INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Analysis of Acoustic information in End-to-End Spoken Language Translation

Gerard Sant[1,2] *Carlos Escolano*[1]

[1] TALP Research Center, Universitat Politècnica de Catalunya, Spain
[2] Barcelona Supercomputing Center, Spain

gerard.muniesa@estudiantat.upc.edu, carlos.escolano@upc.edu

Abstract

End-to-End Transformer-based models are the most popular approach for Spoken Language Translation (SLT). While obtaining state-of-the-art results, we are still far from understanding how these models extract acoustic information from the data and how they are transformed into semantic representations.

In this paper, we seek to provide a better understanding of the flow of acoustic information along speech-to-text translation models. By means of the Speaker Classification and Spectrogram Reconstruction tasks, this study (i) interprets the main role of the encoder with respect to the acoustic features, (ii) highlights the importance of the acoustic information throughout the model and its transfer between encoder and decoder, and (iii) reveals the significant effect of downsampling convolutional layers for learning acoustic features. (iv) Finally, we also observe the existence of a strong correlation between the semantic domain and the speakers' labels in MuST-C.

Index Terms: Spoken Language Translation, Interpretability of Acoustic information.

1. Introduction

In recent years, end-to-end Spoken Language Translation (SLT) models have gained popularity in the research community [1]. This trend was highly influenced by the release of Transformer [2], as it revolutionized the MT field, also impacting speech processing [3]. In contrast to text tasks, which deal with sequences of a discrete nature, speech signals are collected at a digitalization frequency typically between 16-48KHz, resulting in extremely long discrete waveforms. Most approaches use the mel-spectogram of the signal to overcome the long sequences problem. This results in smaller sequences with higher frequency content. In addition, they further reduce the sequence length by collapsing adjacent vectors in a fixed way [4, 5, 6] or by using pretrained compression modules [7, 8, 9].

A common characteristict of end-to-end systems is that they learn latent representations directly from raw data without any specific feature engineering. Whereas the temporal dimension of speech sequences has been studied extensively through interpretability studies [10] or efficient transformer alternatives [11, 12], the feature space, as well as the importance and benefit of acoustic information in end-to-end models, have not been practically explored [13].

Therefore, the contributions of this paper in the field of SLT are:

- To provide an interpretation of how acoustic information is processed throughout the training by the different components of the model's architecture.
- Analyze the importance of Convolutional layers on SLT performance. We observe that these layers play an important role in both subsampling and acoustic feature learning.
- Highlight the bias and effect of the domain in some of the most used datasets in the SLT task.

[0]Work done by Gerard Sant at UPC.

2. Related Work

As mentioned, one of the main challenges in end-to-end SLT is the large sequence length. [4] proposed the use of two one-dimensional convolutional layers prior to the Transformer Encoder over the temporal dimension to reduce the length of the mel spectrogram. [12] observed the presence of temporal redundancy and unnecessary computations throughout the encoder layers. The temporal correlation of speech sequences was further studied by [10], who observed the increasing importance of local context throughout the transformer encoder layers.

A few feature engineering attempts have been made in the Automatic Speech Recognition (ASR) task. For instance, [14] proposed to aggregate speaker information along the encoder layers via i-vectors. Similarly, [15] tried to reinforce the speaker information in the Conformer Encoder [16] by computing the cosine similarity between the representation predicted by a speaker classifier and the real speaker. However, only [13] performed feature engineering based on the amount of speaker information present in the conformer encoder. Specifically, they proposed (i) to stimulate the speaker information at the beginning of the encoder layers minimizing the error of a speaker classifier and (ii) reducing the speaker information in the last layers of the encoder using gradient reversal [17] in the previous configuration.

3. Proposed Methodology

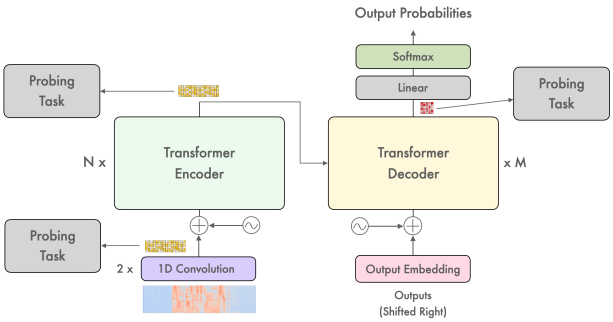

Figure 1: *Proposed scheme to observe the flow of the speaker information along the SLT Transformer. At the output of each layer, a Speaker MLP classifier is trained with the labels provided by the dataset.*

10.21437/Interspeech.2023-2050

3.1. Acoustic Information Tracking

The proposed scheme (Fig. 1) for measuring the information flow along the encoder is built on top of the Speech-to-Text Transformer [6], which reduces the normalized mel-spectogram length of the input waveform. Inspired by studies in Text-to-Speech multispeaker [18], where the resulting speech is composed of the text representation of the message (linguistic information) and the prosody [19] of the speaker in question (acoustic information), the speaker information will be used as the ground truth of our measurement [13]. Therefore, we will train a speaker classifier at the output of each layer of the model, whose accuracy will provide a notion of the amount of acoustic information present in the representation of its input.

In addition, by means of two 1D transposed convolutional layers, we intend to reconstruct the input spectogram to corroborate the results obtained by the previous analysis.

3.2. Robustness of Acoustic Information

To perform SLT, the encoder must learn a transformation of the data from an acoustic-based representation, such as a mel-spectrogram, to a semantic representation required to condition the textual translation's generation. This experiment aims to examine the significance of acoustic information at each layer of the model by introducing a feed-forward (FFN) bottleneck in a low-dimensional space. If a layer of the model does not require certain acoustic features, the probing task will show a decrease in performance that will persist throughout the network. However, if performance improves in subsequent layers, it indicates that the model is relearning some of the acoustic information and that it remains pertinent to performing the task.

4. Experimental Details

Data. For our experiments we are using the English→German (En-De) direction of (i) MuST-C [20], which is based on 408 hours of TED talks and (ii) CoVoST-2[1] [21], a large-scale multilingual speech translation corpus covering translations from 21 languages into English.

The en-de MuST-C test split consists of a total of 2587 utterances spoken by 27 speakers, which corresponds to 98.5 samples per speaker. However, being the CoVoST-2 test partition larger in both size and number of speakers, most speakers have only one utterance. Therefore, a test split has been created with 27 speakers, each with between 98 and 99 samples, from the train partition. The speakers of the new test split have been removed from the train partition. In order to measure the impact of speech domain on learned representation, we also generate a Synthetic MuST-C test set using a Tacotron 2 [22] model trained on Librispeech [23][2], using the same speaker for all utterances, only removing 38 utterances that where not adequately recognized due to background noise.

Baseline. The small S2T-Transformer[3] (S2T-Transformer-S) has 12 encoder layers and 6 decoder layers, with a dimensionality $d = 256$. Four heads are used in the layers' attention modules. The feed-forward layers have a hidden dimensionality of 2048. A 2-layer convolutional network with 1024 internal channels, output dimensionality of 256, kernel sizes of 5 and stride 2 processes the 80-dimensional log-Mel spectrograms. The S2T-Transformer-SP and S2T-Transformer-XS are identical to the baseline but with 16 and 6 encoder layers respectively. The XS architecture also reduces to 3 the number of decoder layers.

Probing Tasks Two tasks are proposed to measure acoustic information, speaker classifier, and Spectrogram reconstruction. For the speaker classifier, a single-layer Perceptron with a ReLU as activation function has been chosen. With the exception of the input tensor classifier with a hidden size of 80, all classifiers have a hidden size of 256.[4]

For the Spectrogram reconstruction task, generators consisting of 2 deconvolution layers are trained. These generators reconstruct the utterance's mel-spectrogram from the output of the attention blocks and convolutional subsampler from the trained SLT models encoder. Parameter-wise, each generator consists of a deconvolutional layer with input size 256 and hidden size 1024 and a second deconvolutional layer with input size 1024 and output size 80. Both layers have stride 2 and kernel size 5, to recover the frequential and temporal dimensions of the real mel-spectrogram.

Bottleneck Architectures. These modify the baseline by adding a bottleneck to the output of a layer of the model (§3.2). For the convolutional layers, the FFN is placed after the last layer, and for the attention blocks, after their FFN and residual connection. Using FNN, the bottleneck first projects the input to a reduced feature space, specifically with $d_{latent} = 128$. It then projects back the latent space to the model dimension.[5]

Training. To train all translation models we use the label smoothed cross entropy loss and the Adam optimizer with a base learning rate of 0.002, a 10000 step warm-up and an inverse square root scheduler. We use a maximum of 20000 tokens and an update frequency of 8. Training is stopped after 50000 updates. The encoders are initialized from the same model configuration (except for the learning rate, with 0.001), pretrained on the ASR part of the data [4]. The data preprocessing has been performed according to the guidelines provided in the framework itself, therefore, the target vocabularies are learned with SentencePiece for MuST-C, with a size of 8000, while for CoVoST-2 char-based vocabulary have been used.

The speaker classifiers for each layer have been trained for 200 iterations using the Adam optimizer with a learning rate of 0.001. We used 75% of the utterances of the corresponding dataset test partition for training, while the rest of the samples were used for its evaluation.

The transposed 1D convolutional layers have been trained on the CoVoST-2 train split for 50 epochs to reconstruct the input spectrogram using the MSE loss function. The Adam optimizer has been utilized for this training, with a learning rate of 0.003.

Evaluation. For the SLT evaluation, we average the 10 best checkpoints on the development set. The evaluation is performed by measuring the BLEU [26]. Accuracy and F1 Score are used to evaluate the speaker classification.[6] The evaluation of the spectrogram reconstruction has relied on Mean Square Error (MSE).

[1]When collecting CoVoST-2, speakers are recorded saying random sentences, thus the correlation between domain and speaker is assumed to be minimal.

[2]available at coqui-ai [24]

[3]We train the s2t_transformer_s architecture from FAIRSEQ [25]

[4]Note that the classification has been performed only with the test partition of the datasets.

[5]Note that this layer is trained jointly with the model.

[6]Since the classification has 27 speakers, a random classifier would give an accuracy of 3.7% , which means that the classifier would have been unable to find speaker information in the input representation.

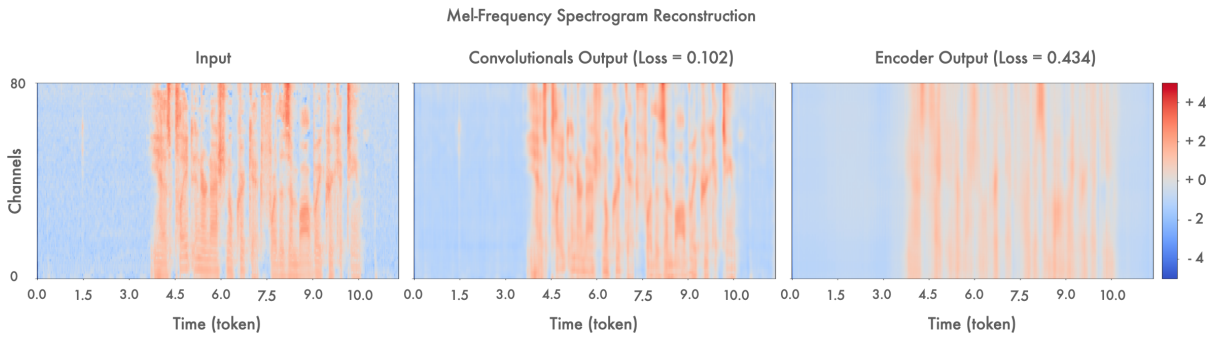

Figure 2: *Left, the mel-spectrogram computed from the dataset. Center, the reconstructed mel-spectrogram generated after the convolutional layers. Right, the reconstructed mel-spectrogram generated from the encoder's output.*

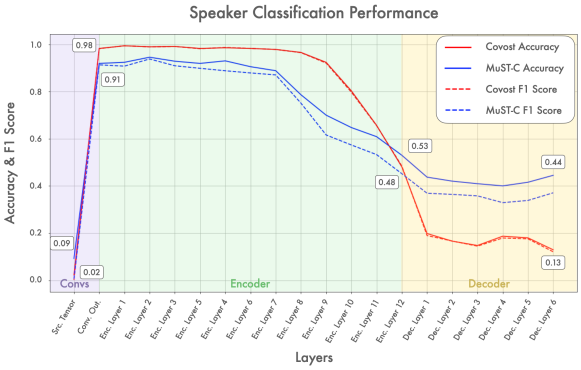

Figure 3: *Performance of the speaker classification based on the output representations of the baseline layers.*

5. Results

5.1. Acoustic Information Flow Analysis

First, we can observe from Fig. 3 how the 1D-convolutional layers not only reduce the input sequence length but also project it to a highly acoustic-rich representation space in which the classifiers achieve more than 90% accuracy on both datasets. The Transformer Encoder maintains a constant high acoustic content representation space throughout the $\approx 60\%$ of its layers, which is then reduced to $\approx 50\%$ of speaker information at the end of the Encoder. This suggests that the first layers of the Transformer Encoder learn to de-cluster the acoustic content so that the last ones are able to eliminate the acoustic information not contributing to the final objective (i.e., speaker information). By measuring the amount of acoustic information through speaker classification, a reduction in accuracy may indicate that the Encoder is creating a semantic-based space of acoustic representations. This is consistent with the improvement in ASR performance obtained by [13] after boosting the speaker information in the early Conformer layers and decreasing it in the later ones, thus creating a more speaker-agnostic representation space. In addition, this initial prioritization of acoustic information could explain the observations of [10], where the attention modules of the early Encoder layers assign similar importance among all the tokens.

Second, analyzing the decoder part of the baseline, a significant discrepancy is observed in the patterns obtained by both datasets. On the one hand, the acoustic information in the decoder of the CoVoST-trained architecture suffers from a substantial drop (from 48% at the output of the Encoder to 20% at the output of the first decoder layer). The last decoder representation maintains a slight amount of acoustic information, reinforcing the information transfer motivation of the end-to-end approach.

However, the model trained on MuST-C suffers only from a slight decrease of acoustic information in the decoder stage, reaching a 44% at its output, which highlights the high correlation of the speaker's labels with the domain of their semantics later explained (§5.3).

5.2. Impact of Acoustic Information

Model depth. One of the main research questions of this work is understanding how the acoustic information is represented through the transformer attention layers. Studying the flow of acoustic information in the different variations of the S2T Transformer, we can observe how the pattern described in (§5.1) is maintained after increasing the depth of the model, performing a more progressive drop in the last encoder layers until reaching values similar to the baseline (from an accuracy of 0.89 at the output of the convolutions to 0.58 by the end of the Encoder). On the other hand, the amount of acoustic information contained in the Encoder is significantly lower at smaller depths (With accuracies between 0.53 and 0.23 for XS models).

Since at lower depths the encoder arbitrarily removes the acoustic content from the beginning, while at greater depths it remains very high, the concept of the encoder as the one in charge of initially de-clustering the acoustic content and then removing the non-relevant information in the last layers is reinforced. [7]

Bottlenecks. Fig. 4 shows the evolution of the acoustic information as a function of the bottleneck position at the encoder stage. As discussed in (§3.2) and inspired by [27], bottlenecks are used to decrease the representation size, forcing the model to discard less useful information for the final task.

Using a bottleneck reduces the acoustic content at the corresponding layer in both datasets. Afterward, most models attempt to recover an acoustically rich space. Therefore, there is a beneficial transfer of this information at the end-to-end models. Moreover, except for the model with the bottleneck in layer 7, a certain pattern is perceived between the amount of acoustic information in the last encoder layer and the performance of the models trained on CoVoST-2, where for acoustic contents lower than 15% they show a performance drop between 1 to 2.16 BLEU. On the other hand, due to the high correlation of the semantic domain and the speaker label discussed in (§5.3), this pattern is not seen for the models trained with MuST-C.

[7] Detailed results and tables provided as supplemental multimedia material.

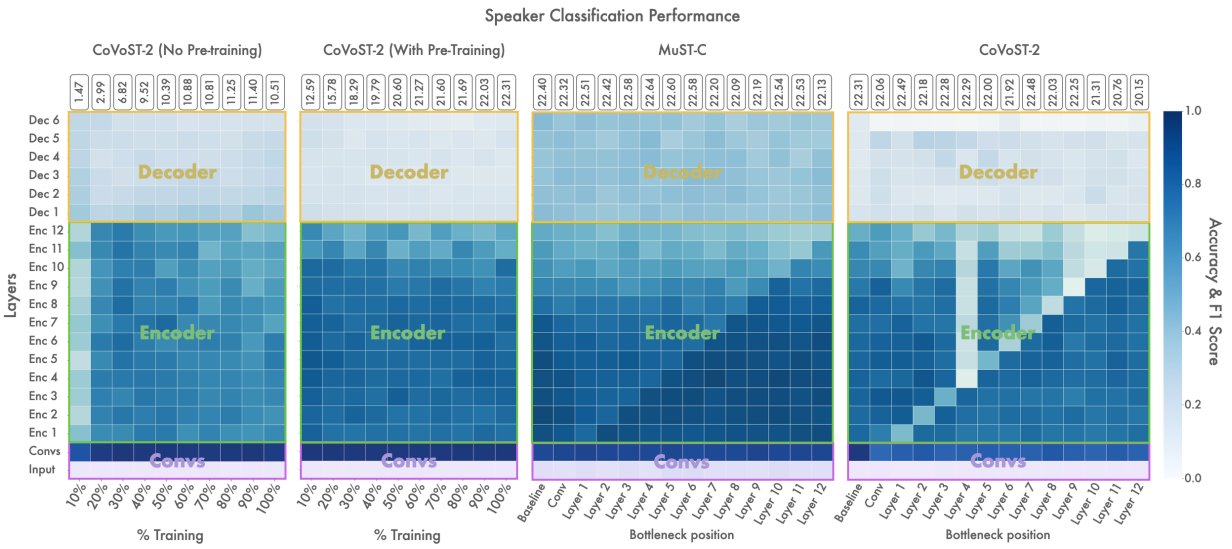

Figure 4: *Accuracy of the speaker classifiers of each layer and BLEU performance of the models in Speech Translation's. Left: Speaker classification accuracy at different points of the training, with and without ASR pre-training. Right: MuST-C and CoVoST-2 performance with different bottleneck positions.*

Convolutional Layers. Speaker classification results suggested that Convolutional layers not only reduce the temporal dimension of the input mel-spectrogram but also act as a feedforward network that enriches the input representation, of 256 dimensions, by over-parametrizing it into a higher-dimensionality of 1024 dimensions. This intuition is enforced further by the spectrogram reconstruction results. Fig. 2 shows how the spectrogram reconstruction follows a curve similar to Speaker representation, where the convolutional layers provide the best results, with an average MSE distance of 0.102, that gradually increases until the last encoder layer, which achieves an MSE distance of 0.434. Analyzing the outputs, we observe that while all layers are able to identify the silences in the spectrogram, by the end of the encoder, a significant amount of detail on the parts that include speech is missing from the reconstructed spectrogram. To test this hypothesis, we trained an additional SLT model without over-parametrization and observed a decrease in translation performance of 1.80 BLEU, from 22.31 to 20.51 BLEU. These results indicate that convolutional layers and this over-parametrization play a significant role in how the model learns acoustic features from the data.

ASR Pre-training To analyze the impact of ASR pre-training, we studied the speaker-classification performance during different steps of the training with and without pre-training. Fig. 4 left shows that without the pre-training, the model cannot accurately predict the speaker until 20% of the training, where we observe high accuracies throughout the entire encoder. As the training progresses, the accuracy of the last encoder layers decreases, indicating that the model shifts to a more semantic representation.

5.3. Semantics & Speaker Label Correlation

As seen in Figures 5 and 4, MuST-C shows a strong correlation between the semantic domain of speech and the speaker label, unlike architectures trained with CoVoST-2, results in considerably high accuracies in the decoding part of the models. This idea is reinforced by the results obtained with (i) the synthesized MuST-C test split, where speaker classification improves across model layers (reaching an accuracy of 23% and 29% at the encoder and decoder outputs, respectively) and (ii) by TF-IDF coding, where classification reaches an accuracy of 32% and an F_{Score} of 30%. For both experiments, classifiers do not have acoustic information to distinguish between different speakers, so their predictions are based exclusively on semantic content. Rather than providing results typical of a random classification, the performance obtained by the semantic classifiers coincides with the difference in acoustic information present at the end of both baselines.

6. Conclusions

This paper analyzed the acoustic information flow along the Spoken Language Translation (SLT) models. The main conclusion of the work could be summarized in four main blocks:

Firstly, the paper provided a view of the encoder as the one responsible for declustering the acoustic information and shifting to a more semantic representation. Secondly, to the best of our knowledge, we have been the first observing a beneficial and necessary transfer of acoustic information between the encoder and decoder in end-to-end models, showing that the model retained it even after trying to reduce it via bottlenecks. Thirdly, the paper observed the importance of convolutional layers on SLT architectures by performing downsampling and over-parametrizing the representation. Finally, the paper highlighted the need to find data collection techniques whose labels do not depend heavily on the semantic domain and thus ensured that the encoder do not rely on non-acoustic information that might impair its generalization capability.

Overall, the paper offered valuable insights into the acoustic information flow in SLT models, useful for future research.

7. Acknowledgements

This work was funded by Spanish State Research Agency (AEI) project PID2019-107579RB-I00 (AEI/10.13039/501100011033) and the "European Union NextGenerationEU/PRTR" under the project ROB-IN (PLEC2021-007859)

8. References

[1] A. Anastasopoulos, D. Chiang, and L. Duong, "An unsupervised probability model for speech-to-translation alignment of low-resource languages," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1255–1263. [Online]. Available: https://aclanthology.org/D16-1133

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[3] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting transformer to end-to-end spoken language translation," in *INTERSPEECH 2019*. International Speech Communication Association, 2019. [Online]. Available: http://hdl.handle.net/11582/319654

[4] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-end automatic speech translation of audiobooks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6224–6228.

[5] M. A. Di Gangi, M. Negri, R. Cattoni, R. Dessi, and M. Turchi, "Enhancing transformer for end-to-end speech-to-text translation," in *Proceedings of Machine Translation Summit XVII: Research Track*. Dublin, Ireland: European Association for Machine Translation, Aug. 2019, pp. 21–31. [Online]. Available: https://aclanthology.org/W19-6603

[6] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, and J. Pino, "Fairseq S2T: Fast speech-to-text modeling with fairseq," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 33–39. [Online]. Available: https://aclanthology.org/2020.aacl-demo.6

[7] E. Salesky, M. Sperber, and A. W. Black, "Exploring phoneme-level speech representations for end-to-end speech translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1835–1841. [Online]. Available: https://aclanthology.org/P19-1179

[8] B. Zhang, I. Titov, B. Haddow, and R. Sennrich, "Adaptive feature selection for end-to-end speech translation," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2533–2544. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.230

[9] M. Gaido, M. Cettolo, M. Negri, and M. Turchi, "CTC-based compression for direct speech translation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, Apr. 2021, pp. 690–696. [Online]. Available: https://aclanthology.org/2021.eacl-main.57

[10] B. Alastruey, J. Ferrando, G. I. Gállego, and M. R. Costa-jussà, "On the locality of attention in direct speech translation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 402–412. [Online]. Available: https://aclanthology.org/2022.acl-srw.32

[11] S. Papi, M. Gaido, M. Negri, and M. Turchi, "Speechformer: Reducing information loss in direct speech translation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1698–1706. [Online]. Available: https://aclanthology.org/2021.emnlp-main.127

[12] G. Sant, G. I. Gállego, B. Alastruey, and M. R. Costa-jussà, "Multiformer: A head-configurable transformer-based model for direct speech translation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 277–284. [Online]. Available: https://aclanthology.org/2022.naacl-srw.34

[13] W. Zhou, H. Wu, J. Xu, M. Zeineldeen, C. Lüscher, R. Schlüter, and H. Ney, "Enhancing and adversarial: Improve asr with speaker labels," *arXiv preprint arXiv:2211.06369*, 2022.

[14] Y. Zhao, C. Ni, C.-C. Leung, S. R. Joty, E. S. Chng, and B. Ma, "Speech transformer with speaker aware persistent memory." in *INTERSPEECH*, 2020, pp. 1261–1265.

[15] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-end speaker-attributed ASR with transformer," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 4413–4417. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-101

[16] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech," in *Proc. Interspeech 2020*, 2020, pp. 4387–4391. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1251

[19] T. B. Mokgonyane, T. J. Sefara, M. J. Manamela, T. I. Modipa, and M. S. Masekwameng, "The effects of acoustic features of speech for automatic speaker recognition," in *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, 2020, pp. 1–5.

[20] R. Cattoni, M. A. Di Gangi, L. Bentivogli, M. Negri, and M. Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Computer Speech & Language*, vol. 66, p. 101155, 2021.

[21] C. Wang, A. Wu, and J. Pino, "Covost 2: A massively multilingual speech-to-text translation corpus," 2020.

[22] 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.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[24] G. Eren and The Coqui TTS Team, "Coqui TTS," 1 2021. [Online]. Available: https://github.com/coqui-ai/TTS

[25] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[27] M. Geva, A. Caciularu, K. Wang, and Y. Goldberg, "Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 30–45. [Online]. Available: https://aclanthology.org/2022.emnlp-main.3