When End-to-End is Overkill: Rethinking Cascaded Speech-to-Text Translation

Anonymous submission to Interspeech 2024

Abstract

Though end-to-end speech-to-text translation has been a great success, we argue that the cascaded speech-to-text translation model still has its place, which is usually criticized for the error propagation between automatic speech recognition (ASR) and machine translation (MT) models. In this paper, we explore the benefits of incorporating multiple candidates from ASR and self-supervised speech features into MT. Our analysis reveals that the primary cause of cascading errors stems from the increased divergence between similar samples in the speech domain when mapped to the text domain. By including multiple candidates and self-supervised speech features, our approach allows the machine translation model to choose the right words and ensure precise translation using various speech samples. This strategy minimizes error spread and takes advantage of large ASR and MT datasets, along with pre-trained ASR/MT models, while addressing associated issues.

Index Terms: speech-to-text translation,machine translation

1. Introduction

In recent years, the academic community has been intrigued by the rapid advancement of end-to-end speech-to-text translation models [1]. These efficient encoder-decoder architectures provide a direct avenue for translating speech, bypassing the need for complex intermediate symbolic representations. However, the arduous task of assembling and curating end-to-end data poses a significant challenge, entailing considerable costs and extensive efforts. These end-to-end methods need the careful selection of high-quality data or argumentation [2, 3, 4, 5, 6], encompassing both speech and translated transcripts, and the scrupulous exclusion of erroneous examples.

Nonetheless, cascaded speech-to-text translation models have encountered substantial criticism due to an intrinsic shortcoming called "cascaded loss" or "error propagation". Studies on ASR+MT systems have explored various methods to enhance the integration of ASR output lattices into MT models [7, 8, 9]. To mitigate error propagation, several approaches [10, 11, 12, 13, 14, 15, 16, 17] have been proposed to integrate ASR and MT models for end-to-end models, which necessitate the addition of supplementary modules and substantial additional training. In contrast, the method proposed in our research employs an n-best strategy that does not require additional parameters and can significantly enhance performance with minimal fine-tuning.

Recent studies [18, 19] have demonstrated the performance improvements achieved by scaling up pre-trained models for downstream natural language processing tasks. However, developing advanced speech translation models is a difficult task, requiring significant computational resources due to the enormous volume of training data and complex model intricacies.

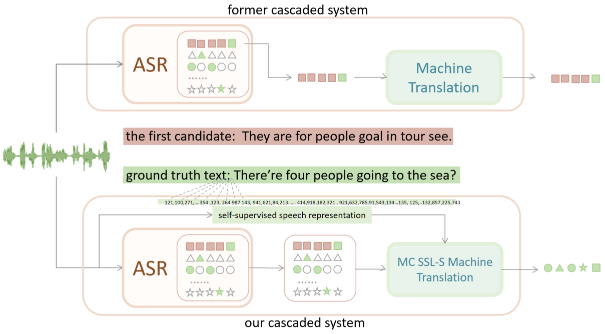

Figure 1: *The correct words are scattered among various candidates, while the former cascaded system directly selected the first candidate, resulting in similar pronunciation errors in the ASR output that are further propagated through the translation model, causing cascading losses.*

To fully exploit the potential of pre-trained MT and ASR models, we present a novel perspective on error propagation and the preservation of essential speech information. We propose the idea of utilizing multiple ASR candidates for machine translation, integrated with self-supervised speech representations, to enhance the accuracy of the translation. Our comprehensive analysis reveals the primary causes of error propagation in cascaded systems, which originate from the misalignment between the acoustic and semantic dimensions of speech. Factors such as homophones with different meanings and word elisions contribute to inaccuracies in ASR results, which consequently propagate to the machine translation model. Furthermore, we explore the use of self-supervised language representations to preserve fine-grained linguistic information in speech.

Our model has the following advantages:

1. Our model achieves the best performance among cascaded models in the Speech-to-Text(S2T) translation task.
2. Our method can leverage variously known ASR and machine translation (MT) pre-trained models. Specifically, it can effortlessly adapt to different model architectures without the need to adjust the model parameters.
3. Unlike end-to-end models, our approach does not require an extensive large amount of costly <speech, transcript, target>paired data.
4. Our model demonstrates rapid training speed and attains exceptional results with minimal data utilization for fine-tuning the Machine Translation (MT) model.

2. Analysis

To see where the error propagation lies in the cascaded system and how ASR errors propagate to MT, we pose two questions: 1) Can the top-ranked ASR candidate cover all the lexicons? 2) Is the top-ranked ASR candidate always the best translation result?

2.1. Preliminary experiments

In the cascaded system, we extract the top 20 results based on the scores from the ASR system and then select the top n candidates. We calculate the lexical overlap between these candidates $\{c_1, c_2 ... c_n\}$ and the ground truth text $\{gt\}$ of the ASR, which means the length of set of all words in candidates $\{w | w \in c_k, 1 \leq k \leq n\}$ intersected with $\{w | w \in gt\}$ divided by the length of the latter. In Table 1, "average" refers to the average lexical overlap between each candidate and the ground truth. At the same time "cumulative" represents the lexical overlap when considering a combination of n candidates with the ground truth. We observe that as n increases, although the average lexical overlap decreases, the cumulative overlap improves. This indicates that apart from the top-ranked candidate, the ASR system fails to include some vocabulary, which remains in the lower-ranked candidates.

Table 1: *Lexical overlap between ASR candidates and GT*

n **candidates**	**Average**	**Cumulative**
1	92.0%	92.0%
5	90.0%	94.3%
10	89.4%	95.0%
20	89.5%	95.5%

Additionally, we use a trained MT model to translate and calculate BLEU scores for the top 5 candidates based on ASR scores. We then analyze the index and percentage of the highest BLEU score and find that only 45.35% of the candidates with the highest BLEU score corresponded to the candidate with the lowest word error rate. It indicates that the top-ranked ASR candidate always performs the best translation result.

We also compare the BLEU score of the translation for the candidate with the lowest word error rate with the average BLEU score of the top 5 translations. The BLEU result of the translated best ASR candidate is 36.0, while the BLEU result of the translated first ASR candidate is 32.9. This indicates that combining multiple candidates can significantly improve translation results.

Table 2: *The proportion of the ASR result index of the best candidate based on the BLEU score of the translation result*

Best BLEU idx	1	2	3	4	5
percentage(%)	45.35	16.73	14.07	12.12	11.74

2.2. Source of cascade loss

Since the best result in the ASR stage does not necessarily represent the best result in an end-to-end manner, there must be a point where cascaded errors occur. Therefore, we delve deeper into analyzing these cascaded errors and make the hypothesis that the source of cascaded loss mainly arises from discrepancies between the pronunciation space and semantic space and that ASR models encounter difficulties in selecting results that most accurately match the language patterns.

ASR is trained using paired speech and text data, but the language patterns it captures are not as rich as those in translation models. Therefore, when the recognition results are similar, it is difficult for the ASR model to select the candidate that best aligns with human common sense and grammar based on scores alone.

As an example, consider the phrase "has put the race on the top." The highest-scoring candidate in ASR recognizes "race" as "rays," which contradicts common sense. However, subsequent candidates include various results with similar pronunciations, such as "race," "raised," and "raise." Another example is "Recording the transaction in an immutable distributed ledger," while the highest-scoring result in ASR is "Recording the transaction in an immutable distributed lecture," which is not a common expression. The subsequent candidates include "legend," "literature," "letter," "ledger," and other results.

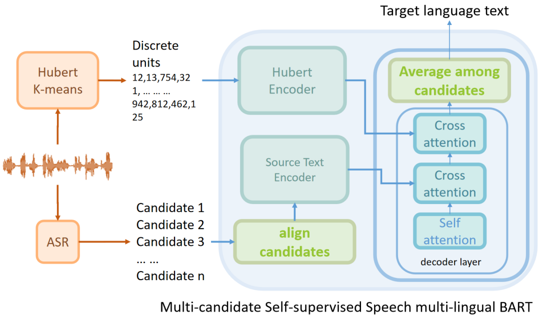

Figure 2: *Overview of our proposed MC-sslS system*

3. Method

To leverage the powerful capability of machine translation in capturing semantic patterns, we propose utilizing multi-candidate ASR inputs and averaging attention computation in the MT model. Furthermore, to address the issue of error propagation in ASR caused by homophones, we employ self-supervised speech representations to enhance accuracy. The combination of methods we propose is represented by the model depicted in Figure 2.

3.1. Multi-candidate from ASR

The correspondence between the speech domain and the semantic domain is not always perfect. However, compared to machine translation models, ASR models have limited ability to capture semantic patterns effectively. Consequently, ASR is prone to mapping speech samples that are acoustically similar to semantically distant outputs. Moreover, ASR cannot selectively choose samples that align more closely with human language conventions and patterns. For a given speech input, the correct vocabulary might be dispersed among multiple candidates generated during beam search. Unfortunately, previous cascaded systems only consider the top-scoring candidate and pass it to the machine translation model, resulting in error propagation within the cascaded system. Our cascaded model first uses Wenet [20] to perform ASR on GigaST [21]; we store the top 20 ASR result sentences based on the cumulative log probability value from the end of beam search.

3.1.1. *Aligned by common substrings*

We follow the following steps to align the candidates:

1. The top n ranked texts $t_k (1 \leq k \leq n)$ consisting of n_k words $(w_1, w_2 .... w_{n_k})$ are selected (in our following experiments, we set n to be 5).
2. The dynamic programming algorithm for finding the longest common substrings [22] is used for $n-1$ times, in the preprocessing stage before putting into the model. There are $n-1$ processes in total. We denote t_i^{m+1} as the aligned result of t_i after the $m^{th} (1 \leq m \leq n-1)$ process. t_1^1 is equal to t_1.
3. During the m^{th} process, (1) we calculate the longest common subsequences of $\{t_1^m, t_{m+1}\}$, and get $\{t_1^{m+1}, t_{m+1}^{m+1}\}$ by connecting the substrings and the largest length of uncommon substrings. (2) The common substrings are aligned, and the remaining parts are padded with "unk" tokens. (3) When $2 \leq m \leq n-1$, the $\{t_2^m ... t_{m+1}^m\}$ are padded with "unk" tokens at the same indexes where $\{t_1^m\}$ is padded to $\{t_1^{m+1}\}$. Then, one process ends.
4. After $n-1$ processes, the aligned and padded texts of the same length are used as input to the attention-based machine translation model.

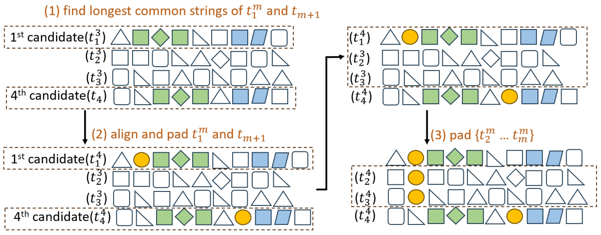

Figure 3: *Above is how the 3^{rd} process is calculated. After finding the longest common subsequences, candidates are aligned and padded. The orange circles denote "unk" tokens.*

3.1.2. *Average attention among candidates*

Thus, we propose an innovative approach, incorporating multiple ASR candidates into a single attention-based machine translation model. Here we use mBART as the backbone architecture. The multi-candidate model can share identical parameters with the translation model, differing only in its attention-averaging technique.

During the training stage, this method averages attention to get $\mathrm{A}'(Q)$ at the sentence level for multiple input candidates corresponding to the same translation result. The average calculation is performed only once after all decoder layers, just before the final layer normalization, and then assign $\mathrm{A}'(Q)$ to $\mathrm{A}_{candidate_i}(Q)$. We find that calculating only once at the final layer yields the best results.

$$\mathrm{A}_{candidate_i}(Q) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

$$\mathrm{A}'(Q) = \frac{1}{n_{\text{candidate}}} \sum_{i=1}^{n_{\text{candidate}}} \mathrm{A}_{candidate_i}(Q) \tag{2}$$

During the inference stage, the attention mechanism is utilized to compute the average attention score for beams that share the same candidate sequence number. This aggregation of attention scores forms a pool of candidate beams. Consider there are m beams in total. Each time predicting the next token, the n candidates generate mn results in total. For each candidate, there are m beams, and the $\mathrm{A}_{beam_k}(Q)$ is calculated as follows. The calculation of attention occurs at the same position and training stage, after all decoder layers, just before the final layer normalization. Attention is calculated once for each token generated. Subsequently, the other calculations are the same as regular beam searches.

$$\mathrm{A}_{beam_k}(Q) = \frac{1}{n_{\text{candidate}}} \sum_{i=1}^{n_{\text{candidate}}} \mathrm{A}_{candidate_i\_beam_k}(Q) \tag{3}$$

It's worth noting that this does not add any parameters to the MT model and can be used without further training, compared with [7, 8, 9]. It is a lightweight method that can be readily adapted to various attention-based architectures.

3.2. Acoustic and linguistic features fusion

The process of converting audio recognition to text and then to machine translation further results in the loss of a large amount of acoustic information in speech. We propose a multi-candidate self-supervised learning speech machine translation model(shown in Figure 2) to enhance the accuracy of filtering correct results by leveraging language patterns from machine translation and fusing acoustic and linguistic features, following the approach utilized in previous works [23] and utilizing HuBERT [24] for generating target self-supervised discrete units. This choice was influenced by the superior performance demonstrated by HuBERT in various tasks such as ASR, spoken language modeling, and speech synthesis, as shown in by [25, 26, 27]. It outperforms other unsupervised representations, including VAE-based representations employed in [28, 29].

We follow [23] to use the 11^{th} layer of HuBERT as input, which contains richer linguistic information, and transform them into tokens with a word list of 1000 for the number of speech units using the K-means model trained on English speech. Due to the high length of the original unit, we reduced the consecutive repetitive units to a single unit. We create a new trainable vocabulary for word embeddings. After being encoded, it interacts with the decoder's cross-attention to obtain deeper language information. The calculation process is the same as that of how the output of the source text encoder computes cross-attention with the production of self-attention. The rest of the process is similar to that of the multi-candidate machine translation model.

4. Experiments & Results

4.1. Data

For ST datasets, we use GigaST [21], a large-scale pseudo speech translation (ST) corpus containing 7.5M en-zh pairs. It is created by translating the text in GigaSpeech [30], an English ASR corpus, into German and Chinese. The training set is translated by a robust machine translation system, and the test set is translated by humans.

4.2. Model setup

4.2.1. *Cascaded system*

Our cascaded model first uses Wenet [20] to perform ASR on Gigaspeech with 7.4M valid en-zh pairs, and we store the speech recognition texts corresponding to the top 20 cumulative log probability value rankings at the end of beam search. The top n ranked texts are input as encoders into the mBART-based machine translation model. We do not use any self-supervised speech models to initialize the ASR encoder. The encoder is trained from scratch following the WeNet paper. Taking inspiration from the approach proposed by [23], We use the multilingual HuBERT (mHuBERT) model and K-means model to encode source speech into a vocabulary of 1000 units. The mHuBERT and the K-means models are learned from the combination of English, Spanish, and French unlabeled speech data from VoxPopuli [31], while we use them to encode English speech only.

4.2.2. Multi-candidate mBART

The multi-candidate mBART is built upon the mBART model [32], which is an open-sourced sequence-to-sequence denoising auto-encoder pre-trained on extensive monolingual corpora from multiple languages using the BART objective [33]. The training of mBART involves the application of BART to large-scale monolingual corpora across various languages.

We fine-tune the mBART model on 8 A100 GPUs. Our model configuration follows[1]to use the same parameters for fine-tuning the mBART model.

On the GigaST dataset, we compare our method with the original cascaded system and two end-to-end ST systems, including SSL-Transformer and Speech-Transformer [21].

Table 3: *Giga-ST main results: "MC" denotes the method of multi-candidate in Section 3.1.2, "Alignment" denotes the method in Section 3.1.1, "sslS" denotes the method of fusing acoustic and linguistic features in Section 3.2*

Settings	Models	BLEU score
(1)	Machine translation	40.2
End-to-End Model		
(2)	SpeechTransformer	36.3
(3)	SSL-Transformer	38.0
Cascaded Model		
(4)	Wenet + mBART	36.8
(5)	(4) + MC	36.9
(6)	(4) + MC + Alignment	37.8
(7)	(4) + MC + Alignment + sslS	38.1

4.3. Results

4.3.1. Comparison with baselines

The following experiments demonstrate the effectiveness of our approach.

1. Comparing settings (6) and (4), it shows that the Alignment and Multi-candidate methods enhance the performance of the conventional cascaded system in setting (4).
2. When comparing settings (7), (6), and (3), it shows that fusing acoustic and linguistic features improves the performance of the conventional cascaded system in setting (4), making it comparable to the end-to-end model.

Moreover, while the transition from high-quality machine translation models to improved multi-candidate translation models requires no parameter changes, in the era of emerging and popular large language models, it is evident that our approach holds more excellent practical value and prospects than the End-to-End model. Furthermore, our model achieves a BLEU score of 37.3 after just one epoch of training, requiring only one hour on our settings. With a concise duration of fine-tuning, it surpasses the performance of former cascaded models, demonstrating the effectiveness of our approach.

In machine translation, for instance, mBART [32] can be pre-trained by denoising complete text data from various languages. It is capable of transferring knowledge to language pairs without parallel text or those not included in the pre-training corpus. Languages that are not part of the pre-training corpus can benefit from machine translation, which strongly suggests that the initialization process is, to some extent, language-agnostic. Pre-training can capture common patterns in text. This further highlights the advantages and potential of cascaded models.

Interestingly, as we increase the training steps used for machine translation, the gap between the multi-candidate cascaded system and the machine translation model gradually diminishes. In contrast, the gap between the multi-candidate cascaded system and the single-candidate cascaded system grows more prominent. This discovery sheds light on the potential of harnessing multi-candidate utilization within established commercial machine translation models.

4.3.2. Ablation

Comparing settings (5) and (6) shows that the Alignment method is crucial to make the multi-candidate method effective. Without lexical alignment, the sentence-level Multi-candidate average attention struggles to select the correct candidate words.

4.3.3. Case study

By employing the multi-candidate strategy, we can observe that among the top five candidates based on their scores, there exist samples that deviate from conventional human language expressions. Nevertheless, in the attention mechanism of the machine translation process, words that align more closely with human expression and convey correct semantic meaning receive greater attention. As for the example of "Where the Golgi apparatus, sometimes called the Golgi body, receives them." The first candidate from ASR misrecognizes the pronunciation of the "golgi apparatus" as two non-existent words, "golgy" and "golji". However, the candidate with a BLEU score of 100 is ranked fifth. The machine translation model leverages rich text patterns through multi-candidates, allocates more attention to the correct candidate "Golgi apparatus" and effortlessly selects the proper translation, regarding the example of "Recording the transaction in an immutable distributed ledger" while the subsequent candidates include "legend", "literature", "letter", "ledger" and other alternatives. Our model successfully selects the correct phrase "distributed ledger" in terms of vocabulary collocation.

5. Conclusion

Our analysis pinpoints factors contributing to error propagation in cascaded systems, such as pronunciation disparities and semantic differences. Our multi-candidate approach notably enhances speech-to-text (S2T) translation, bridging the S2T-T2T gap without altering model parameters. Our research deepens our understanding of error propagation and linguistic information loss, thereby improving speech translation. With enhanced ASR and MT resources, our multi-candidate method narrows the S2T-T2T divide, providing increased accuracy and efficiency, all without additional parameters or modules.

[1] https://github.com/facebookresearch/fairseq/blob/main/examples/mbart/README.md

6. References

[1] A. Bérard, O. Pietquin, C. Servan, and L. Besacier, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *arXiv preprint arXiv:1612.01744*, 2016.

[2] S. Popuri, P.-J. Chen, C. Wang, J. Pino, Y. Adi, J. Gu, W.-N. Hsu, and A. Lee, "Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation," *arXiv preprint arXiv:2204.02967*, 2022.

[3] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, "Leveraging weakly supervised data to improve end-to-end speech-to-text translation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7180–7184.

[4] M. C. Stoian, S. Bansal, and S. Goldwater, "Analyzing asr pretraining for low-resource speech-to-text translation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7909–7913.

[5] J. Pino, Q. Xu, X. Ma, M. J. Dousti, and Y. Tang, "Self-training for end-to-end speech translation," *arXiv preprint arXiv:2006.02490*, 2020.

[6] C. Wang, A. Wu, J. Pino, A. Baevski, M. Auli, and A. Conneau, "Large-scale self-and semi-supervised learning for speech translation," *arXiv preprint arXiv:2104.06678*, 2021.

[7] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] V. H. Quan, M. Federico, and M. Cettolo, "Integrated n-best reranking for spoken language translation." in *Interspeech*, 2005, pp. 3181–3184.

[9] R. Ma, H. Li, Q. Liu, L. Chen, and K. Yu, "Neural lattice search for speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7794–7798.

[10] N. Bertoldi and M. Federico, "A new decoder for spoken language translation based on confusion networks," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* IEEE, 2005, pp. 86–91.

[11] D. Beck, T. Cohn, and G. Haffari, "Neural speech translation using lattice transformations and graph networks," in *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, 2019, pp. 26–31.

[12] M. Sperber, G. Neubig, N.-Q. Pham, and A. Waibel, "Self-attentional models for lattice inputs," *arXiv preprint arXiv:1906.01617*, 2019.

[13] S. Peitz, S. Wiesler, M. Nußbaum-Thom, and H. Ney, "Spoken language translation using automatically transcribed text in training," in *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*, 2012.

[14] Q. Cheng, M. Fang, Y. Han, J. Huang, and Y. Duan, "Breaking the data barrier: Towards robust speech translation via adversarial stability training," *arXiv preprint arXiv:1909.11430*, 2019.

[15] M. A. Di Gangi, R. Enyedi, A. Brusadin, and M. Federico, "Robust neural machine translation for clean and noisy speech transcripts," *arXiv preprint arXiv:1910.10238*, 2019.

[16] S. Dalmia, B. Yan, V. Raunak, F. Metze, and S. Watanabe, "Searchable hidden intermediates for end-to-end models of decomposable sequence tasks," *arXiv preprint arXiv:2105.00573*, 2021.

[17] H. Inaguma, S. Dalmia, B. Yan, and S. Watanabe, "Fast-md: Fast multi-decoder end-to-end speech translation with non-autoregressive hidden intermediates," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 922–929.

[18] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[19] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang *et al.*, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[20] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," *arXiv preprint arXiv:2102.01547*, 2021.

[21] R. Ye, C. Zhao, T. Ko, C. Meng, T. Wang, M. Wang, and J. Cao, "Gigast: A 10,000-hour pseudo speech translation corpus," *arXiv preprint arXiv:2204.03939*, 2022.

[22] P. Charalampopoulos, T. Kociumaka, S. P. Pissis, and J. Radoszewski, "Faster Algorithms for Longest Common Substring," in *29th Annual European Symposium on Algorithms (ESA 2021)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), P. Mutzel, R. Pagh, and G. Herman, Eds., vol. 204. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, pp. 30:1–30:17. [Online]. Available: https://drops-dev.dagstuhl.de/entities/document/10.4230/LIPIcs.ESA.2021.30

[23] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang *et al.*, "Direct speech-to-speech translation with discrete units," *arXiv preprint arXiv:2107.05604*, 2021.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. 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.

[25] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[26] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[27] 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," *arXiv preprint arXiv:2104.00355*, 2021.

[28] A. Tjandra, S. Sakti, and S. Nakamura, "Speech-to-speech translation between untranscribed unknown languages," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 593–600.

[29] C. Zhang, X. Tan, Y. Ren, T. Qin, K. Zhang, and T.-Y. Liu, "Uwspeech: Speech to speech translation for unwritten languages," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, 2021, pp. 14 319–14 327.

[30] G. Chen, S. Chai, G. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang *et al.*, "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *arXiv preprint arXiv:2106.06909*, 2021.

[31] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[32] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pretraining for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020.

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.