM CORRECT TARGET REPLIES OF TEST DATASET.

Model	BLEU	ROUGE-L	Dist-2(%)	Ppl
Original	-	-	37.4	-
pre(text)	1.89	0.163	33.8	6.58
ft(text)	6.11	0.231	23.2	4.23
pre(speech)	1.38	0.142	27.3	3.39
ft(speech)	1.99	0.168	23.5	3.43

D. Objective evaluations

To evaluate the performance of the proposed model, we first conducted an objective evaluation. The objective measurements were BLEU, ROUGE-L, Distinct-2, and Perplexity (Ppl). Distinct measures the amount of the repetition and generic replies by calculating the proportion of unique n-grams in the generated replies. We used 2-gram (Dist-2). Ppl measures the model's fitness to the test dataset and is calculated from the probability when the sequences are generated. Note that Ppl of conventional text-dialogue models and that of the proposed models cannot be directly compared because the former is from the probability of SentencePiece tokens, and the latter is from quantized acoustic features.

We used the test dataset from FavoriteThingsChat for the evaluation. Since the output of the proposed models is discrete symbols obtained from continuous acoustic features, we cannot directly compare the performance between the proposed model and the conventional text-dialogue model. To avoid this problem, the objective measurements of the proposed model were obtained from the ASR transcriptions. We transcribed generated speech-replies from the proposed model using the Conformer-based ASR model [20] trained with the CSJ corpus. The character error rate was 8.0%, calculated from 500 sentences selected from the test dataset by comparing the output from ASR and transcriptions by an annotator.

Table I lists the results of the objective evaluations, where pre(text/speech) denotes pre-training models(text/speech), and ft(text/speech) denotes fine-tuned models(text/speech). Original was calculated from the correct target replies of the test dataset. The results of the pre-trained models (pre (text) and pre (speech)) indicate that the proposed model scored close to the conventional model in objective evaluations. In other words, our proposed model can directly generate speech-reply from the text-dialogue context. Next, the results of fine-tuning models (ft (text) and ft (speech)) indicate that BLEU and ROUGE-L improved after fine-tuning in both the conventional and proposed models. This suggests that the conventional and proposed models can acquire the dialogue domain from the fine-tuning dataset. However, Distinct was lower after fine-tuning. The domain specificity of fine-tuned models could result in lower diversity in generated speech.

TABLE II
PREFERENCE TEST BETWEEN GENERATED REPLIES FROM CONVENTIONAL TEXT-DIALOGUE AND PROPOSED MODELS FOR 250 REPLIES.

Proposed model vs conventional model	Win	Tie	Lose
	19.2%	25.6%	55.2%

E. Subjective evaluations

We conducted subjective evaluations on generated replies to confirm their qualities. Objective evaluations are known to be suboptimal for dialogue generation [21]. We conducted three evaluations: a preference test between the models, SSI (sensibleness, sensitivity, and interestingness) evaluation [1], [22] on generated replies, and a listening test to measure the naturalness of generated speech. The fine-tuned text and speech models were evaluated using test data from FavoriteThingsChat.

1) Preference test: We conducted the preference test to compare the naturalness of replies. The test data were 250 dialogues randomly chosen from the test dataset. To exclude the effect of ASR errors, an annotator transcribed generated speech-replies from the proposed model. The participants were three native Japanese who did not know which model the reply was from. Each was presented with the text-dialogue context and replies from both models, and evaluated on the basis of win (reply from the proposed model is better), lose (reply from the conventional text-dialogue model is better), and tie (both models are equally good or bad).

Table II lists the results of the preference test. The replies generated from the conventional text-dialogue model were more preferred. However, 44.8% (19.2% + 25.6%) of the replies generated from the proposed model scored the same or higher in naturalness. This indicates that the proposed model generates fair-quality replies, comparable to the conventional model in half the cases.

2) SSI test: We conducted the SSI evaluation to measure the more detailed quality of the generated replies. Sensibleness measures whether a model's replies make sense in context and do not contradict anything said earlier. Specificity measures whether a model's replies are specific to the dialogue context. Interestingness measures the attractiveness of the replies. The test data were 500 cases (250 from the conventional text-dialogue model and 250 from the proposed model) the same chosen from the test dataset. The same three participants for the preference test independently scored the SSI of the dialogue pairs on a scale of 5 (good sensibleness/specificity/interestingness) to 1 (no sensibleness/specificity/interestingness).

Table III lists the results of the SSI evaluation. The replies from the conventional text-dialogue model had higher scores in SSI, the same as in the preference test. Obtained scores of specificity and interestingness would be affected by sensibleness. To exclude this effect, we also calculated the scores of replies with sensibleness scores higher than 2. These scores are in brackets. The gap in sensibleness is smaller but still exists in specificity and interestingness. Therefore, we can explain why the replies from the conventional text-dialogue model are preferred with their simpler contents and less interestingness as well as the grammatical or contextual errors.

TABLE III
SSI TEST FOR GENERATED TARGET REPLIES FROM CONVENTIONAL TEXT-DIALOGUE AND PROPOSED MODELS. SCORES IN BRACKETS ARE SCORES EXCLUDING LOW SENSIBLE REPLIES.

Model	Sensibleness	Specificity	Interestingness
Text model	3.51 (4.10)	2.88 (3.24)	2.52 (2.86)
Proposed model	2.86 (3.78)	2.30 (2.82)	1.96 (2.40)

TABLE IV
NATURALNESS SCORES WITH CONFIDENCE INTERVAL 95%.

Model	MOS-naturalness
text-to-speech	3.98 ± 0.15
text-to-speech via quantization	2.99 ± 0.15
proposed	3.05 ± 0.14

3) Naturalness of synthesized speech: Finally, we conducted a MOS (mean opinion score) test to evaluate the naturalness of the speech samples generated from the proposed model. We also prepared two cases to assess (1)a gap in speech quality between TTS and the proposed model and (2) degradation in speech quality in vector quantization. The synthesized speech from the TTS described in Sec. IV-A2 is for the first case, and the synthesized speech from the TTS via vector quantization is for the second case. Manual transcriptions obtained from speech-replies of the proposed model were used for the TTS input to exclude the effect of the speech content. The participants were 11 native Japanese. Each participant evaluated 20 speeches under 3 conditions and rated their naturalness on a scale from 5 (very natural) to 1 (very unnatural). Table IV lists MOS scores of synthesized speech. The naturalness of the proposed model and that of the TTS via quantization was comparable, although that of the TTS without vector quantization is the best. This indicates that the dominant factor of this degradation is the vector quantization of acoustic features. In other words, the proposed model can successfully reproduce the nature of the training data in terms of speech quality.

TABLE V
DIALOGUE CONTEXT AND GENERATED TOKENS OF ”そうなんですね(I SEE)”. QA IS INDEX OF QUANTIZED ACOUSTIC FEATURES, AND SP IS THAT OF SENTENCEPIECE FOR TRANSCRIBED TEXT.

Case1	
context	個人的に北海道の旭山動物園が好きです (Personally, I like Asahiyama Zoo in Hokkaido.)
transcribed	そうなんですね。私も大好きです (I see. I love it too.)
tokens(SP)	16644
tokens(QA)	1952 1459 339 3295 3477 3733 2188 1740 2669 962 1908 2286 3931 747 915 132 3828 1985 3745 964
Case2	
context	食器入れて食洗器用の洗剤を指定の場所に入れてボタン押すだけでいいんですよ! (All you have to do is put the dishes in, put the dish detergent in the designated place and push the button!)
transcribed	そうなんですね。ありがとうございます (I see. Thank you.)
tokens(SP)	16644
tokens(QA)	928 4019 339 3173 3733 2197 2188 133 621 3236 2254 457 1403 2347 3347 3444 2046 3273 1697 3524

V. Discussion

Objective and subjective evaluations showed that the proposed model could directly generate replies from text-dialogue contexts. However, it was equal to that of the conventional text-dialogue model. To explore this, we analyzed generated sequences from two points; the length of generated sequence and diversity of expression.

Regarding the first point, Table V (Case 1) shows the transcribed text (transcribed), SentencePiece for the transcribed text (tokens(SP)), and tokens of generated quantized acoustic features (tokens(QA)) obtained from the reply "そうなんですね ("I see." in Japanese). We can see that just a single SentencePiece token expresses the reply with 20 tokens of quantized acoustic features. The tokens from the conventional text-dialogue model are chunks of characters, but those of the proposed model are shorter chunks representing 50 ms (four frames of 12.5 ms). The generated sequences from the proposed model are longer than the text-dialogue model. This could lead to more difficult reply generation, similar to the generation of acoustic features being more difficult than that of phoneme sequence in image-to-speech tasks [11].

Regarding the second point, we compared quantized acoustic feature sequences expressing the same speech content. Cases 1 and 2 in Table V show text-dialogue contexts and generated sequences. We observed that different contexts could generate speeches transcribed into the same SentencePiece but have a different quantized acoustic feature sequence. One reason for this would be the diversity of prosody such as F0 and speaking rate in acoustic features. Since a melspectrogram, which we used as an acoustic feature, includes not only speech content but also pitch (F0) information, speech with different prosody has different acoustic features. As a result, generated quantized acoustic feature sequences differ while their speech content is the same. These reasons could make training the proposed model more difficult and result in lower quality in generated replies compared with the conventional text-dialogue model. One promising approach to solve this problem would be acquiring better discrete representations than the simple LBG-based clustering. For example, representation such as disentangled representations for speech content, prosodic information, and speaker identity proposed by Polyak et al. [23] could lessen the effect of prosody and lead to constructing a better speech-dialogue model. Using better representation also showed better performance in direct speech-to-speech translation [12].

VI. Conclusions

We proposed a novel approach that generates speech-reply directly from text-dialogue context, bypassing text and phonemes. From the results of objective and subjective evaluations, the proposed model can successfully generate speech-replies directly from text-dialogue contexts only. However, the quality gap between the proposed model and a conventional text-dialogue model from the viewpoint of sensibleness, specificity, and interestingness still exists.

For future work, we will explore fine-tuning the proposed pre-trained model using natural speech-dialogue data, although we used only generated speech-dialogue data by TTS in this study. We will also apply other clustering approaches to obtain better discrete symbols disentangling the speech content and prosody. Construction of an end-to-end speech-dialogue model similar to speech-to-speech translation [12] is also for future work.

References

[1] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *Arxiv e-prints*, 2020.

[2] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in *EACL*, 2021, pp. 300–325.

[3] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "PLATO: Pre-trained dialogue generation model with discrete latent variable," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 85–96.

[4] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An end-to-end conversational style matching agent," in *Proceedings of the 19th International Conference on Intelligent Virtual Agents*, 2019, pp. 111–118.

[5] J. Fraser, I. Papaioannou, and O. Lemon, "Spoken conversational ai in video games: Emotional dialogue management increases user engagement," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 2018, pp. 179–184.

[6] M. Yamanaka, Y. Chiba, T. Nose, and A. Ito, "A study on a spoken dialogue system with cooperative emotional speech synthesis using acoustic and linguistic information," in *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. Springer, 2018, pp. 101–108.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[8] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.

[9] J. Effendi, S. Sakti, and S. Nakamura, "End-to-End image-to-speech generation for untranscribed unknown languages," *IEEE Access*, vol. 9, pp. 55 144–55 154, 2021.

[10] X. Wang, S. Feng, J. Zhu, M. Hasegawa-Johnson, and O. Scharenborg, "Show and speak: Directly synthesize spoken description of images," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4190–4194.

[11] X. Wang, J. Van Der Hout, J. Zhu, M. Hasegawa-Johnson, and O. Scharenborg, "Synthesizing spoken descriptions of images," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3242–3254, 2021.

[12] A. Lee, P.-J. Chen, C. Wang, J. Gu, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W.-N. Hsu, "Direct speech-to-speech translation with discrete units," *Arxiv e-prints*, 2021.

[13] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model," in *INTERSPEECH 2019*, 2019, pp. 1123–1127.

[14] T. Kano, S. Sakti, and S. Nakamura, "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 958–965.

[15] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2018, pp. 66–71.

[16] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.

[17] H. Sugiyama, M. Mizukami, T. Arimoto, H. Narimatsu, Y. Chiba, H. Nakajima, and T. Meguro, "Empirical analysis of training strategies of transformer-based japanese chit-chat systems," *Arxiv e-prints*, 2021.

[18] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[20] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.

[21] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *EMNLP*, 2016.

[22] A. D. Cohen, A. Roberts, A. Molina, A. Butryna, A. Jin, A. Kulshreshtha, B. Hutchinson, B. Zevenbergen, B. H. Aguera-Arcas, C. ching Chang, C. Cui, C. Du, D. D. F. Adiwardana, D. Chen, D. D. Lepikhin, E. H. Chi, E. Hoffman-John, H.-T. Cheng, H. Lee, I. Krivokon, J. Qin, J. Hall, J. Fenton, J. Soraker, K. Meier-Hellstern, K. Olson, L. M. Aroyo, M. P. Bosma, M. J. Pickett, M. A. Menegali, M. Croak, M. Díaz, M. Lamm, M. Krikun, M. R. Morris, N. Shazeer, Q. V. Le, R. Bernstein, R. Rajakumar, R. Kurzweil, R. Thoppilan, S. Zheng, T. Bos, T. Duke, T. Doshi, V. Prabhakaran, W. Rusch, Y. Li, Y. Huang, Y. Zhou, Y. Xu, and Z. Chen, "LaMDA: Language models for dialog applications," *Arxiv e-prints*, 2022.

[23] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *INTERSPEECH 2021*, 2021, pp. 3615–3619.