Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation

Won Ik Cho[1], *Donghyun Kwak*[2], *Ji Won Yoon*[1], *Nam Soo Kim*[1]

Department of Electrical and Computer Engineering and INMC, Seoul National University[1]
Search Solution Inc.[2]

wicho@hi.snu.ac.kr, donghyun.kwak@navercorp.com, jwyoon@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

Speech is one of the most effective means of communication and is full of information that helps the transmission of utterer's thoughts. However, mainly due to the cumbersome processing of acoustic features, phoneme or word posterior probability has frequently been discarded in understanding the natural language. Thus, some recent spoken language understanding (SLU) modules have utilized end-to-end structures that preserve the uncertainty information. This further reduces the propagation of speech recognition error and guarantees computational efficiency. We claim that in this process, the speech comprehension can benefit from the inference of massive pre-trained language models (LMs). We transfer the knowledge from a concrete Transformer-based text LM to an SLU module which can face a data shortage, based on recent cross-modal distillation methodologies. We demonstrate the validity of our proposal upon the performance on Fluent Speech Command, an English SLU benchmark. Thereby, we experimentally verify our hypothesis that the knowledge could be shared from the top layer of the LM to a fully speech-based module, in which the abstracted speech is expected to meet the semantic representation.
Index Terms: spoken language understanding, pretrained language model, cross-modal knowledge distillation

1. Introduction

Speech and text are two representative medium of language. Speech, which is delivered mainly via waveform, can be projected to text with the help of automatic speech recognition (ASR). On the contrary, the text is represented visually in letters and is easily digitized to Unicode. It is deemed a lot more beneficial to use text in language comprehension, due to its transmission of information being less uncertain.

Despite the shared semantic representation between those two [1], especially in engineering studies, they are treated as the data of different modality. In this regard, in contemporary speech-based natural language understanding (NLU) and slot filling tasks, main approaches have exploited either ASR-NLU pipeline [2] or end-to-end speech processing [3, 4, 5]. The former, which is conventional, is partially improvable and explainable, while the latter is in fashion since it can mitigate the effect of ASR errors that can be cascaded.

In this paper, we combine the two approaches in a cross-modal viewpoint. Given original speech, its ground truth script, and the target intent, we transfer knowledge from the inference process of the pre-trained language model (LM) to the speech understanding (Figure 1). The core idea is setting a meeting place for the representation from the acoustic data and that from the digitized text, in other words, where the phonetic and lexical data coincide in terms of semantics. In this way, we compensate for the roughness of low-level processing of speech engineering, at the same time benefiting from the text-based inference.

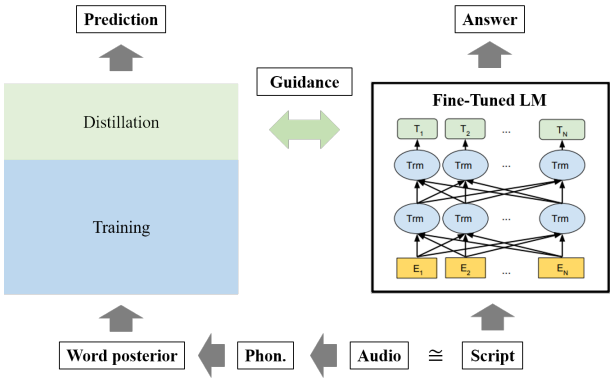

Figure 1: *A brief architecture of the proposed distillation scheme on an end-to-end SLU module. The diagram on the right side is adopted from [6].*

The contribution of this study is as the followings:

- Leveraging the high-performance inference of text-based fine-tuned LM to an end-to-end spoken language understanding via cross-modal knowledge distillation (KD)
- Verifying the effect of KD with the performance on widely-used intent identification and slot-filling dataset
- Suggesting the loss function and KD weight scheduling that can be effective in speech data shortage scenarios

2. Related Work

Comprehending the directive utterances in terms of intent argument has been vastly investigated so far, whether it be a text or speech input. While the systems with either input aim to execute similar tasks, the speech-based one inevitably requires more delicate handling that owes to the signal-level features.

2.1. Conventional pipeline

In conventional settings, spoken language understanding (SLU) is divided into ASR and NLU. ASR is a procedure that transcribes speech into text, and in NLU, the resulting text is analyzed to yield the intent arguments [2]. This cascade structure has also been widely used in other spoken language processing tasks, including speech translation [7] and intention understanding [8]. It provides a transparent analysis since the modules are distinct, that one can easily recognize the issue and make module-wise enhancement. However, as [9] pointed out in the recent study on speech translation, mainly three limitations lie in the pipeline: 1) time delay of cascading, 2) parameter redundancy by module separation, and 3) amplification of ASR errors. Though solutions such as N-best are effective, it is still probable that the last factor induces performance degradation.

2.2. End-to-end approaches

To cope with the disadvantages above, in up-to-date SLU, the inference has been performed in an end-to-end manner, wrapping up the ASR and NLU process. Advanced from the early approaches that directly infer the answer from signal level features [10] or jointly trains ASR and NLU components [3], recent ones use word posterior-level [4] or phoneme posterior-level [5] pre-trained modules to deal with the shortage of labeled speech resources. The amount of abstraction differs, but the approaches above share the ultimate goal of correctly inferring the argument, usually via slot-wise intent classification.

2.3. Pre-trained language models

Lately, a recurrent neural network (RNN) [11] and Transformer [12]-based pre-trained LMs [13, 6] have shown powerful performances over various tasks. Moreover, task-specific training is available by merely adding a shallow trainable layer on the top of the pre-trained module and undertaking a lightweight fine-tuning. However, so far, few end-to-end SLU approaches have taken advantage of them [14] mainly because the inference requires an explicitly text-format input, which necessitates an accurate ASR. Followingly, the task turns into a conventional pipeline problem, deterring the cross-modality.

2.4. Knowledge distillation of LMs

Though the aforementioned limitation is probable, it is a significant loss for the whole SLU inference to renounce the comprehensive and verified information processing of the pre-trained LMs. Is there any approach we can leverage the guaranteed performance? Knowledge distillation (KD) can be one solution [15]. It is widely used for model compression, but its scheme of minimizing the logit-wise difference can be adopted in the transfer [16] or cross-modal [17] learning as well. Notably for the Transformer [12]-based pre-trained LMs that occupy a massive volume, recent model compression work proposed the condensation schemes adopting bidirectional long short-term memory (BiLSTM) [18] or thinner Transformer layers [19]. In this paper, we plan to inherit them along with the philosophy of cross-modal distillation.

3. Proposed Method

The core content of our proposal is leveraging the pre-trained LM [6] to SLU via cross-modal fine-tuning, where the tuning is executed in the form of distillation [18, 19].

3.1. Motivation

In [1], it is demonstrated in detail how the spoken language and written one share knowledge in abstracting the features. Beyond the lexical features, which are a mere correspondence of a phoneme sequence, written language contains the tonal symbols (e.g., *pinyin*) or punctuation marks, which regard various prosodic features of the speech. Thus, we hypothesized that (1) the integration of both modalities affects a speech-based analysis in a positive way.

Consequently, we noted that it had been experimentally displayed that the text-level features reach a state-of-the-art performance within NLU tasks if combined with a pre-trained LM [6], while yet the speech-oriented models can get little from it. It is not unnatural to expect that (2) the speech processing can be boosted by NLU via some possible form of knowledge sharing.

In summary, taking into account (1) and (2), we aimed to transfer implicit linguistic processing in LMs (that can help understand the spoken language) to an SLU module, without an explicit process of speech-to-text transformation.

3.2. Materialization

The next step is materializing the architecture. Here we refer to two kinds of key papers, namely cross-modal KD for speech translation [20] and LM compression [18].

Cross-modal KD is an ambiguous term at a glance since it is difficult to define what the modality is. Thus, we here regard speech and text to incorporate different modality, though in our task, both lead to the same type of inference (intent understanding). Similar to [20], where a student speech translation model learns from the prediction of a teacher machine translation module, our SLU model takes advantage of the logit inference of a fine-tuned Transformer-based LM [6].

In this process, we employ detailed compressing procedures of a Transformer LM [18], both regarding the model architecture and loss functions. At the very first phase, a pre-trained LM, e.g., bidirectional encoder representation from Transformers (BERT) [6], is fine-tuned with the ground truth, eventually making up a teacher model (though with different modality). Consequently, at the end-to-end SLU training phase, which utilizes a frozen pre-trained acoustic module [21, 4], the loss function is updated with the knowledge distilled from the teacher. Here, knowledge is represented as a loss, which indicates the gap between the logit layers of both modules.

To wrap up, leveraging pre-trained LM to an end-to-end SLU in our approach includes *LM fine tuning* and *distillation from LM to SLU*.

3.3. Model construction

The final step is constructing the concrete structure of KD, where the teacher pre-trained LM [6] utilizes text input, and the student adopts a speech instance [4], while two share the same type of prediction [18]. In this process, we set rules of thumb to leverage the given structure and training resources as efficiently as possible. Since one of our aims is to make the best of verified ready-made solutions, we integrated the released structures, the specifications follow as:

- Backbone student model adopts ASR pre-trained module [21] and RNN-based intent classifier [4], which respectively yields word posterior sequence and slot-wise predictions.
- For the teacher model, the pre-trained BERT is utilized without additional modification, and the fine-tuning only exploits a freely available benchmark.
- In addition to the cross-entropy (CE) function that is used as the loss of an end-to-end SLU module, a KD loss is augmented to the total loss to reflect the influence of the teacher in the student training phase.

In sharing the knowledge, as mentioned above, the guidance is conveyed from the upper components of the fine-tuned BERT logit layers so that the student coincides with the representation that comes from the text input. Unlike the raw-text-friendly input layers of LM, we believe that the upper layers are the parts where the abstracted textual information possibly meets the spoken features.

More specifically, the shared knowledge can be represented as a *regulation* (loss function) that the teacher model gives to the

student in the training phase, which leads the tutee to a desirable direction. The notation for the total loss function is as follows:

$$L = \alpha_t * L_{ce} + \beta_t * L_{kd} \quad (1)$$

where t is a scheduling factor and $\alpha_t + \beta_t = 1$. α_t and β_t, here denoted as KD weight, are hyper-parameters that decide the influence of L_{ce} and L_{kd} respectively, which can be either fixed or dynamically updated.

Detailed on the losses, L_{ce} is a CE between the answer labels and the predicted logits of the SLU component, as in (2), where $f_{(\cdot)}$ is a logit representation and Y is the target label. L_{kd} is either a mean-squared error (MSE) or smoothed L_1 loss (MAE) between the predicted logits of SLU component and the fine-tuned BERT, adopted based on [18] and [22] respectively. In (3), D determines the type of distance (e.g., MSE, MAE):

$$L_{ce} = CE(f_{SLU}, Y) \quad (2)$$

$$L_{kd} = D(f_{SLU}, f_{BERT}) \quad (3)$$

In BERT fine-tuning, we adopt two kinds of engineering to investigate the teacher models of diverse performance. For a less accurate one, we build a fully connected (FC) layer on the top of [CLS] representation of BERT [6], while for the stronger model, we set FC layers for all the output representations of BERT and then apply a max pooling. We call the former *teacher* and the latter *professor* henceforth, considering the difference in training accuracy of both.

Furthermore, to leverage the *teacher* and *professor* model simultaneously, we mix up the loss that comes from each network to make up a hybrid case as in (4):

$$\begin{aligned} L_{kd} = (1-\gamma) * D(f_{SLU}, f_{teacher}) \\ + \gamma * D(f_{SLU}, f_{professor}) \end{aligned} \quad (4)$$

where $\gamma = 0$ denotes *only teacher* and $\gamma = 1$ *only professor*. For $0 < \gamma < 1$, *hybrid*, we apply the batch-wise intent error rate, $\gamma = err$, inspired be [23]. This implies that the *professor* models teaches more than *teacher* for the challenging samples.

4. Experiment

4.1. Dataset

Following the previous end-to-end SLU papers [4, 5, 24], we use the Fluent Speech Command (FSC) dataset proposed in [4]. It incorporates 30,874 English speech utterances annotated with three slots, namely *action*, *object*, and *location*. For example, for "*Turn the lamp off.*", we have slots filled as {*action*: decrease, *object*: lamp, *location*: none}, while "*Increase the temperature in the bedroom*" fills the *location* slot.

We adopt this dataset for three reasons; first, the amount of speakers and speech utterances is substantial, and second, the corpus incorporates fairly complex query-answer pairs; total 248 phrasings with 31 unique intents. Above all, the dataset is publicly available. These qualify the dataset for a benchmark, over other speech command datasets such as ATIS [25]. The FSC specification can be found in [4].

4.2. Implementation

In our experiment, we referred to three released implementations: (i) a full end-to-end SLU module utilizing FSC[1], (ii) a freely available pre-trained BERT-Base[2], and (iii) a recipe providing task-specific BERT-to-BiLSTM distillation[3]. With (i) as a backbone, we distilled the *thinking* of (ii) to the RNN encoder-decoder of (i) in the training phase. The overall procedure follows (iii), which performs a (text-only) BERT-to-BiLSTM distillation and reaches quite a standard (e.g., over ELMo [13]).

[1]https://github.com/lorenlugosch/end-to-end-SLU/

[2]https://github.com/huggingface/transformers

[3]https://github.com/pvgladkov/knowledge-distillation

4.2.1. Teacher training and baselines

Three types of systems are mainly considered. The first type denotes the teacher, namely *pretrained LMs (BERT) fine-tuned with the ground truth (GT) script*, which require an accurate script as an input. Teacher training was done with the whole FSC scripts, tokenized via word piece model [26] of BERT-Base, maximum length 60. For *teacher* we achieved the train error rate of 3.74%, and for *professor*, 0.19%. Both reached the test error rate of 0.00%.

For the teachers, if ASR output transcriptions are fed as input, we acquire the systems of the second type; an *ASR-NLU pipeline*, a common baseline. We did not re-train the ASR module with FSC, and instead used recently distributed Jasper [27] module with high accuracy (LibriSpeech [28] WER 3.61%). It was observed that *teacher* gets test error rate of 18.18%, while *professor* reaches 16.75%, showing slightly more robustness.

The last type of models are speech-based ones: a *word posterior-based RNN end-to-end* [4] and a *phoneme posterior-based Transformer architectures* [5]. In specific, [4] exploits the intermediate layer of ASR pretrained model, and besides, [5] trains new BERT [6] or ERNIE [29]-like networks with the phoneme posterior of the acoustic model component as an input. Both utilize the non-textual representation in a task-specific tuning, and especially [5] performs a large-scale.phone-level pretraining. Unlike [4], which we train as well in our environment, for [5], the reported results are adopted from the original paper, especially the highest among all the settings. For the test of these models, only the speech inputs are utilized.

4.2.2. The proposed

We compare the above approaches to the proposed scheme. As stated in Section 3.3, the whole process resembles [4], only with the difference in the total loss L. Mainly three factors determine L_{kd}: *who teaches*, *how the loss is calculated*, and *how much the guidance influences*. The first one regards the source of distillation, namely *teacher* and *professor*. The second is upon D, MSE or MAE. The last denotes the scheduling on α_t and β_t.

On the last topic, where α_t and β_t sets the KD weight, we perform three scheduling strategies regarding the temporal factor, namely t the epochs.

$$\begin{aligned} &(a) \quad \beta_t = err_{t,batch} (= 1 - acc_{t,batch}) \\ &(b) \quad \beta_t = exp(1-t) \\ &(c) \quad \beta_t = 0.1 * max(0, -|t-\mu|/(0.5*\mu) + 1) \end{aligned} \quad (5)$$

First one is the aforementioned err, adopted as (5a), which depends upon the training intent error rate per batch. Qualitatively, it regards well-classified samples contribute more to the training, as suggested in [23]. Second one is the exponential decay (*Exp.*), calculated as (5b) where the teacher influence falls down exponentially and mechanically depending on the epoch. The rest is the triangular scheduling (*Tri.*) which is inspired by [30], defined as (5c) for $\mu = T/2$ and T the maximum number of epochs. 0.1 was multiplied to compensate for the scale of KD loss compared to CE. Unlike *Exp.* where the teacher warms up the parameters at the very early phase, in *Tri.*, the student learns by itself at first and the teacher intervenes in the middle.

4.3. Result and discussion

4.3.1. Teacher performance

Overall, it is verified that the fine-tuned text models show significance with the ground truth script, albeit *teacher* failed to reach the performance of some end-to-end SLU models in terms of training accuracy. The text-based systems seem to face less amount of uncertain representations in the training phase. Besides, *professor* far outperforms *teacher* in training accuracy, as shown in Section 4.2.1. However, since the performance is not sustained in the ASR context, we set a baseline for ASR-NLU with the borrowed value [5] (Table 1).

4.3.2. Comparison and analysis

The results show that the distillation affects if the setting is considerate, 1.19% to 1.02% at best (Table 1). Though some do not display the enhancement probably for the sensitivity of the test set, we obtained the performance regarding phoneme BERT (1.05%) [5] for certain cases, namely utilizing *teacher* and *hybrid*. Though we could not achieve the current best-known state that adopts ERNIE (0.98%), one of ours with MAE reached slightly beyond BERT. We acquired around 15% reduced error rate via simple distillation to the vanilla SLU model.

It is notable that *professor* does not necessarily present the best teaching, in correspondence with the recent findings of [31]. It was also observed that the *professor* distillation spent much more epochs for the student to reach the fair accuracy in the training phase. In this regard, for *data shortage scenario #1*, even *hybrid* (where *professor* influences much) failed to converge, with *err* scheduling that had yielded the best performance. This implies that the distillation should be more like guidance, not just a harsh transfer, if the resource is scarce.

The decision of loss function is also the part we scrutinized in this study considering the previous research on Speech BERT [22]. It has been empirically shown that MAE can compensate for the different natures of the speech and text data. This is not significant in the *whole-data* scenario (Table 1), where the overfitting is less probable. However, in data shortage scenarios, adopting MSE failed to guarantee the usefulness of distillation as a helper, inducing degradation or collapse (Table 2). We assume that this is a matter of the boosted scale of the loss, that comes from the different levels of uncertainty of both modalities, which appears even with MAE sometimes (*scenario #2*).

4.3.3. Data shortage scenarios and scheduling

We checked that the proposed method is also useful in the case where the amount of text data dominates the speech, by restricting the usage of speech-text pairs to 10% and 1% in the training phase (Table 2). Given the identical test set for all the scenarios, the amount of error reduction became more visible as the data decreased. For instance with *teacher*, MAE, and *Exp.*, we obtained 0.9%p error rate reduction for *whole-data*, 0.16% for 10% scenario, and 0.44% for 1% scenario.

Under shortage, the scheduling seems to matter more than the *whole-data* case. At first we suspected that *err* or *Tri.* would show considerable performance. However, for the both scenarios, exponential decay (*Exp.*) exhibited the significance compared to the others, given MAE and *teacher* distillation. This means that early influence and fading away can lead the student to better direction if the resource is not enough (*Exp.* $>$ *err*, *Tri.*). The teaching should be moderate (*teacher* $>$ *hybrid*), and the transfer of loss should be restricted in some circumstances (e.g., β_t = *err* in *scenario #2*) to prevent the collapse.

Test error rate (%)	***Reported & done***			
ASR-NLU	16.75 (Done) / **9.89** (Reported by [5])			
Lugosch et al. [4]	1.20 / **1.19** (Done)			
Wang et al. [5]	**1.05** (BERT) / **0.98** (ERNIE)			
Proposed	$\beta_t = 0.1$ MSE	$\beta_t = 0.5$ MSE	β_t = *err* MSE	β_t = *err* MAE
Distill-*Teacher* ($\gamma = 0$)	1.19	1.19	**1.05**	1.18
Distill-*Professor* ($\gamma = 1$)	1.18	1.19	1.13	1.18
Distill-*Hybrid* (γ = *err*)	1.13	1.13	**1.05**	**1.02**

Table 1: *Results of the whole-data scenario.*

Test error rate (%)	***MSE*** *(err)*	***MAE + Schedulings*** *err*	*Exp.*	*Tri.*
Distill-*Teacher* ($\gamma = 0$)	**1.05**	1.18	1.10	**1.05**
Distill-*Professor* ($\gamma = 1$)	1.13	1.18	1.18	1.08
Distill-*Hybrid* (γ = *err*)	**1.05**	**1.02**	1.08	1.08
Data shortage #1	***10%*** *(10 random subsets)*			
Lugosch et al. [4]	2.10 / **2.04** (Done)			
Distill-*Teacher* ($\gamma = 0$)	2.32	2.00	**1.88**	1.98
Distill-*Hybrid* (γ = *err*)	$\times$	2.06	2.01	1.98
Data shortage #2	***1%*** *(20 random subsets)*			
Lugosch et al. [4]	**17.22** (Done)			
Distill-*Teacher* ($\gamma = 0$)	$\times$	$\times$	**16.88**	17.27

Table 2: *Distillation influences in the data shortage scenarios with various scheduling schemes. We set [4] as baseline for the shortage scenarios.* $\times$ *denotes the failure of convergence.*

4.3.4. Knowledge sharing

One may ask whether the distillation is truly a sharing of knowledge, since it can be interpreted as merely supervising the student based on relatively accurate logits. Also, in view of probabilistic distribution, some outputs regarding *confident* inferences might be just considered as the hard-labeled answer itself. However, in quite a few cases, logits can reflect the extent each problem is difficult for the teacher. We believe that such information is intertwined with the word-level posterior, which incorporates the uncertainty of speech processing as well.

5. Conclusion

In this paper, we materialized the speech to text adaptation by an efficient cross-modal LM distillation on an intent classification and slot filling task, FSC. We found that cross-modal distillation works in SLU, and more significantly in speech data shortage scenarios, with a proper weight scheduling and loss function. It also appears that an uncompetitive teacher conveys more useful knowledge. As future work, we plan to decompose the layer-wise information hierarchy of pre-trained LMs that the SLU systems might leverage beyond logit-level representations.

6. Acknowledgements

This research was supported by NAVER Corp. The authors appreciate Hyoungseok Kim, Gichang Lee, and Woomyoung Park for constructive discussion. Also, the authors greatly thank Sang-Woo Lee, Kyoung Tae Doh, and Jung-Woo Ha for helping this research.

7. References

[1] F. N. Akinnaso, "On the differences between spoken and written language," *Language and speech*, vol. 25, no. 2, pp. 97–125, 1982.

[2] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," *arXiv preprint arXiv:1609.01454*, 2016.

[3] 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," *arXiv preprint arXiv:1809.09190*, 2018.

[4] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[5] P. Wang, L. Wei, Y. Cao, J. Xie, and Z. Nie, "Large-scale unsupervised pre-training for end-to-end spoken language understanding," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7999-8003.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation," *arXiv preprint arXiv:1802.03142*, 2018.

[8] W. I. Cho, J. Cho, W. H. Kang, and N. S. Kim, "Text matters but speech influences: A computational analysis of syntactic ambiguity resolution," *arXiv preprint arXiv:1910.09275*, 2019.

[9] Y. Liu, J. Zhang, H. Xiong, L. Zhou, Z. He, H. Wu, H. Wang, and C. Zong, "Synchronous speech recognition and speech-to-text translation with interactive decoding," *arXiv preprint arXiv:1912.07240*, 2019.

[10] Y. Qian, R. Ubale, V. Ramanaryanan, P. Lange, D. Suendermann-Oeft, K. Evanini, and E. Tsuprun, "Exploring asr-free end-to-end modeling to improve spoken language understanding in a cloud-based dialog system," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 569–576.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[14] Z. Zhang, Z. Zhang, H. Chen, and Z. Zhang, "A joint learning framework with bert for spoken language understanding," *IEEE Access*, vol. 7, pp. 168 849–168 858, 2019.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[16] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[17] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[18] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from bert into simple neural networks," *arXiv preprint arXiv:1903.12136*, 2019.

[19] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[20] Y. Liu, H. Xiong, Z. He, J. Zhang, H. Wu, H. Wang, and C. Zong, "End-to-end speech translation with knowledge distillation," *arXiv preprint arXiv:1904.08075*, 2019.

[21] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[22] Y.-S. Chuang, C.-L. Liu, and H.-Y. Lee, "Speechbert: Cross-modal pre-trained language model for end-to-end spoken question answering," *arXiv preprint arXiv:1910.11559*, 2019.

[23] H.-G. Kim, H. Na, H. Lee, J. Lee, T. G. Kang, M.-J. Lee, and Y. S. Choi, "Knowledge distillation using output errors for self-attention end-to-end models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6181–6185.

[24] E. Palogiannidi, I. Gkinis, G. Mastrapas, P. Mizera, and T. Stafylakis, "End-to-end architectures for asr-free spoken language understanding," *arXiv preprint arXiv:1910.10599*, 2019.

[25] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[26] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[27] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, and R. T. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," *arXiv preprint arXiv:1904.03288*, 2019.

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

[29] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, "Ernie: Enhanced language representation with informative entities," *arXiv preprint arXiv:1905.07129*, 2019.

[30] J. Yang, M. Wang, H. Zhou, C. Zhao, Y. Yu, W. Zhang, and L. Li, "Towards making the most of bert in neural machine translation," *arXiv preprint arXiv:1908.05672*, 2019.

[31] J. W. Yoon, H. Lee, H. Y. Kim, W. I. Cho, and N. S. Kim, "TutorNet: Towards flexible knowledge distillation for end-to-end speech recognition," *arXiv preprint arXiv:2008.00671*, 2020.