Semantic Data Augmentation for End-to-End Mandarin Speech Recognition

Jianwei Sun, Zhiyuan Tang, Hengxin Yin, Wei Wang, Xi Zhao, Shuaijiang Zhao, Xiaoning Lei, Wei Zou, Xiangang Li

KE Holdings Inc, Beijing, China

{sunjianwei006, tangzhiyuan001, zhaoshuaijiang001, zouwei026, lixiangang002}@ke.com, hengxinyin@csu.edu.cn, 2181002081@cnu.edu.cn, zxcdlucy@163.com

Abstract

End-to-end models have gradually become the preferred option for automatic speech recognition (ASR) applications. During the training of end-to-end ASR, data augmentation is a quite effective technique for regularizing the neural networks. This paper proposes a novel data augmentation technique based on semantic transposition of the transcriptions via syntax rules for end-to-end Mandarin ASR. Specifically, we first segment the transcriptions based on part-of-speech tags. Then transposition strategies, such as placing the object in front of the subject or swapping the subject and the object, are applied on the segmented sentences. Finally, the acoustic features corresponding to the transposed transcription are reassembled based on the audio-to-text forced-alignment produced by a pre-trained ASR system. The combination of original data and augmented one is used for training a new ASR system. The experiments are conducted on the Transformer[2] and Conformer[3] based ASR. The results show that the proposed method can give consistent performance gain to the system. Augmentation related issues, such as comparison of different strategies and ratios for data combination are also investigated.

Index Terms: Speech recognition, End-to-end, Data augmentation, Transposition

1. Introduction

In recent years, end-to-end models have become a primary paradigm for automatic speech recognition (ASR) applications, generally under the framework of Connectionist Temporal Classification (CTC) loss function[4, 5], attention mechanism[6, 7, 8] or joint CTC-attention[9] alongside with a variety of neural network structures, e.g., LSTM[10], Transformer[2] and Conformer[3]. The training of end-to-end ASR systems usually consumes a large amount of data to avoid overfitting, while it's unreasonable to obtain too much expensive human-labeled speech data. Data augmentation is a quite promising approach to alleviate this problem, by increasing the quantity of speech data simply based on existed materials.

Many data augmentation methods have been proposed for ASR, typically conducted on either acoustic or linguistic level. Acoustic augmentation usually increases the variations of existed acoustic signals or features via strategies such as vocal tract length perturbation[11], noise insertion[12], speed perturbation[13], acoustic room simulator[14] and SpecAugment[1]. The resulted acoustic inputs share the same labels as original. On the other hand, linguistic augmentation generates new transcriptions firstly, then obtains corresponding acoustic inputs by splicing or synthesis. For example, neural text-to-speech synthesis has been widely used as a data augmentation approach for ASR, by introducing more acoustic and

linguistic diversity necessary for robustness or leveraging more training data for low-resource scenario [15, 16, 17, 18]. As in the work of [19], code-switching scenario was investigated and audios were spliced based on phone alignment, and also synthesized with new text generated by word translation or word insertion. When it comes to Mandarin ASR, the linguistic augmentation methods can be very different because of the specific language properties owned by Chinese. For example, Chinese verb never changes its form regardless of any tenses and there is no distinction between singular and plural for Chinese noun or pronoun. Those specific properties give many inspirations in the developing of linguistic-level data augmentation for Mandarin ASR.

In this paper, we propose a simple yet effective data augmentation technique based on semantic transposition of the transcriptions via syntax rules for end-to-end Mandarin ASR. Precisely, the transposition strategies are designed based on Chinese syntax rules to make sure the resulted sentences still conform to formal syntax, regardless of whether the original meaning changes or not. Some transposition strategies are presented, such as placing the object in front of the subject or swapping the subject and the object. For the process of data augmentation, the transcriptions are firstly segmented based on part-of-speech tags, then selected transposition strategies are conducted on them. Finally, the acoustic features corresponding to the transposed transcription are reassembled based on the frame level forced-alignment produced by a pre-trained GMM-HMM ASR system. The combination of original data and augmented one is used for training a new end-to-end ASR system. Although the splicing process of new acoustic inputs is similar to that in the work of [19], the latter one simply generates new texts by replacing the English segment of one code-switching utterance with that of another utterance, and our proposed method produces new texts by exchanging places of words with more flexible rules inside the single utterance.

The rest of the paper is organized as follows: Section 2 describes the details about semantic data augmentation, including the designing of transposition strategies and the process of data augmentation. Section 3 reports the experimental performance, and introduces some analysis including the topic of visualization, comparison of different rules and combination ratios between augmented data and original one. The paper is concluded in Section 4 with some future plan.

2. Semantic Data Augmentation

2.1. Chinese Transposition Strategies

The syntax rules of Mandarin Chinese are flexible, but they share a basic rule which can be summarized as "(attribute)

subject + [adverbial] predicate ⟨complement⟩ + (attribute) object"[20, 21]. In daily communication, Mandarin speakers often use some syntax rules to change the order of words in the sentence, such as inverted sentence. The common syntax rules of Mandarin are showed in Table 1. For example, "我今天要去公园" (I am going to the park today) can be changed to "公园我今天要去" (I am going to the park today) through the syntax rule R2, and "我今天很高兴" (I am very happy today) can be changed to "我很高兴今天" (I am very happy today) through the syntax rule R7. Some rules keep the original meaning of the utterance after transposition, like inverted sentence, while others may change the meaning, providing more acoustic and linguistic diversity for model training. For example, "我很喜欢朋友" (I like my friends very much) can be changed to "朋友很喜欢我" (Friends like me very much) via the syntax rule R1. Furthermore, if the attributes of the subject and the object are exchanged with each other, the semantics of the sentence itself will be enriched. For example, the sentence "我很帅他很丑" (I am handsome and he is ugly) can be change to "我很丑他很帅" (I am ugly and he is handsome). There are also some strategies that make the meaning of resulted text irrational but still meeting the syntax rules, e.g., "我要去公园" (I am going to the park) can be changed to "公园要去我" (The park is going to me) by syntax rule R5.

The Mandarin text can be easily expanded in terms of sentence structure and semantics through the syntax rules, alleviating the lack of training corpus to a certain extent. In this paper, the part-of-speech tagging tool CoreNLP[22] is used to segment and tag the text. Based on the presented syntax rules in Table 1, the sentence pattern and semantics of the text are expanded to enrich our linguistic corpus.

Table 1: *Common syntax rules of Mandarin Chinese*

Rules	Details
R1	Switch the position of subject and object
R2	Put the object at the beginning of the sentence
R3	Put the attributive at the beginning of a sentence
R4	Switch the positions of adjectives and adverbs next to each other
R5	Put predicate at the beginning of a sentence
R6	Swap the positions of two adjectives
R7	Put the adverbial at the beginning of a sentence

2.2. Augmentation Procedure

After transposing existed transcriptions, related speech features are resembled based on frame-level forced-alignment produced by a pre-trained ASR system. The details involve 3 steps as shown in Figure 1.

2.2.1. Framewise forced-alignment

In order to get acoustic features corresponded to the transposed transcriptions, in this paper we resemble the existed features based on framewise forced-alignment. Precisely, we can extract the acoustic segments associated with any specific words via forced-alignment, and then resulted acoustic segments in one sentence can be reordered according to the transposition of its text. There are a variety of methods to align transcriptions and acoustic features. Moreno et al. [23] obtained text-to-speech alignment by selecting words as anchor points. Similarly, Haubold and Kender [24] used a set of small acoustic units
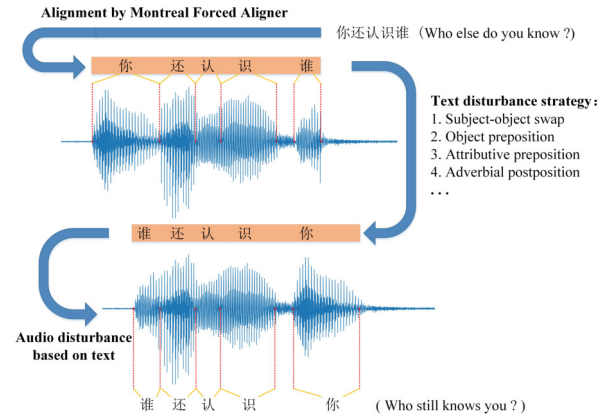

Figure 1: *The flow-chart of Semantic Data Augmentation.*

as anchor points to align audio and text. Hoffmann and Pfister [25] proposed an HMM-based sentence-level forced-alignment method through a phoneme model set. Ahmed et al. [26] used syllable information in speech and text to achieve sentence-level alignment. Martin [27] used a new generation of acoustic analysis software to complete the alignment of text and audio. In this paper, we apply commonly used GMM-HMM ASR system to achieve forced-alignment[28].

2.2.2. Part-of-speech tagging and text expanding

The part-of-speech tagging tool CoreNLP performs word segmentation on the text. Sentence components of texts are analyzed based on the part-of-speech tags. Table 2 shows the mapping relationship between the part-of-speech tags and sentence patterns. According to Table 2, the sentence pattern corresponding to each text can be found, and then the text is transposed according to the syntax rules as shown in Table 1. If the part-of-speech tag of the sentence is only a single noun or adjective without complete sentence patterns, the text should not be transposed to avoid semantic confusions or errors.

2.2.3. Feature resembling

After completing semantic augmentation for linguistic data, the segments of audios are reordered and resembled accordingly. Finally, the augmented data will be combined with existed data to train a new ASR model.

3. Experiment

3.1. Dataset

We perform the experiments on two corpora: HKUST and AISHELL-1. The HKUST was collected and transcribed by Hong Kong University of Science and Technology (HKUST), which contains 873 calls (about 150 hours) in training set, 4000 utterances in development set and 24 calls (about 5,000 utterances) in test set. The AISHELL-1 contains 120,098 utterances in training set, 14,326 utterances in developments set and 7,176 utterances in test set.

All experiments are conducted using 40-dimensional log-Mel filter bank features, computed with a 25ms window and shifted every 10ms. The features are normalized via global mean subtraction and global variance normalization.

Table 2: *The mapping relationship between part-of-speech tags and sentence patterns*

Part-of-speech	Sentence pattern
(noun) + (verb) + (noun)	subject (noun) + predicate (verb) + object (noun)
(noun) + (adverb) + (adverb) + (adjective)	subject (noun) + adverbial (adverb + adverb) + attribute (adjective)
(noun) + (adverb) + (verb) + (noun)	subject (noun) + adverbial (adverb) + predicate (verb) + object (noun)
(noun)	single noun
(adjective)	single adjective

3.2. Model

We use end-to-end ASR based on two promising neural structures, i.e. Transformer and Conformer, to verify our proposal with the tool Espnet[29].

For the two self-attention based encoder-decoder models, the encoder includes 12 layers and the decoder 6 layers. Besides, the number of filters in convolutional network for Conformer is 512, the head number of multi-head attention is 4, the dimension of input and output in feedforward network is both 512, and the dimension of inner layer in feedforward network is 2,048. The language models for Transformer and Conformer are RNN-LM and Transformer-LM respectively.

Warming-up learning rate strategy is applied to train both Transformer and Conformer. Besides, we train the Transformer models 20 epochs in total with batch size 32 and train the Conformer models 50 epochs in total with batch size 32. Considering tractability and rationality in the text transposition, the first four syntax rules showed in Table 1 are selected to do semantic data augmentation.

3.3. Performance

Table 3 shows the Character Error Rate (CER) results on the test sets of HKUST and AISHELL-1 with both Transformer and Conformer, including the state-of-the-art results from Espnet. In addition to different neural structures for acoustic and language models, we also investigate the influence of semantic data augmentation on both components respectively, i.e., both components can be trained with original data or augmented data. For end-to-end ASR, language model can even be omitted.

It can be seen that the best results are obtained with our proposed semantic data augmentation (noted as SA), and the performance can be even improved when the augmentation is applied on either the acoustic model or the language model. It can be explained that semantic data augmentation expands both acoustic and linguistic data, which gives diversity on different levels and the improvements stack on each other.

3.4. Analysis

3.4.1. Visualization

To better understand how the proposed augmentation improves the system's performance, we visualize the spectrogram of a test sample and its labels alongside with recognized ones by Transformer-based ASR with or without semantic data augmentation, as shown in Figure 2. We segment the ground-truth labels and recognized ones through the part-of-speech tagging tool CoreNLP for better comparison. We can see that the baseline model recognizes "觉得-哪部-手机" as "觉得呢-不少机", and does not accurately recognize the relatively complex part-of-speech of the sentence, while the model with semantic data augmentation can better identify the phrases such as "哪部" and "手机". Also, spectrogram shows more segments along the hor-

izontal time axis, indicating more complex syntax structure, so more similar data for training is better for recognizing this kind of samples. Overall, the improvement can be attributed that the proposed augmentation method provides more diversity on both acoustic and linguistic level.
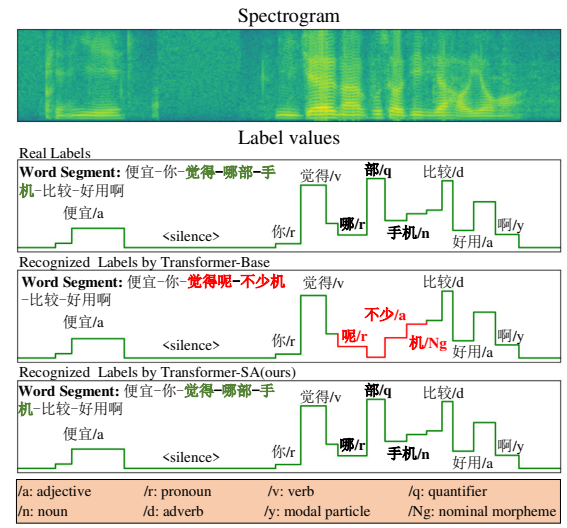

Figure 2: *Typical test sample with recognized labels by Transformer-based ASR with or without semantic data argumentation (noted as SA).*

3.4.2. Comparison of rules

We compare the impact of the selected four syntax rules for semantic data augmentation algorithm on performance of both Transformer and Conformer, as shown in Table 4 and Table 5, respectively. It can be seen that different syntax rules have different effects on the performance of different models. None of the four rules can improve the system consistently, even for the same model, while when applying the four syntax rules together, consistent performance gains are obtained. We argue that the applying of only a single rule may destroy the balance of different syntax patterns, but more syntax rules can give more syntax diversity.

3.4.3. Data combination ratio

In order to further study the influence of the impact of different syntax rules on the performance of ASR model, we investigate different data combination ratios for original data and augmented data. Table 6 shows the performance of the Transformer-based ASR with different training data combina-

Table 3: *CER(%) with or without semantic data augmentation (SA) using different acoustic and language models.*

Acoustic model	Language model		HKUST		AISHELL-1	
	structure	text	dev	test	dev	test
Transformer[30]	RNN	Raw	None	23.7	6.0	6.7
Transformer(ours)	RNN	None	22.1	21.8	6.5	7.0
		Raw	19.8	20.9	**6.1**	7.0
		SA	21.8	20.8	6.3	6.8
Transformer-SA(ours)	RNN	None	20.8	21.9	6.3	7.1
		Raw	20.4	21.1	6.3	6.9
		SA	**19.4**	**20.7**	**6.1**	**6.6**
Conformer(Espnet)	Transformer	Raw	None	None	4.3	4.6
Conformer(ours)	Transformer	None	16.7	18.6	5.0	5.5
		Raw	16.3	18.5	4.3	4.6
		SA	16.6	18.6	5.0	5.5
Conformer-SA(ours)	Transformer	None	16.6	18.6	4.5	4.9
		Raw	16.4	18.5	4.5	4.9
		SA	**15.9**	**18.3**	**4.2**	**4.5**

Table 4: *CER(%) of four different syntax rules for semantic augmentation algorithm on Transformer.*

R1	R2	R3	R4	HKUST		AISHELL-1	
				dev	test	dev	test
✗	✗	✗	✗	19.8	20.9	**6.1**	7.0
✓	✗	✗	✗	20.1	**20.7**	6.2	**6.5**
✗	✓	✗	✗	20.2	20.9	6.5	6.7
✗	✗	✓	✗	20.1	21.0	**6.1**	7.0
✗	✗	✗	✓	19.8	20.9	**6.1**	6.6
✓	✓	✓	✓	**19.4**	**20.7**	**6.1**	6.6

Table 5: *CER(%) of four different syntax rules for semantic augmentation algorithm on Conformer.*

R1	R2	R3	R4	HKUST		AISHELL-1	
				dev	test	dev	test
✗	✗	✗	✗	16.3	18.5	4.3	4.6
✓	✗	✗	✗	15.7	18.5	**4.2**	4.6
✗	✓	✗	✗	**14.7**	18.7	**4.2**	4.6
✗	✗	✓	✗	16.3	18.6	4.3	4.6
✗	✗	✗	✓	16.6	19.2	4.3	4.7
✓	✓	✓	✓	**15.9**	**18.3**	**4.2**	**4.5**

tion ratios under the four syntax rules. It can be further confirmed that different syntax rules for semantic data augmentation may have different impact, but a relatively balanced ratio setup often results in better performance.

4. Conclusions

In this work, we proposed a semantic data augmentation algorithm for end-to-end Mandarin speech recognition. This algorithm uses part-of-speech tagging tool CoreNLP to segment the transcriptions, and then applies transposition strategies, such as placing the object in front of the subject or swapping the subject and the object, on the segmented sentences to give more diver-

Table 6: *Comparison of CER(%) under different data combination ratios with Transformer.*

Data combination ratios					HKUST		AISHELL-1	
Raw	R1	R2	R3	R4	dev	test	dev	test
1	0	0	0	0	19.8	20.9	**6.1**	7.0
0.8	0.2	0	0	0	20.1	**20.7**	6.2	**6.5**
0.8	0	0.2	0	0	20.2	20.9	6.5	6.7
0.8	0	0	0.0	0	20.1	21.0	**6.1**	7.0
0.8	0	0	0	0.2	19.8	20.9	**6.1**	6.6
0.8	0.05	0.05	0.05	0.05	**19.4**	**20.7**	**6.1**	6.6
0.7	0.15	0	0	0.15	19.8	20.8	**6.1**	6.6
0.4	0.2	0.2	0.1	0.1	21.2	21.8	7.3	7.9

sity on the linguistic level. Finally, new acoustic features with more variations can be reassembled based on framewise forced-alignment produced by a pre-trained GMM-HMM ASR system. The method can produce consistent performance gain for different end-to-end ASR systems, i.e., Transformer and Conformer based systems in this paper. In the future, we will investigate more complex syntax rules and even grammar rules for semantic data augmentation, and apply the proposed method to other languages, such as English.

5. References

[1] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[2] 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.

[3] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[4] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[6] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *arXiv preprint arXiv:1506.07503*, 2015.

[7] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," *arXiv preprint arXiv:1508.01211*, 2015.

[8] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[9] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[10] A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *2013 IEEE workshop on automatic speech recognition and understanding*. IEEE, 2013, pp. 273–278.

[11] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, 2013.

[12] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," in *INTERSPEECH*, 2017.

[15] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 996–1002.

[16] R. Gokay and H. Yalcin, "Improving low resource turkish speech recognition with data augmentation and tts," in *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*. IEEE, 2019, pp. 357–360.

[17] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7069–7073.

[18] C. Du and K. Yu, "Speaker augmentation for low resource speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7719–7723.

[19] C. Du, H. Li, Y. Lu, L. Wang, and Y. Qian, "Data augmentation for end-to-end code-switching speech recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 194–200.

[20] Y. Shi, *The establishment of modern Chinese grammar: The formation of the resultative construction and its effects*. John Benjamins Publishing, 2002, vol. 59.

[21] C. Ross and J.-h. S. Ma, *Modern Mandarin Chinese grammar: A practical guide*. Routledge, 2017.

[22] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[23] P. J. Moreno, C. Joerg, J.-M. V. Thong, and O. Glickman, "A recursive algorithm for the forced alignment of very long audio segments," in *Fifth International Conference on Spoken Language Processing*, 1998.

[24] A. Haubold and J. R. Kender, "Alignment of speech to highly imperfect text transcriptions," in *2007 IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 224–227.

[25] S. Hoffmann and B. Pfister, "Text-to-speech alignment of long recordings using universal phone models." in *INTERSPEECH*, 2013, pp. 1520–1524.

[26] I. Ahmed, S. K. Kopparapu, T. Innovation, L. Mumbai, Y. Park, and T. West, "Technique for automatic sentence level alignment of long speech and transcripts." in *INTERSPEECH*, 2013, pp. 1516–1519.

[27] P. Martin, "Winpitchpro-a tool for text to speech alignment and prosodic analysis," in *Speech Prosody 2004, International Conference*, 2004.

[28] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2011.

[29] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 552–568.

[30] H. T. H. T. K. Shigeki, C. Nanxin and et al, "A comparative study on transformer vs rnn in speech applications," *arXiv preprint arXiv:2005.08100*, 2019.