INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

End to end transformer-based contextual speech recognition based on pointer network

Binghuai Lin[⋆] *, Liyuan Wang* [⋆]

Smart Platform Product Department, Tencent Technology Co., Ltd, China
{binghuailin, sumerlywang}@tencent.com

Abstract

Most spoken language assessment systems rely on the text features extracted from the automatic speech recognition (ASR) transcripts and thus depend heavily on the accuracy of the ASR systems. Automatic speech scoring tasks such as reading aloud and spontaneous speech are commonly provided with the prompts in advance to guide test takers' answers, which contain information that should be included in the answers (e.g., listening passage, and sample response). Utilizing these texts to improve ASR performance is of great importance for these tasks. In this paper, we develop an end-to-end (E2E) ASR system incorporating contextual information provided by prompts. Specifically, we add an extra prompt encoder to a transformer-based E2E ASR system. To fuse the probabilities of the ASR output and the prompts dynamically, we train a soft gate based on the pointer network with carefully constructed prompt training corpus. We experiment the proposed method with data collected from English speaking proficiency tests recorded by Chinese teenagers from 16 to 18 years old. The results show the improved performance of speech recognition with a nearly 50% drop in word error rate (WER) utilizing prompts. Furthermore, the proposed network performs well in rare word recognition such as locations and personal names.

Index Terms: speech recognition, end-to-end, transformer, pointer network, contextual information

1. Introduction

Automatic speech scoring has become an essential and popular tool for computer-assisted language learning (CALL). It is initially aimed at evaluating the second language (L2) learners' pronunciation in restricted testing tasks such as Repeat Sentences and Read-by-Words. With the development of ASR technology, automatic speech scoring has also been transferred to more open-ended testing tasks such as Questions on Image and Detailed Response to Topic [1]. These speech scoring tasks commonly depend heavily on the text features generated from the ASR transcripts. It has been shown the ASR errors have an enormous impact on the performance of an automated speech assessment [2]. Some prompted texts, related topics, or reference answers are commonly provided in restricted spoken tasks such as reading aloud and sentences verbatim, and less restricted tasks such as oral composition, leading to the limited actual responses of the test takers. One example of prompts in the spoken test is shown in the Figure 1. Utilizing these prompts to improve the performance of ASR is consequential and feasible for better automatic speech scoring.

With the development of deep neural network, the E2E ASR systems have prevailed in recent studies. Instead of separate models including the acoustic model (AM), the pronunciation model (PM), and the language model (LM) in traditional speech recognition systems, the E2E ASR systems optimize them jointly using a single network. The E2E models can be classified into three categories: (1) connectionist temporal classification (CTC) model [3]; (2) recurrent neural network (RNN) transducer [4]; and (3) attention-based models [5][6]. Recently, attention-based models have achieved state-of-art results in speech recognition. The self-attention-based network called transformer becomes popular in ASR research due to faster training with more parallelization [7]. These E2E models simplify the training process of ASR. However, joint optimization models make the language modeling with far fewer text data, which are limited to speech transcripts of training utterance. Consequently, it leads to worse performance of E2E ASR systems in rare, context-dependent words and phrases [8].

Section1: Read aloud the following sentence.
Prompt: What do colors tell us about ourselves and the world around us?

Section2: You will have one minute to prepare and another minute to talk about the following topics in at least five sentences.
Prompt: mother, cook, clean the room, family, have the dinner

Figure 1: *Example of spoken language testing prompts*

Many studies focus on incorporating external LMs to E2E systems. These approaches usually train an external LM and incorporate it into the E2E system. Shallow fusion combined predicted probability of the decoder and the external LM during inference [9][10]. Deep fusion concatenated the output hidden states of the attention-based system and the pre-trained LM, and fine-tuned the combining parameters [11]. Cold fusion trained the E2E systems from scratch with a fixed pre-trained LM model [12]. Instead of directly fusing a better external LM to the system, component fusion first decoupled the language modeling component with the E2E system before the fusing phase based on the cold fusion mechanism [13]. To make use of the prompts provided with the testing questions, a contextual language model conditioned on the video metadata by a hybrid pointer network was proposed and proved to improve the performance of ASR [14]. However, the separate optimization of the LM and E2E model may not achieve the best results. Other methods include direct utilization of this contextual information in the E2E systems. The previous work modified the prediction network of RNN transducer, producing output conditioned on both the previously predicted labels, as well as the keywords with an attention-based decoder [15][16]. The model called Contextual Listen, Attend and Spell (CLAS) was proposed to incorporate contextual information with an additional bias-encoder combined by an attention mechanism and trained

⋆ equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-774

jointly [17]. These studies based on RNN transducer or LAS focused on the incorporation of contextual information in the recurrent decoder called long short-term memory (LSTM), which inspire us to incorporate prompts to the transformer-based ASR system. Instead of compressing the contextual information into one vector as previous works, which may lead to disambiguation of similar sounding phrases, e.g. the prompt 'john' and 'Joan' [18], we consider incorporating the probability of individual words in the prompts into the final output of ASR.

In natural language processing (NLP), pointer network is a neural architecture that learns pointers to positions of an input sequence by the SoftMax probability distribution based on the attention mechanism [19]. It is referred to as copying mechanism in text generation and has been widely used in dialog generation [20]. It has been shown to be effective in solving the problem of rare words that have very few or no occurrences in training data [21]. It is quite likely that responses of the spoken language testing contain rare words during automatic speech scoring, and pointer network can be helpful to facilitate rare word recognition in such situations.

In this paper, we propose a transformer-based contextual E2E speech recognition incorporating prompt information. We add an extra encoder to embed prompts and fuse the output probability of the raw ASR output and prompts with dynamic weights based on the pointer network. The contextual E2E ASR model is fine-tuned based on a pre-trained ASR model based on the carefully constructed data. Experimental results show the improvement of ASR performance based on the data from automatic speech scoring. Our proposed network has three major differences compared with previous work. First, we fuse the probability of each token in the prompts to the final ASR output based on the attention mechanism instead of compressing the contextual information (prompt embedding) into one vector. Second, as the training of the prompt encoder decouples with the raw ASR training, we can merge the prompt encoder into the E2E ASR system quickly based on the carefully-crafted prompt training corpus without training ASR from scratch. Third, we incorporate the pointer network into an E2E ASR model instead of fusing or rescoring based on an additional LM. In section 2, we will introduce the proposed network. In section 3, we will show experimental setups and results, followed by some discussion. We will draw conclusion and future suggestions in section 4.

2. Proposed network

2.1. The transformer model

Transformer was first proposed and applied in machine translation [7] and has achieved competitive results in ASR compared to other E2E models [22]. It is composed of an encoder and decoder. The encoder takes frame acoustic features as input $X = \{x_1, x_2, \cdots, x_n\}$, and outputs the contextual representation of acoustic features $H = \{h_1, h_2, \cdots, h_n\}$. The decoder generates sequence output $Y = \{y_1, y_2, \cdots, y_m\}$, conditioned on the contextual representation from the encoder as shown in Eq. (1):

$$P(Y|X) = \prod_{t=1}^{m} P(y_t|H, y_1, y_2, \cdots, y_{t-1}) = \prod_{t=1}^{m} P(y_t|z_t) \tag{1}$$

The hidden state of decoder z_t is the tth representation derived from the last decoder self-attention and the encoder-decoder attention layers. The attention mechanism can be defined as Eq. (2):

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

where Q, K, V are queries, keys of dimension d_k and values of dimension d_v.

2.2. The Contextual Transformer

The E2E ASR model based on the traditional transformer is designed with acoustic features and the previously generated tokens as input. It can't utilize the information directly when given prompts. In this paper, we propose a contextual E2E ASR model, which can fuse the prompted information dynamically. The proposed contextual E2E ASR model is shown in Figure 2.

To make use of the information from prompted texts, we add an extra encoder to the raw transformer. We denote the sequence of prompts as $C = \{c_1, c_2, \cdots, c_k\}$. The sequence is embedded into the hidden representations as $H^c = \{h_1^c, h_2^c, \cdots, h_k^c\}$ by an attention-based encoder.

The information we want to introduce into the E2E ASR system is the occurrence probability of a particular word in the prompts, which is obtained by a scaled dot-product attention structure. We treat the hidden states of decoder z_t as queries, and the representations of the prompt encoder H^c as keys of dimension d_t. The final probability distribution of each token in the prompts are defined as Eq. (3):

$$P^c = \text{softmax}(\frac{z_t(H^c)^T}{\sqrt{d_t}}) \tag{3}$$

where P^c is defined as Eq. (4). k in Eq. (4) denotes the number of tokens in the prompt.

$$P^c = \{P_1^c, P_2^c, ...P_k^c\} \tag{4}$$

As the prompts may not always contain useful information for the ASR system, we employ the pointer network to dynamically fuse the output distribution of the prompts and the raw ASR output. We generate a value ranging from 0 to 1 to determine the fusion weight of the occurrence probability of the prompts and the raw ASR output. The weight is computed as Eq. (5):

$$P_t^{gen} = \sigma(W_c z_t^c + W_d z_t) \tag{5}$$

where z_t is the last hidden states of the decoder at time step t. z_t^c is the combined representation of each token in the prompt with respective weights based on the attention mechanism at tth time step defined in Eq. (6), where i is the ith token of the prompt. W_c and W_d are randomly-initialized values determining the weights of z_t^c and z_t.

$$z_t^c = \sum_{i=0}^{k} P_i^c h_i^c \tag{6}$$

The final output distribution of E2E ASR are defined as the weighted summation of the raw ASR output and prompts as shown in Eq. (7):

$$P(y_t)^{gen} = (1 - P_t^{gen}) \sum_{i:y_i=y_t} P_i^c + P_t^{gen} P(y_t|z_t) \tag{7}$$

where P_i^c is the ith output distribution of prompts, and $P(y_t|z_t)$ is the vocabulary distribution of the raw ASR output. The final output probability of the word y_t is the combination of the word y_t's raw probability from the ASR output and its corresponding probability from the prompt encoder if the word appears in the prompt.

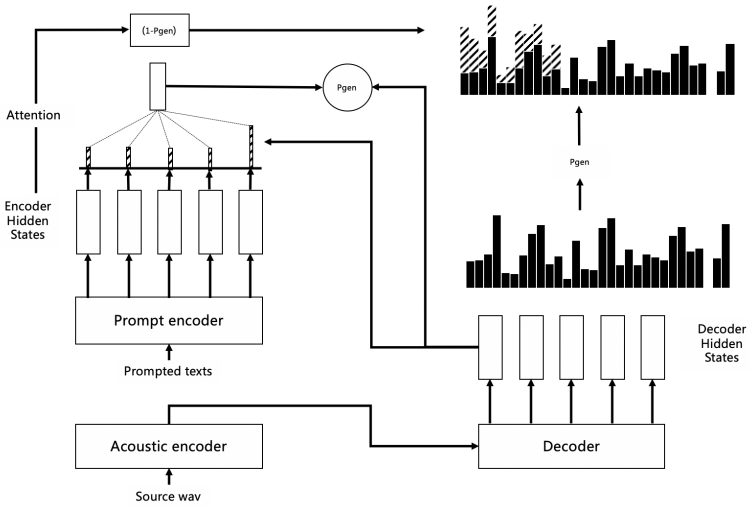

Figure 2: *The contextual E2E ASR model*

2.3. Training and inference

As the training process of an E2E ASR model is time-consuming and could be difficult to converge with structure modification, we first pre-train a vanilla E2E ASR model without the prompt encoder based on available audio-transcript data. Then we adopt a two-stage strategy to fully train the prompt encoder and make better use of the pre-trained E2E ASR model.

Specifically, the raw ASR model is pre-trained with training data consisting of paired audio-transcripts with the training loss defined as Eq. (8):

$$L_{ASR} = \sum_{t=1}^{T} -y_t \log P(y_t|z_t) \tag{8}$$

At the first stage of training the prompt encoder, we train the prompt encoder with fixed weights of encoder and decoder of the pre-trained ASR model with the same training corpus but shuffled transcripts as input to the prompt encoder. Specifically, we take the shuffled transcripts and acoustic features as input to the prompt encoder and ASR encoder, respectively, to predict corresponding correct sequences of texts from the prompts without fusion with the raw ASR output as in Eq. (3). The training loss is then defined as Eq. (9):

$$L_{prompt} = \sum_{t=1}^{T} -y_t \log P(y_t^c|y_{shuffle}^p, z_t) \tag{9}$$

where $y_{shuffle}^p$ is the shuffled transcripts corresponding to the given audio, and $P(y_t^c|y_{shuffle}^p, z_t)$ is the tth predicted distribution of the prompt as defined in Eq. (4).

Human transcripts: My name is sally where are you from
Raw ASR transcripts: My names is sallie what are you for

Methods for constructing prompt training data:
1. Shuffling human transcripts: Name my sally where is you are from
2. Correcting ASR prediction errors: Name sally where we from
3. Replacing human transcript words randomly: You apple is sallie what are you like

Figure 3: *Examples of the constructed training prompts*

At the second stage, to make the network dynamically choose the output from prompts and the raw ASR output, we train the fusion weights of the contextual E2E ASR system, which works as a soft switch between the output probabilities of them. The prompts are constructed from the raw training ASR data. Specifically, we focus on mimicking two critical conditions for the performance of our fused system. On the one hand, when the output of the raw ASR is wrong and prompts contains the ground truth, it is desirable that the soft switch is capable of putting more weights on the prompt encoder. To achieve this, we utilize particular pairs of audios and texts with significant prediction errors when using the raw ASR model, and construct prompts correcting these ASR prediction errors. On the other hand, when the output of ASR is correct and prompts are wrong, the switch needs to put more weights on the raw ASR output. It can be achieved by replacing prompts with words randomly chosen from the vocabulary table. We also freeze the weights of the raw ASR model. The loss of the second step is defined as Eq. (10):

$$L_{fuse} = \sum_{t=1}^{T} -y_t \log P^{gen}(y_t) \tag{10}$$

where $P^{gen}(y_t)$ is defined in Eq. (7). We demonstrate three methods of constructing prompt training corpus as shown in Figure 3.

During inference, we take audios and extra prompts as input to the contextual E2E ASR for prediction.

3. Experiments

3.1. Corpus description

Our training data consist of both native and non-native speech. We use the whole 960 hours of the LibriSpeech corpus [23] as the native data. The 1000 hours non-native data are collected from Chinese middle school students and hand-transcribed.

We conduct experiments based on three testing datasets: (1) synthetic data; (2) data from one actual spoken test for Chinese middle school students; and (3) data consisting of rare words such as location and personal names.

We synthesize data in four different ways for comprehensive testing. The raw test data are composed of $1,000$ audios

with an average length of 30 seconds and human transcripts. We construct four datasets with audios and four different kinds of prompts: (1) true reference prompts (True Ref), which are the corresponding transcripts or keywords for the chosen audios with significant word errors when predicted by the raw ASR model; (2) wrong reference prompts (Wrong Ref), which are constructed from words randomly chosen from the vocabulary and not included in the corresponding transcripts; (3) shuffled reference prompts (Shuffled Ref), which are constructed from shuffled corresponding transcripts; and (4) mixed reference prompts (Mixed Ref), which are constructed by mixing true references and wrong references. Examples of constructed prompts are shown in Figure 4.

Transcripts:
My name is sally where are you from

Constructed Prompts:
True Ref: Name sally where
Wrong Ref: His what kind apple bad error
Shuffled Ref: Name sally where my you are from is
Mixed Ref: sally where what nose child

Figure 4: *Examples of the constructed testing prompts*

To test model performance in spoken test tasks, data from one typical spoken test of Chinese middle school students are collected and transcribed. The testing provides multiple keywords and reference answers to limit the responses of oral presentation. The data are composed of 500 sentences with average length of 60 seconds. To test model performance for rare words, we collect $1,000$ sentences, each of which contains at least one location or personal name.

3.2. Experiments and results

Our experiments are performed using the ESPnet E2E speech processing toolkit [24]. The transformer based ASR consists of a 12-layer encoder and a 6-layer decoder. The dimension of attention is 512 and the number of attention heads is 8. The prompt encoder is composed of a 2-layer transformer encoder.

To further validate the performance of the proposed network, we compare our network with the shallow fusion [9][10]. The shallow fusion of E2E ASR model fuses the pre-trained LM at the beam-search decoding stage of transformer as defined in Eq. (11). As the E2E ASR model used in our experiments is trained based on subword units, the LM is trained based on subword units as well.

$$y_t = argmax(\log P(y_t|z_t) + \lambda \log P_{LM}(y_t)) \qquad (11)$$

The experimental results based on synthetic data are shown in Table 1. From the results, we can see a nearly 50% drop in WER when given correct prompts, indicating effectiveness of our proposed model. The soft switch in our proposed model works well when given mixed and wrong prompts. The small gap between the results with True Ref and Shuffled Ref indicates insignificant impact of the word order in a prompt.

We conduct experiments based on the dataset collected from the spoken English testing. We compare our results (Contextual Transformer) with the aforementioned shallow fusion. Besides WER, we also focus on the word error rate of given prompts (KWER), which are composed of keywords and key sentences. The results are shown in table 2, which demonstrate

Table 1: *Results of synthetic data*

Model	WER
Raw ASR	14.4
True Ref	7.4
Wrong Ref	14.6
Mixed Ref	8
Shuffled Ref	7.8

the better performance of our proposed model. We didn't observe sharp drops in WER and KWER as in table 1 due to the diversity of students' responses. However, it is still promising to utilize this mechanism for better downstream automatic spoken scoring.

Table 2: *Results of spoken test data*

Model	WER	KWER
Raw ASR	23.4	15.5
Shallow Fusion	23.2	12.1
Contextual Transformer	22.7	11.3

Table 3: *Results of rare words*

Model	WER	KWER
Raw ASR	22.3	20.3
Shallow Fusion	21.2	15.8
Contextual Transformer	19.6	9

It is quite common that the test topics are related to a person or location, and the ASR model usually performs worse in rare word recognition. Here we demonstrate the effectiveness of the proposed network in such situations. The results are shown in table 3. From the results, we can see the proposed network perform well in rare word recognition. We show some examples in the rare word recognition in Figure 5.

Human Transcripts: My name is sally where are you from
Raw ASR output: My name is sallie where are you from
Prompt: Sally
Contextual ASR: My name is sally where are you from

Human Transcripts: Thanks rita bye
Raw ASR output: Thanks reader bye
Prompt: Rita
Contextual ASR: Thanks rita bye

Figure 5: *Examples of rare words*

4. Conclusion

In this paper, we propose a contextual E2E ASR model based on the transformer. We utilize a prompt encoder to embed the prompts and combine it with the E2E ASR framework by a pointer network. The network is fine-tuned based on a pre-trained ASR model. Experimental results show the improved performance of contextual transformer given prompts. The proposed method can also be used to tackle the problem of rare word recognition. In the future, we will experiment the contextual transformer with more comprehensive downstream scoring tasks.

5. References

[1] J. Cheng, Y. Z. Dantilio, X. Chen, and J. Bernstein, "Automatic assessment of the speech of young english learners," in *Proceedings of the ninth workshop on innovative use of NLP for building educational applications*, 2014, pp. 12–21.

[2] J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 294–299.

[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[4] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[8] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang, "Shallow-fusion end-to-end contextual biasing." in *Interspeech*, 2019, pp. 1418–1422.

[9] I. Williams, A. Kannan, P. S. Aleksic, D. Rybach, and T. N. Sainath, "Contextual speech recognition in end-to-end neural network systems using beam search," in *Interspeech*, 2018, pp. 2227–2231.

[10] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang, "Shallow-fusion end-to-end contextual biasing," in *Interspeech*, 2019, pp. 1418–1422.

[11] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[12] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[13] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, "Component fusion: Learning replaceable language model component for end-to-end speech recognition system," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5361–5635.

[14] D.-R. Liu, C. Liu, F. Zhang, G. Synnaeve, Y. Saraf, and G. Zweig, "Contextualizing asr lattice rescoring with hybrid pointer network language model," *arXiv preprint arXiv:2005.07394*, 2020.

[15] Y. He, R. Prabhavalkar, K. Rao, W. Li, A. Bakhtin, and I. McGraw, "Streaming small-footprint keyword spotting using sequence-to-sequence models," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 474–481.

[16] M. Jain, G. Keren, J. Mahadeokar, and Y. Saraf, "Contextual RNN-T for open domain ASR," *arXiv preprint arXiv:2006.03411*, 2020.

[17] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: end-to-end contextual speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*. IEEE, 2018, pp. 418–425.

[18] U. Alon, G. Pundak, and T. N. Sainath, "Contextual speech recognition with difficult negative training examples," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6440–6444.

[19] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Advances in neural information processing systems*, 2015, pp. 2692–2700.

[20] S. He, C. Liu, K. Liu, and J. Zhao, "Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 199–208.

[21] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[22] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[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)*. IEEE, 2015, pp. 5206–5210.

[24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen *et al.*, "ESPnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.