INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Cross-Modal Semantic Alignment before Fusion for Two-Pass End-to-End Spoken Language Understanding

Lingyan Huang[1], *Tao Li*[1], *Haodong Zhou*[2], *Qingyang Hong*[1*], *Lin Li*[2*]

[1]School of Informatics, Xiamen University, China
[2]School of Electronic Science and Engineering, Xiamen University, China
{qyhong,lilin}@xmu.edu.cn

Abstract

The deliberation-based two-pass model that combines both semantic and acoustic information can effectively improve the performance of end-to-end (E2E) spoken language understanding (SLU). However, existing two-pass models usually simply fuse speech embedding and text embedding without taking into account the inherent distinctions between these two modalities. We propose a novel approach named **C**ross-modal **S**emantic **A**lignment before **F**usion (CSAF), which adopt contrastive loss aligning speech and text embeddings before fusing them. We introduce a shared semantic memory transformer to project the embeddings from two modalities into a common semantic space, and a multi-modal gated network to generate the fused embeddings. We conduct experiments on the FSC Challenge test set and SLURP dataset. The results demonstrate that our method can significantly promote intent classification accuracy, achieving an absolute improvement of 3.1% over previous works in the FSC Challenge Utterance Set.

Index Terms: spoken language understanding, deliberation method, contrastive learning, multi-modal information fusion

1. Introduction

Spoken Language Understanding(SLU), which predicts semantic information from the audio signal, is a fundamental element of any spoken dialog system. A traditional SLU system has conventionally been a cascaded architecture comprising automatic speech recognition (ASR) and natural language understanding (NLU). ASR transcribes speech into text while NLU takes the transcription text as input and returns the corresponding intent. However, the two blocks of a cascaded system are generally built and optimized individually, posing several undesirable issues. Firstly ASR transcription errors compromise the performance of the downstream NLU systems. Secondly, the transcribed text unavoidably loses the relevant acoustic information, such as pronunciation and prosody. Due to the above restrictions with cascaded systems, end-to-end (E2E) SLU[1, 2, 3, 4] has attracted considerable attention recently. However, since parsing semantics directly from the speech is a challenging task, it remains a struggle for E2E SLU to outperform their cascaded counterparts.

To address this issue, a series of research on E2E SLU has been proposed. Some approaches[5, 6, 7, 8] attempt to promote the semantic understanding of speech representations by aligning them with text representations extracted from pre-trained language models. Another solution integrates the ASR and NLU networks with an appropriate interface to combine speech and transcription information[9, 10, 11]. Nevertheless, these methods may fail to leverage transcription information that facilitates explaining the system behavior or have the potential to be less robust to ASR errors. [12] proposed a two-pass E2E SLU model with bi-modal input property, which can rewrite the ASR hypothesis making it more robust to ASR errors, but they neglected to consider that the heterogeneities across modalities may adversely affect the modal fusion. [13] also adopted a deliberation-based model and introduced cross-modal attention to bridge the modality gap. However, existing methods have not explored the utility of aligning speech and text representations into a shared semantic space before fusing them.

In this work, we propose a novel approach named **C**ross-modal **S**emantic **A**lignment before **F**usion (CSAF), which makes use of a contrastive loss to align speech embeddings and text embeddings and utilizes a gated multi-modal network[14, 15] to automatically learn the weights of each modality for determining the final fused embeddings. Our model follows the two-pass architecture[12], which generates intent and transcription from audio in the first pass and rewrites the first-pass results in the second pass. It is a challenge to align speech and text embeddings for the two-pass model since the speech encoder used to extract speech embeddings needs to keep frozen in the second pass. Inspired by the recent research[16] on speech translation, we introduce a shared semantic memory transformer with a few shared layers to project embeddings from two modalities into a common semantic space. We evaluate our method on the FSC challenge test set[4, 17] and SLURP dataset[18]. Experimental results demonstrate that our method outperforms the previous strong baselines on accuracy.

Our major contributions can be summarized as follows:

- We propose a novel approach named CSAF, which aligns the representations of speech and text modalities into a shared semantic space before fusing them.
- We introduce a shared semantic memory transformer that enables us to align representations of two modalities into a common semantic space requiring only a few shared layers.
- Experimental results demonstrate that our approach can improve intent recognition accuracy. In particular, it is more robust to ASR errors and has a strong generalization capability for unseen utterances.

2. Proposed Method

2.1. Model architecture

We employ the two-pass model with a fusion module, the architecture of which is illustrated in Figure 1. The first pass model consists of a speech encoder and a first pass decoder. The speech encoder takes a speech feature sequence $X \in \mathbb{R}^{U \times D}$ as input and outputs a speech embedding denoted by $S \in \mathbb{R}^{U \times D}$,

*Corresponding author

10.21437/Interspeech.2023-758

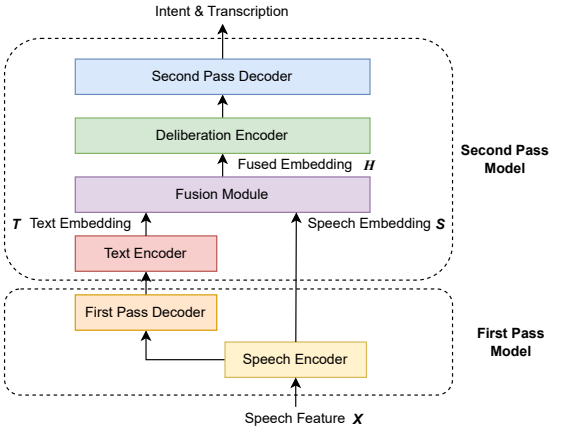


Figure 1: *Our proposed model architecture.*

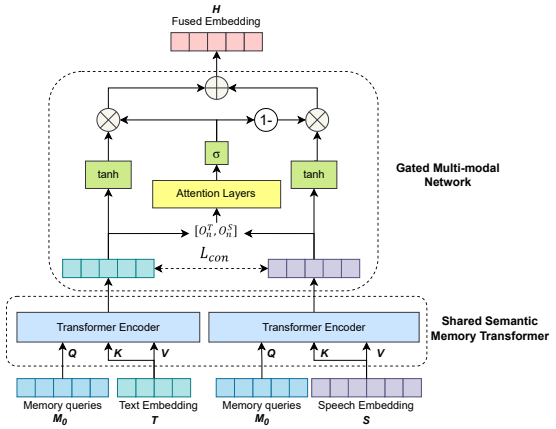


Figure 2: *A detailed diagram of our fusion module.*

where U and D are used to represent sequence length and feature dimension, respectively. Then, this speech embedding S is fed to the first pass decoder to generate $Y^1 = (I^1, y_1^1, \cdots, y_L^1)$ of length $L+1$, which is a combination of intent I and transcription text $(y_1, \cdots, y_L)$. We reserve the ability of the first pass model on predicting intent so that the SLU system is still able to produce low-latency results[12].

The second pass model takes the transcript generated by the first pass model and the speech embedding as input and returns more refined intent and transcription. The ASR transcript is passed to the text encoder to obtain the text embedding $T \in \mathbb{R}^{V \times D}$, where V is the number of tokens in transcription and D is the embedding dimension. The fusion module is intended to convert speech embedding S and text embedding T into a fused embedding $H \in \mathbb{R}^{M \times D}$, where M is the length of the fused embedding. Finally, the fused embedding H is encoded by a deliberation encoder and then decoded by a second pass decoder to produce the ultimate output result $Y^2 = (I^2, y_1^2, \cdots, y_L^2)$. Note that the speech encoder updates parameters only in the first pass and keeps frozen at the second pass to ensure that model can generate transcripts with the same accuracy and latency.

2.2. Fusion module

We design a fusion module that consists of a shared semantic memory transformer and a gated multi-modal network aiming to merge the information of different modalities appropriately. The architecture of the fusion module is shown in Figure 2.

2.2.1. Shared semantic memory transformer

Inspired by [16], we employ a shared semantic memory transformer to align two embeddings from different modalities into a shared cross-modal semantic space. Specifically, we first initialize M learnable memory queries to map speech embeddings and text embeddings in different lengths into constant length M. The shared semantic memory transformer with a similar structure to the transformer[19] takes memory queries as attention queries, while uni-modal embeddings as attention keys and values. This shared transformer is jointly trained on both speech embeddings and text embeddings. In the following, we use the example of speech embedding to describe the computational procedure for obtaining semantic memory through a n-layer shared semantic memory transformer.

$$Q_0 = M_0 \in \mathbb{R}^{M \times D} \tag{1}$$

$$K_i = V_i = S \in \mathbb{R}^{U \times D} \tag{2}$$

$$Z_i = Attn(Q_i, K_i, V_i) + Q_i \in \mathbb{R}^{M \times D} \tag{3}$$

$$Q_{i+1} = O_i^S = FFN(LN(Z_i)) + LN(Z_i) \in \mathbb{R}^{M \times D} \tag{4}$$

where Attn is a multi-head attention, FFN is a position-wise feed-forward network, and LN means layer normalization. M_0 denotes the memory queries, and S denotes the speech embedding. $O_n^S \in \mathbb{R}^{M \times D}$ is the semantic memory of speech embedding. Correspondingly, we can also obtain the semantic memory of text embedding $O_n^T \in \mathbb{R}^{M \times D}$.

We apply a contrastive loss on the semantic memories of two modalities to align them at the sequence level. Different from the previous works[5, 6, 7], which makes pooled speech representation align with the text [CLS] token which commonly is the first token used for classification in pre-trained language model BERT[20], we concatenate semantic memories across feature dimension to acquire sequence-level representations for preserving more information of the utterances. For the batch of size $|\mathcal{B}|$, we represent the concatenated semantic memories O_n^S and O_n^T as $B^S \in \mathbb{R}^{|\mathcal{B}| \times D'}$ and $B^T \in \mathbb{R}^{|\mathcal{B}| \times D'}$ respectively. The contrastive loss is computed as:

$$\mathcal{L}_{con} = -\frac{1}{|\mathcal{B}|} \sum_{i}^{|\mathcal{B}|} \log \frac{e^{\cos(B_i^S, B_i^T)/\tau}}{\sum_{j \neq i}^{|\mathcal{B}|} e^{\cos(B_i^S, B_j^T)/\tau}} \tag{5}$$

where $\cos(\cdot, \cdot)$ means cosine similarities, and τ is a temperature hyperparameter.

During the training process with contrastive loss $\mathcal{L}_{con}$, the speech encoder and text encoder remain frozen, and only the shared semantic memory transformer is employed to learn alignment.

2.2.2. Gated multi-modal network

Intuitively, the distinct information carried by speech and text possibly has diverse influences on understanding the semantics of an utterance. Motivated by this, we introduce a gated multi-modal network[14, 15]to control contributions of two aligned embeddings O_n^S and O_n^T. The attention layers are two fully

connected layers following [15]. We compute attention weights k between O_n^T and O_n^S as:

$$k = \sigma(F_{att}([O_n^T, O_n^S])) \tag{6}$$

where σ is a sigmoid operator, and F_{att} are the attention layers. $[\cdot, \cdot]$ means concatenation operator. And then, fused embedding H is obtained by summing the weighted representations of the two modalities as:

$$H = k \odot tanh(O_n^T) + (1 - k) \odot tanh(O_n^S) \tag{7}$$

where $\odot$ denotes the element-wise product.

2.3. Training process

Following previous works[12, 21] on two-pass systems, the training process of our model can be described as two stages. In stage one, similar to the hybrid CTC/attention model[22], we use CTC loss and attention loss to optimize the speech encoder and first pass decoder as:

$$\mathcal{L}_{CTC} = -\log p_{CTC}(Y|X) \tag{8}$$

$$\mathcal{L}_{att} = -\sum_{i}^{L+1} \log p(y_i|y_1, \cdots, y_{i-1}, X) \tag{9}$$

$$\mathcal{L}_{1-pass} = \mathcal{L}_{SLU} = \alpha\mathcal{L}_{CTC} + (1 - \alpha)\mathcal{L}_{att} \tag{10}$$

where α is a hyperparameter. In stage two, we train the fusion module, deliberation encoder, and second pass decoder with the SLU task and alignment task as shown below:

$$\mathcal{L}_{2-pass} = \lambda\mathcal{L}_{SLU} + (1 - \lambda)\mathcal{L}_{con} \tag{11}$$

where λ is a hyperparameter.

3. Experimental Setups

3.1. Datasets

Our experiments are performed on two publicly available datasets: FSC[4] and SLURP[18]. FSC is a popular benchmark dataset for SLU evaluation, which consists of 30,043 spoken commands for a smart home or virtual assistant. There are a total of 31 distinct intents. Since the first pass model with speech encoder and decoder can achieve 99.7% intent classification accuracy on the original test split[23], we experiment with the Challenge[17] test split containing Challenge Speaker Set with unseen speakers and Challenge Utterance Set with unseen utterances. SLURP is a recently challenging dataset that has higher complexity in its lexical richness, syntactic structure, and semantic content. It contains 72,277 utterances and 18 unique scenarios with 46 defined intents. For the SLURP dataset, we additionally use the extra SLURP-Synth with 69,253 synthesized speech.

3.2. Model training and hyperparameters

Our models are built on top of the Espnet-SLU toolkit[23]. Excluding the fusion module, all configuration of our model is similar to the previous two-pass E2E SLU system[12]. The model details and hyperparameters are different between the two datasets. For the FSC dataset, we use HuBERT[24] that pre-trained through self-supervision as a feature extractor to improve the acoustic modeling. We apply 12 transformer encoder blocks and 6 transformer decoder blocks with 256-dimensional hidden sizes and 4 attention heads for two encoder-decoder structures of our model. BERT[20] is used as the text encoder. The shared semantic memory transformer closely resembled a 3-layer transformer with 256-dimensional hidden sizes and 4 attention heads, and the number of its memory queries is set to 64. The attention layers have 32 and 512 units respectively, and a fully connected layer converts the representation dimensionality from 256 to 512 before fusing. The hyperparameters α, λ and τ are set to 0.5, 0.5 and 1. For the SLURP dataset, there are some different model setups. Firstly, we apply FBank features as speech features. Secondly, the speech/deliberation encoder consists of 6-layer conformer[25] with 512-dimensional hidden sizes and 8 attention heads. Lastly, the hyperparameters α, λ and τ are 0.3, 0.1 and 0.1, respectively.

ASR transcripts of the FSC dataset are generated by a pre-trained ASR model trained on Gigaspeech[26] dataset, that effectively improves semantic understanding results of the second stage[12]. The entire model is trained with an Adam[27] optimizer with learning rate of 2e-4 and 25k warm-up updates. In addition, we utilize the SpecAugment[28] and label smoothing[29] techniques during training.

Table 1: *Performance comparison in intent classification accuracy (%) of our proposed two-pass model with CSAF and other baselines.*

Model	FSC Challenge		SLURP
	Utt	Spk	
Only audio			
Espnet-SLU[23]	78.5	97.5	86.3
Audio and Transcripts			
Two-pass[12]	<u>82.3</u>	**98.1**	86.6
Two-pass w/ CA	79.9	97.4	<u>87.1</u>
Two-pass w/ CSAF(Ours)	**85.4**	<u>97.8</u>	**87.2**

4. Results and Analysis

4.1. Main results

The results of our experiments on the FSC Challenge dataset and SLURP dataset are reported in Table 1. For evaluation, we employ intent classification accuracy as metric. It is observed that our model achieves remarkable improvements on three datasets compared with an only acoustic-based strong baseline Espnet-SLU[23] which is the same as the first-pass of the proposed system. To further demonstrate the advantages of the proposed CSAF method, we implement the two-pass model[12], which concatenates speech embedding and text embedding in the time dimension and feeds the concatenated embedding to deliberation encoder. Our method outperforms the two-pass model on FSC Challenge Utterance Set and SLURP dataset, obtaining an absolute improvement of 3.1% and 0.6%. On FSC Challenge Speaker Set, the proposed method also achieves competitive performance. We observe that our approach greatly reduces misidentification of the sentence "open language settings" with the intent 'change_language_none_none' as 'change_language_chinese_none', which is a common error in the two-pass model because "settings" is an unseen word. To measure the performance of our CSAF method against another fusion method adopted by [13], we construct a two-pass model with cross-modal attention (CA) fusion method and dis-

Table 2: *Ablation study for the CSAF method. Different variants of the CSAF method are evaluated, including removing contrastive loss or gated network and replacing contrastive loss.*

Model	FSC Challenge		SLURP
	Utt	Spk	
CSAF	**85.4**	97.8	**87.2**
w/o Gate	80.9	98.0	86.7
w/o $\mathcal{L}_{con}$	83.6	**98.1**	86.4
w/ $\mathcal{L}_{con}$-Token	82.9	98.0	86.8
w/ $\mathcal{L}_{con}$-Mem	80.5	97.8	86.7

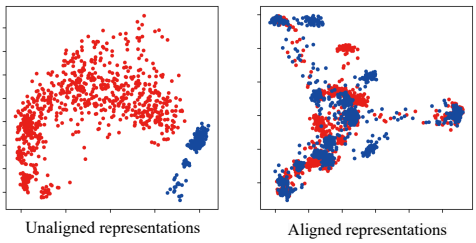
cover that this fusion mechanism only works on the SLURP dataset. With the observation that the ASR model trained on the Gigaspeech dataset gets 29.8 and 35.4 word error rate (WER) on FSC Challenge Utterance and Speaker Set, we argue that the degradation of accuracy on the FSC Challenge dataset is attributable to the high WER of ASR transcripts. It also proves that our proposed CSAF method is more robust against ASR errors.

4.2. Ablation study

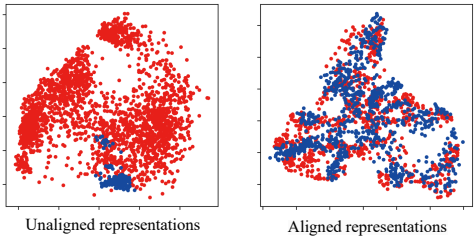
We conduct an ablation study to analyze each component, as shown in Table 2. We notice that any of our variations only leads to fewer fluctuations in the FSC Speaker Set, probably since the performance on this set is already near 100% making it difficult to increase further. We validate the effectiveness of the multi-modal gated network (Gate) and contrastive loss by removing them. Here we can see an accuracy decrease on FSC Utterance Set and SLURP dataset, proving that the gated network and contrastive loss are both beneficial to improve the performance. We also experiment with two alternative contrastive losses $\mathcal{L}_{con}$-Token and $\mathcal{L}_{con}$-Mem. The former means the contrastive loss computed like [8], which aims to align speech embeddings with text embeddings on a token-by-token basis. The latter is adopted by [16], which maximizes the similarity between the same semantic memory element. The results indicate that these two contrastive losses did not achieve the anticipated effect on the FSC Utterance Set, which further demonstrates the effectiveness of our sequence-level contrastive loss.

4.3. Visualization analysis

This work aims to validate that fusing align representations of different modalities contributes to promoting the performance of SLU. To verify that speech and text embeddings are well-aligned into a shared semantic space by a 3-layer shared semantic memory transformer, we visualize the feature distribution of two modalities. We apply principal component analysis (PCA) to reduce the dimension of feature to 2D. The unaligned representations extracted by the uni-modal encoder and the aligned representations are the output embeddings of the shared semantic memory transformer. As shown in Figure 3, after the shared semantic memory transformer, the speech representations and text representations are mapped to the same distribution, which demonstrates that they are well-aligned into the common space. To better investigate whether aligned speech and text representations have the same semantics, we randomly chose several representations with different intents for visualization. Figure 4 illustrates that speech and text representations with the same intent are clustered together.

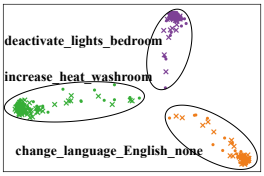


(a) FSC Utterance Set

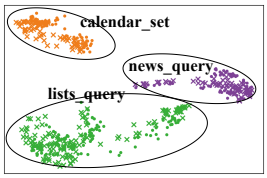


(b) SLURP

Figure 3: *A visualization of speech representations (red) and text representations (blue) by 2-dimensional PCA projection.*



(a) FSC Utterance Set



(b) SLURP

Figure 4: *A visualization of speech representations (".") and text representations ("+") with different intents by 2-dimensional PCA projection.*

5. Conclusion

In this paper, we propose a novel approach called **C**ross-modal **S**emantic **A**lignment before **F**usion (CSAF) for two-pass E2E SLU. Our method aligns the representations of the two modalities into a shared semantic space and then fuses the aligned representations. We employ a shared semantic memory transformer to obtain fixed-length semantic memories of the two modalities and introduce a contrastive loss to align them. Besides, we use a multi-modal gated network, which enables the model to calibrate unimodal embeddings based on the contributions and create a common representation with them. We conduct experiments on two public datasets: FSC Challenge and SLURP. The experimental results demonstrate that our model outperforms the previous strong baselines. In the future, we plan to explore how to leverage the pre-training task to improve the performance of our model in low-resource scenarios.

6. Acknowledgements

Thanks to the National Natural Science Foundation of China (Grant No. 62276220, No.62001405 and No.61876160) for funding.

7. References

[1] P. Haghani, A. Narayanan, M. Bacchiani, G. Chuang, N. Gaur, P. Moreno, R. Prabhavalkar, Z. Qu, and A. Waters, "From audio to semantics: Approaches to end-to-end spoken language understanding," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 720–726.

[2] E. Morais, H.-K. J. Kuo, S. Thomas, Z. Tüske, and B. Kingsbury, "End-to-end spoken language understanding using transformer networks and self-supervised pre-trained features," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7483–7487.

[3] M. Radfar, A. Mouchtaris, and S. Kunzmann, "End-to-End Neural Transformer Based Spoken Language Understanding," in *Proc. Interspeech 2020*, 2020, pp. 866–870.

[4] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech Model Pre-Training for End-to-End Spoken Language Understanding," in *Proc. Interspeech 2019*, 2019, pp. 814–818.

[5] B. Agrawal, M. Müller, S. Choudhary, M. Radfar, A. Mouchtaris, R. McGowan, N. Susanj, and S. Kunzmann, "Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7157–7161.

[6] S. Cha, W. Hou, H. Jung, M. Phung, M. Picheny, H.-K. J. Kuo, S. Thomas, and E. Morais, "Speak or Chat with Me: End-to-End Spoken Language Understanding System with Flexible Inputs," in *Proc. Interspeech 2021*, 2021, pp. 4723–4727.

[7] Y. Zhu, Z. Wang, H. Liu, P. Wang, M. Feng, M. Chen, and X. He, "Cross-modal Transfer Learning via Multi-grained Alignment for End-to-End Spoken Language Understanding," in *Proc. Interspeech 2022*, 2022, pp. 1131–1135.

[8] V. Sunder, E. Fosler-Lussier, S. Thomas, H.-K. Kuo, and B. Kingsbury, "Tokenwise Contrastive Pretraining for Finer Speech-to-BERT Alignment in End-to-End Speech-to-Intent Systems," in *Proc. Interspeech 2022*, 2022, pp. 2683–2687.

[9] S. Seo, D. Kwak, and B. Lee, "Integration of pre-trained networks with continuous token interface for end-to-end spoken language understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7152–7156.

[10] Y. Chen, W. Lu, A. Mottini, L. E. Li, J. Droppo, Z. Du, and B. Zeng, "Top-down attention in end-to-end spoken language understanding," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6199–6203.

[11] M. Rao, A. Raju, P. Dheram, B. Bui, and A. Rastrow, "Speech to Semantics: Improve ASR and NLU Jointly via All-Neural Interfaces," in *Proc. Interspeech 2020*, 2020, pp. 876–880.

[12] S. Arora, S. Dalmia, X. Chang, B. Yan, A. W. Black, and S. Watanabe, "Two-Pass Low Latency End-to-End Spoken Language Understanding," in *Proc. Interspeech 2022*, 2022, pp. 3478–3482.

[13] D. Le, A. Shrivastava, P. D. Tomasello, S. Kim, A. Livshits, O. Kalinli, and M. Seltzer, "Deliberation Model for On-Device Spoken Language Understanding," in *Proc. Interspeech 2022*, 2022, pp. 3468–3472.

[14] J. Arevalo, T. Solorio, M. M. y Gómez, and F. A. González, "Gated multimodal units for information fusion," *ArXiv*, vol. abs/1702.01992, 2017.

[15] M. S. Saeed, M. H. Khan, S. Nawaz, M. H. Yousaf, and A. Del Bue, "Fusion and orthogonal projection for improved face-voice association," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7057–7061.

[16] C. Han, M. Wang, H. Ji, and L. Li, "Learning shared semantic space for speech-to-text translation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2214–2225.

[17] S. Arora, A. Ostapenko, V. Viswanathan, S. Dalmia, F. Metze, S. Watanabe, and A. W. Black, "Rethinking End-to-End Evaluation of Decomposable Tasks: A Case Study on Spoken Language Understanding," in *Proc. Interspeech 2021*, 2021, pp. 1264–1268.

[18] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6000–6010.

[20] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[21] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.-Y. Liu, "Deliberation networks: Sequence generation beyond one-pass decoding," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1782–1792.

[22] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[23] S. Arora, S. Dalmia, P. Denisov, X. Chang, Y. Ueda, Y. Peng, Y. Zhang, S. Kumar, K. Ganesan, B. Yan *et al.*, "Espnet-slu: Advancing spoken language understanding through espnet," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7167–7171.

[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] 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.

[26] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. Interspeech 2021*, 2021, pp. 3670–3674.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[29] R. Müller, S. Kornblith, and G. Hinton, "When does label smoothing help?" in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 4694–4703.