MTP-S2UT: ENHANCING SPEECH-TO-SPEECH TRANSLATION QUALITY WITH MULTI-TOKEN PREDICTION

Jianjin Wang[1⋆] *Runsong Zhao*[1⋆] *Xiaoqian Liu*[1] *Yuan Ge*[1] *Ziqiang Xu*[1]
Tong Xiao[1,2†] *Shengxiang Gao*[3] *Zhengtao Yu*[3] *Jingbo Zhu*[1,2]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2] NiuTrans Research, Shenyang, China
[3] Kunming University of Science and Technology, Kunming, China

ABSTRACT

Current direct speech-to-speech translation methods predominantly employ speech tokens as intermediate representations. However, a single speech token is not dense in semantics, so we generally need multiple tokens to express a complete semantic unit. To address this limitation, we introduce multi-token prediction (MTP) loss into speech-to-unit translation (S2UT) models, enabling models to predict multiple subsequent tokens at each position, thereby capturing more complete semantics and enhancing information density per position. Initial MTP implementations apply the loss at the final layer, which improves output representation but initiates information enrichment too late. We hypothesize that advancing the information enrichment process to intermediate layers can achieve earlier and more effective enhancement of hidden representation. Consequently, we propose MTP-S2UT loss, applying MTP loss to hidden representation where CTC loss is computed. Experiments demonstrate that all MTP loss variants consistently improve the quality of S2UT translation, with MTP-S2UT achieving the best performance.

***Index Terms*—** Speech-to-Speech Translation, Discrete Speech Tokens, Speech-to-Unit Translation, Multi-Token Prediction

1. INTRODUCTION

Direct speech-to-speech translation [1] converts source language speech into target language speech, showing promising applications in international conferences, cross-border communication, and travel scenarios. Recent advances in direct speech-to-speech translation with speech tokens (also known as discrete units) have emerged as a dominant research direction [2, 3, 4, 5].

The speech-to-unit translation (S2UT) model [2] serves as a representative architecture within this paradigm. As shown in Fig. 1a, it quantizes target speech into discrete speech tokens via a speech tokenizer [2, 6, 7], performs continuous source speech to discrete target speech tokens conversion using sequence-to-sequence models, and synthesizes target speech from these tokens through a detokenizer.

Compared to text tokens, speech tokens exhibit sparse semantic representations, typically requiring multiple speech tokens to express a single semantic concept [8]. This leads to higher prediction entropy and increases modeling complexity [8]. Multi-token prediction (MTP) offers a promising solution to mitigate this challenge by predicting multiple tokens at each position, allowing these tokens to collectively form complete semantic units. MTP [9] is originally proposed as an auxiliary task in large language models, as illustrated in Fig. 1b. The approach employs multiple output heads to predict multiple future tokens in parallel, thereby enhancing representational capacity and accelerating inference. Subsequently, DeepSeek-V3 [10] further refines the MTP module, as shown in Fig. 1c, by introducing additional inputs and Transformer blocks with a focus on improving model performance. VocalNet [8] pioneers the application of MTP in spoken dialogue models [11], as depicted in Fig. 1d. Building upon DeepSeek-V3's MTP design, VocalNet removes the additional inputs to prevent MTP from degenerating into next-token prediction (NTP), thereby effectively alleviating error propagation in speech token prediction while better capturing local patterns [8].

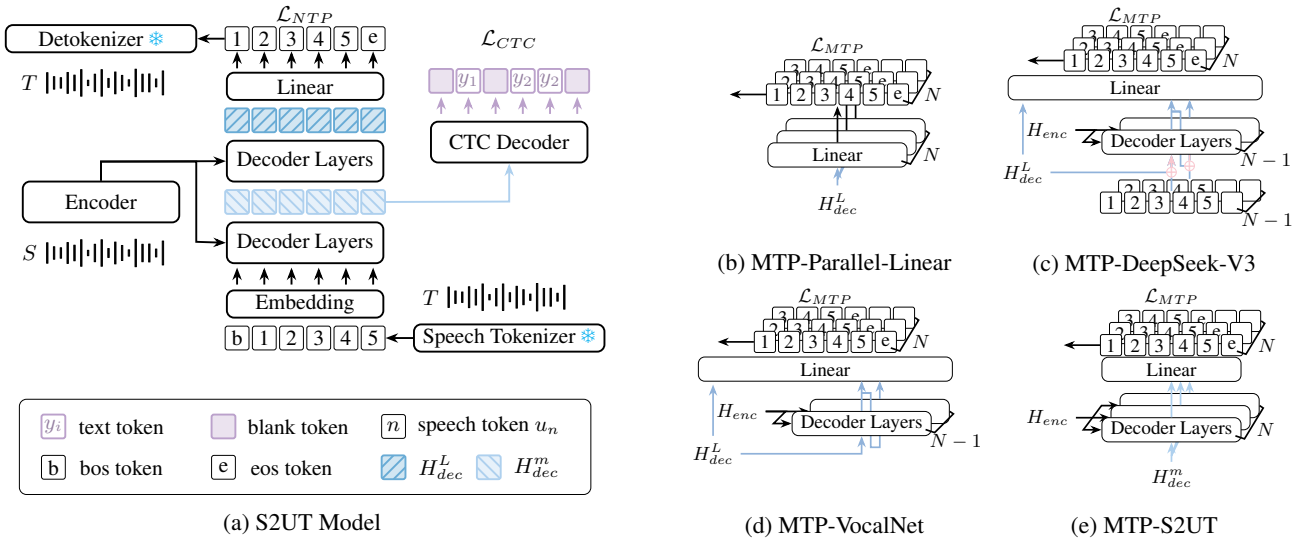
Our work is the first to introduce the MTP loss variants into the S2UT framework. As shown in Fig. 2, experiments reveal that MTP influences CTC-based text decoding, causing forward shifting of text tokens and backward shifting of blank tokens in CTC alignments, which indicates a forward semantic drift in hidden representations. Since a semantic unit typically corresponds to multiple speech tokens and each token can only access semantic information from preceding tokens, the early determination of semantic information helps reduce the uncertainty in predicting the corresponding multiple speech tokens. The hidden layer where CTC loss is computed incorporates both speech and text information and exhibits a significant shifting phenomenon. Therefore, we consider this layer as critical and propose the MTP-S2UT loss, as illustrated in Fig. 1e. We apply the MTP loss to this layer to encourage the model to fuse information from multiple speech tokens and text earlier in the intermediate layers, thereby enhancing the hidden representations in the shallower layers. Results in Section 3.4 show significant improvements on French→English speech translation across all variants, with MTP-S2UT yielding the most pronounced gains, consistent across speech tokenizers and on Spanish→English tasks. This validates that MTP strengthens the hidden representations of speech tokens, with earlier application leading to superior performance. Further analyses in Sections 3.5.1 and 3.5.2 reveal that MTP loss guides text token forward shifting and reduces the uncertainty in speech token prediction, with this phenomenon being particularly pronounced in MTP-S2UT.

Our work validates the effectiveness of MTP loss within the S2UT framework and demonstrates that applying it to the CTC hidden layer further enhances speech translation performance, providing novel insights into the role of MTP in speech translation.

2. METHOD

This section is structured as follows: Section 2.1 reviews the S2UT model. Section 2.2 details the adaptation of existing MTP losses to

⋆ Equal contribution.
† Corresponding author.



(a) S2UT Model (b) MTP-Parallel-Linear (c) MTP-DeepSeek-V3 (d) MTP-VocalNet (e) MTP-S2UT

Fig. 1. Overview of S2UT model and our implementation of 4 MTP loss variants on the S2UT model.

S2UT. Our MTP-S2UT loss is finally introduced in Section 2.3.

2.1. S2UT

A speech-to-speech translation dataset is typically composed of quadruplets in the form of $D = \{(S, X, Y, T)\}$. Here, $S = (s_1, \cdots, s_{|S|})$ represents the source speech, $X = (x_1, \cdots, x_{|X|})$ is the source text, $Y = (y_1, \cdots, y_{|Y|})$ is the target text, and $T = (t_1, \cdots, t_{|T|})$ is the target speech. As shown in Fig. 1a, in the S2UT [2] model, the continuous target speech waveform T is quantized into a sequence of speech tokens $U = (u_1, \cdots, u_{|U|}, e)$ by the speech tokenizer, where e denotes the end-of-sequence (eos) token. The encoder will encode the source speech S into a sequence of hidden states H_{enc}.

Subsequently, the encoded source speech representations H_{enc} and right-shifted sequence $U^{+1} = (b, u_1, \cdots, u_{|U|})$ are fed into the decoder to predict U via cross-attention mechanisms, where b denotes the begin-of-sequence (bos) token:

$$H^0_{dec} = \text{Emb}(U^{+1}) \tag{1}$$

$$H^i_{dec} = \text{DecoderLayer}^i(H_{enc}, H^{i-1}_{dec}) \tag{2}$$

$$\mathcal{L}_{NTP} = -\log P(U|H^L_{dec}) \tag{3}$$

$$\mathcal{L}_{S2UT} = \mathcal{L}_{NTP} + \mathcal{L}_{other} \tag{4}$$

where L is the number of decoder layers and S2UT model employs multi-task training with $\mathcal{L}_{NTP}$ representing the next-token prediction loss, $\mathcal{L}_{other}$ capturing other auxiliary task losses, and $\mathcal{L}_{S2UT}$ as the final combined training objective.

2.2. Multi-token Prediction

Prior works [8, 9, 10] employ MTP loss $\mathcal{L}_{MTP}$ to replace the NTP loss $\mathcal{L}_{NTP}$, applying it exclusively to the final layer hidden representations H^L_{dec}. Under this paradigm, each token position is required to predict the subsequent N tokens:

$$\mathcal{L}_{MTP} = -\sum_{k=0}^{N-1} \log P(U^{-k}|H^L_{dec}) \tag{5}$$

where U^{-k} denotes the sequence U left-shifted by k positions, with insufficient positions padded accordingly.

The difference between the loss functions of the previous three works lies in the modeling of $P(U^{-k}|H^L_{dec})$. As shown in Fig. 1b, **MTP-Parallel-Linear** [9] uses N independent linear heads:

$$P(U^{-k}|H^L_{dec}) = \text{softmax}(W^k H^L_{dec}) \tag{6}$$

As shown in Fig. 1c, **MTP-DeepSeek-V3** [10] employs MTP with teacher-forcing and Transformer blocks. Given that S2UT takes additional H_{enc} as input, we implement MTP-DeepSeek-V3 as follows:

$$H^0_{out} = H^L_{dec} \tag{7}$$

$$H^k_{in} = W^k_{in}[\text{LN}(H^{k-1}_{out}); \text{LN}(\text{Emb}(U^{1-k}))] \tag{8}$$

$$H^k_{out} = \text{Decoder}^k(H_{enc}, H^k_{in}), k > 0 \tag{9}$$

$$P(U^{-k}|H^L_{dec}) = \text{softmax}(W_{out} H^k_{out}) \tag{10}$$

In Fig. 1c, the symbol $\oplus$ represents the fusion operation described in Equation 8, which concatenates the normalized H^{k-1}_{out} and the normalized embedding of U^{1-k} along the feature dimension and reduces the dimension to the original size through a linear layer W^k_{in}, where LN denotes layer normalization, and $[\cdot;\cdot]$ denotes concatenation along the feature dimension. Compared to MTP-Parallel-Linear, MTP-DeepSeek-V3 provides additional U^{1-k} and H_{enc} as inputs, making MTP simpler.

VocalNet [8] argues that using teacher-forcing to input real tokens makes the training objective no different from NTP, failing to alleviate error accumulation and unable to effectively capture local speech structures. Therefore, as shown in Fig. 1d, it removes $\text{Emb}(U^{1-k})$ from MTP-DeepSeek-V3:

$$H^k_{in} = H^{k-1}_{out} \tag{11}$$

It is worth noting that although MTP loss introduces many parameters, they can all be discarded during inference, thus not affecting the time and space consumption during inference.

2.3. MTP-S2UT

To achieve higher translation quality, the S2UT model requires additional text supervision signals to enhance the intermediate representation H^m_{dec}. Connectionist Temporal Classification (CTC) [12] loss $\mathcal{L}_{CTC}$ is typically applied to the decoder intermediate hidden states H^m_{dec} to enhance the semantic information of their representations:

$$\mathcal{L}_{CTC} = -\log P(Y|H^m_{dec}) \tag{12}$$

where $1 \leq m \leq L$. As shown in Fig. 1e, we believe that H^m_{dec} is rich in both textual and speech modal information, therefore applying the MTP loss to H^m_{dec} would be more effective:

$$\mathcal{L}_{MTP-S2UT} = -\sum_{k=0}^{N-1} \log P(U^{-k}|H^m_{dec}) \tag{13}$$

The modeling for $P(U^{-k}|H^m_{dec})$ is as follows:

$$H^k_{out} = \text{Decoder}^k(H_{enc}, H^m_{dec}) \tag{14}$$

$$P(U^{-k}|H^m_{dec}) = \text{softmax}(W_{out}H^k_{out}) \tag{15}$$

This facilitates earlier integration of cross-modal information, enhancing the semantic density of the intermediate representations.

3. EXPERIMENTS

3.1. Data

3.1.1. Dataset

We conduct our experiments on the CVSS-C benchmark [13], a large-scale speech-to-speech translation dataset. We evaluate the MTP loss variants applied to S2UT on the French→English (Fr→En) and Spanish→English (Es→En) translation tasks.

3.1.2. Pre-processing

For the source speech, 80-dimensional mel-filterbank features [14] are computed, followed by global cepstral mean and variance normalization. For the target speech, three tokenizers are evaluated. The first is an unsupervised tokenizer based on k-means clustering (k=1000) applied to mHuBERT features [15]; the resulting discrete units are synthesized using a unit-based vocoder [2]. The other two are supervised tokenizers, namely the $\mathcal{S}^3$ tokenizer [6] and the GLM-4-Voice-Tokenizer [7], with codebook sizes of 6561 and 16384, respectively. For these tokenizers, speech reconstruction is performed in two stages: a flow matching model [16] first generates a mel-spectrogram from the tokens, which is then converted to waveform audio by a vocoder [17]. Finally, the source and target text are tokenized using SentencePiece [18], resulting in a unigram vocabulary of 6000 tokens for each.

3.2. Model settings

The S2UT model employs a 12-layer Conformer [19] encoder with a hidden feature dimension of 256. Two-layer Transformer decoders with identical hidden feature dimensions are connected after the 6th and 8th encoder layers for multi-task learning of source and target language texts, with weights set to 8 for both. The decoder is a Transformer decoder with 6 layers and a hidden feature dimension of 512. A CTC decoder is attached after the 3rd decoder layer for multi-task learning of target language text, with a weight of 1.6.

Tokenizer	Model	Greedy	Beam5	Beam10
$\mathcal{S}^3$ tokenizer	S2UT	17.79	18.98	19.15
	+ MTP-Parallel-Linear	21.34	22.40	22.52
	+ MTP-DeepSeek-V3	23.38	24.25	24.31
	+ MTP-VocalNet	23.29	24.17	24.27
	+ MTP-S2UT	**24.36**	**25.14**	**25.16**
HuBERT with K-means	S2UT	22.02	23.11	23.33
	+ MTP-Parallel-Linear	22.03	23.07	23.10
	+ MTP-DeepSeek-V3	22.73	23.86	23.87
	+ MTP-VocalNet	22.11	23.37	23.60
	+ MTP-S2UT	**23.59**	**24.50**	**24.53**
GLM-4-Voice-Tokenizer	S2UT	21.62	23.08	23.26
	+ MTP-Parallel-Linear	21.92	23.36	23.56
	+ MTP-DeepSeek-V3	22.99	24.27	24.45
	+ MTP-VocalNet	23.55	24.99	25.20
	+ MTP-S2UT	**23.97**	**25.22**	**25.26**

Table 1. ASR-BLEU scores with three speech tokenizers on CVSS-C Fr→En test set under different MTP loss variants.

Model	Greedy	Beam5	Beam10
S2UT	16.67	17.99	18.18
+ MTP-Parallel-Linear	16.83	18.35	18.58
+ MTP-DeepSeek-V3	18.94	20.14	20.31
+ MTP-VocalNet	19.98	21.47	21.69
+ MTP-S2UT	**21.87**	**22.59**	**22.83**

Table 2. ASR-BLEU scores with $\mathcal{S}^3$ tokenizer on CVSS-C Es→En test set under different MTP loss variants.

For the MTP configuration, each speech token predicts the subsequent $N = 7$ speech tokens. MTP-Parallel-Linear utilizes N independent linear layers, while the other three variants employ one shared linear layer and multiple independent decoders. MTP-DeepSeek-V3 additionally shares the word embedding layer of the main network. MTP-S2UT applies the MTP loss at the 3rd layer, which is also the same layer where the CTC loss is applied. In preliminary experiments, we observe that increasing the decoder layers of MTP-DeepSeek-V3 from 1 to 3 layers yields a 0.21 improvement in greedy search. For fair comparison, we uniformly set each decoder to 3 layers. The weight for MTP loss is set to 1.0.

3.3. Evaluation

We evaluate translation quality using ASR-BLEU. This metric is calculated by first transcribing the synthesized target speech into text with an Automatic Speech Recognition (ASR) model, and then computing the BLEU score between the transcription and the ground-truth reference text. For this purpose, we use the oct22 version of the ASR model in the fairseq [20] toolkit.

3.4. Main Results

We train S2UT models on the CVSS-C Fr→En dataset using four MTP loss variants. Table 1 shows that models trained with our MTP-S2UT loss variant achieve the best results across all tokenizers and decoding methods. With the $\mathcal{S}^3$ tokenizer and greedy search, the ASR-BLEU score of S2UT improves from 17.79 to 24.36. Other MTP loss variants also bring gains, though smaller ones. This proves that MTP loss boosts the information in hidden states, which leads to better translation. The effect works best when we boost earlier hidden states.

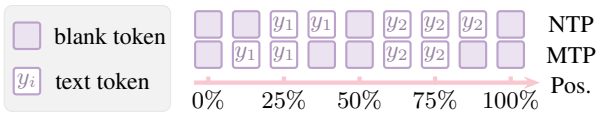


Fig. 2. Example of CTC output sequences decoded from the intermediate hidden states H^m_{dec}. Compared to NTP loss, the model trained with MTP loss produces text tokens y_1 that exhibit an overall forward shift. After each text token's first occurrence, subsequent tokens can access its semantic information, thus we use the first occurrence position to represent the earliest availability of its semantic information. For instance, the first occurrence positions of y_1 and y_2 from the model trained with MTP loss are 12.5% and 62.5%, respectively. Complete statistics are provided in Table 3.

Model	$\mathcal{S}^3$ tokenizer	HuBERT with K-means	GLM-4-Voice-Tokenizer
S2UT	51.011%	49.628%	50.363%
+ MTP-Parallel-Linear	47.725%	47.288%	49.601%
+ MTP-DeepSeek-V3	50.166%	49.267%	50.719%
+ MTP-VocalNet	47.504%	42.486%	48.889%
+ MTP-S2UT	47.382%	44.561%	43.889%

Table 3. Average relative position of the first occurrence of text tokens. If the average position is $> 50\%$, it indicates that text tokens generally appear later in the sequence relative to blank tokens.

We also test on the CVSS-C Es→En dataset. Table 2 shows results that match our Fr→En experiments, proving that MTP loss brings gains across different languages.

3.5. Analysis

3.5.1. CTC Decoding Forward Shift

In theory, MTP loss encourages early planning of future information. We can verify this hypothesis by performing CTC decoding on the intermediate hidden states H^m_{dec}. As shown in Fig. 2, models trained with MTP loss exhibit a forward shift in text tokens compared to those trained with NTP loss, indicating that the semantic information in hidden states is advanced along the sequence dimension.

To quantify this advancement, we calculate the relative position of the first occurrence of all text tokens within the entire sequence. We compute the average position of all text tokens, with results presented in Table 3. The results demonstrate that MTP loss significantly shifts text tokens forward, except for MTP-DeepSeek-V3. This exception occurs because MTP-DeepSeek-V3 utilizes additional input through teacher forcing, providing an extra information source that maintains relatively stable semantics. In contrast, other MTP loss variants drive semantic information to shift forward appropriately to achieve future token prediction.

We hypothesize that the forward shift of semantic information can reduce the model's uncertainty in predicting speech tokens to some extent. For example, *Hello* occupies one text token and multiple speech tokens. If the semantic representation of *Hello* appears at earlier positions among these speech tokens, the model can easily predict the speech tokens corresponding to *Hello*. Conversely, if the semantic representation of *Hello* appears at later positions among these speech tokens, the model exhibits higher uncertainty for the initial speech tokens, as it lacks knowledge about which semantic speech tokens to predict. We will measure the model's uncertainty regarding speech tokens in the next subsection.

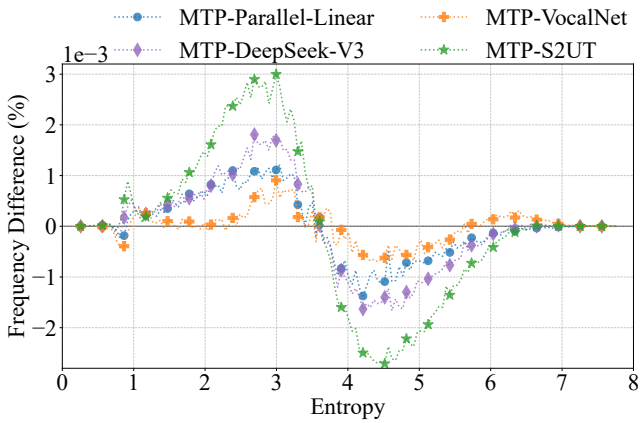


Fig. 3. Entropy distribution of 1.2M speech token predictions in S2UT models trained with MTP loss. Speech tokens are from the CVSS-C Fr→En test set using the $\mathcal{S}^3$ tokenizer. All frequencies are presented relative to the baseline (with NTP loss) for enhanced visualization clarity.

3.5.2. Speech Token Uncertainty

We employ entropy to measure the uncertainty of the model's predictions for speech tokens:

$$\text{Entropy}(\text{Token}) = -\sum_{z \in \mathcal{Z}} p(z) \log p(z) \tag{16}$$

where $\mathcal{Z}$ is the set of all possible tokens, and $p(z)$ is the probability that the token is z.

We compute the entropy of five models over 1.2 million tokens using the $\mathcal{S}^3$ tokenizer on the CVSS-C Fr→En test set. To better compare the impact of MTP loss on S2UT models, we subtract the frequency distribution of the S2UT baseline from the frequency distributions of the four MTP loss variants. As shown in Fig. 3, all MTP loss variants lead to increased frequency in low-entropy regions and decreased frequency in high-entropy regions for S2UT models, indicating that MTP loss indeed reduces uncertainty during the prediction process. Among these variants, our MTP-S2UT demonstrates the most significant and pronounced reduction in uncertainty.

4. CONCLUSIONS

This paper introduces multi-token prediction (MTP) into the S2UT framework and proposes a novel MTP-S2UT loss applied at the intermediate CTC layer, significantly enhancing translation quality by encouraging earlier and richer fusion of semantic information across speech and text modalities. Experimental results on Fr→En and Es→En tasks demonstrate consistent and substantial improvements across multiple speech tokenizers and decoding strategies, with MTP-S2UT achieving the highest gains. Our analysis reveals that MTP not only reduces predictive uncertainty for speech tokens but also induces a forward shift in CTC alignments, indicating more efficient semantic planning. This work validates the effectiveness of MTP in speech-to-speech translation and highlights the importance of early intermediate layer enrichment, paving the way for more powerful and efficient direct S2UT models in the future.

5. REFERENCES

[1] Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Interspeech 2019*, 2019, pp. 1123–1127.

[2] Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu, "Direct speech-to-speech translation with discrete units," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, Eds., Dublin, Ireland, May 2022, pp. 3327–3339, Association for Computational Linguistics.

[3] Hirofumi Inaguma, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, and Juan Pino, "UnitY: Two-pass direct speech-to-speech translation with discrete units," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, Eds., Toronto, Canada, July 2023, pp. 15655–15680, Association for Computational Linguistics.

[4] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al., "Seamlessm4t: massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[5] Shaolei Zhang, Qingkai Fang, Shoutao Guo, Zhengrui Ma, Min Zhang, and Yang Feng, "StreamSpeech: Simultaneous speech-to-speech translation with multi-task learning," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar, Eds., Bangkok, Thailand, Aug. 2024, pp. 8964–8986, Association for Computational Linguistics.

[6] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al., "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[7] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang, "Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot," *arXiv preprint arXiv:2412.02612*, 2024.

[8] Yuhao Wang, Heyang Liu, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, and Yu Wang, "Vocalnet: Speech llm with multi-token prediction for faster and high-quality generation," *arXiv preprint arXiv:2504.04060*, 2025.

[9] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve, "Better & faster large language models via multi-token prediction," in *Forty-first International Conference on Machine Learning*, 2024.

[10] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al., "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[11] Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al., "Wavchat: A survey of spoken dialogue models," *arXiv preprint arXiv:2411.13577*, 2024.

[12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[13] Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen, "CVSS corpus and massively multilingual speech-to-speech translation," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, Eds., Marseille, France, June 2022, pp. 6691–6703, European Language Resources Association.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.

[15] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[16] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023.

[17] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 17022–17033, Curran Associates, Inc.

[18] Taku Kudo and John 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*, Eduardo Blanco and Wei Lu, Eds., Brussels, Belgium, Nov. 2018, pp. 66–71, Association for Computational Linguistics.

[19] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020*, 2020, pp. 5036–5040.

[20] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.