

# Multilingual Approach to Joint Speech and Accent Recognition with DNN-HMM Framework

Yizhou Peng\*, Jicheng Zhang\*, Haobo Zhang\*, Haihua Xu†, Hao Huang\*, Sheng Li‡ and Eng Siong Chng†

\* School of Information Science and Engineering, Xinjiang University, Urumqi, China

† School of Computer Science and Engineering, Nanyang Technological University, Singapore

‡ National Institute of Information and Communications Technology, Kyoto, Japan

**Abstract**—Human can recognize speech, as well as the peculiar accent of the speech simultaneously. However, present state-of-the-art ASR system can rarely do that. In this paper, we propose a multilingual approach to recognizing English speech, and related accent that speaker conveys using DNN-HMM framework. Specifically, we assume different accents of English as different languages. We then merge them together and train a multilingual ASR system. During decoding, we conduct two experiments. One is a monolingual ASR-based decoding, with the accent information embedded at phone level, realizing word-based accent recognition (AR), and the other is a multilingual ASR-based decoding, realizing an approximated utterance-based AR. Experimental results on an 8-accent English speech recognition show both methods can yield WERs close to the conventional ASR systems that completely ignore the accent, as well as desired AR accuracy. Besides, we conduct extensive analysis for the proposed method, such as transfer learning with out-domain data exploitation, cross-accent recognition confusion, as well as characteristics of accented-word.

## I. INTRODUCTION

Automatic speech recognition (ASR) has been significantly improved thanks to the deep neural network techniques [1], [2], [3], [4]. However, there still remains challenges in speech recognition area. For instance, state-of-the-art ASR performance would be severely degraded when it recognizes noisy speech [5], [6]. Another limitation is that most present ASR systems are only capable of recognizing monolingual speech. However, as globalization trends are deepened, multilingual ASR systems [7], [8], [9] are greatly required.

Except for the linguistic content that is recognized by ASR system, speech also conveys other important information about the speaker, such as voice, emotion, gender, accent information etc. As a result, recognition of such information is also widely studied [10], [11], [12], [13], [14], [15], [16] in speech research area.

Though speech related research is widely studied, few efforts are focused on combining different recognition tasks as mentioned to study jointly. For example, one rarely combines speech recognition with gender recognition, or combines speech recognition with accent recognition. More often

than not, they are studied separately. This contradicts our human speech recognition behavior. Intuitively, while human recognize speech content, they can also recognize the gender or accent, or even speaker emotion of the incoming speech simultaneously.

In this paper, we propose a joint approach to performing speech and accent recognition using DNN-HMM modeling framework, namely TDNNf [17]. One of the advantages for the proposed method is that it can be easily deployed for real-time speech and accent recognition simultaneously. Our RTF is  $\sim 0.4$  under normal 2.6 GHz CPU setting. Meanwhile, we also submit another paper using Transformer-based End-to-end (E2E) approach [18]. Though the E2E approach can yield better results taking advantage of out-domain data, the main limitation of the method is much slower than TDNNf method,

The paper is organized as follows. Section II is to review prior related work. Section III presents the proposed joint speech and accent recognition method. Section IV is our data specification. Section V briefs the overall experimental setup and Section VI presents the results of the proposed methods. After that, we perform further analysis in Section VII, and draw conclusions in Section VIII.

## II. RELATED WORK

Recently, multilingual ASR has been drawing wide attention, but most of the works are only concerned with bilingual case, such as code-switching ASR [19], [20], [21]. In this paper, our ASR system is a multilingual ASR system, which is instead built with 8 accented English corpora. Besides, different from code-switching ASR, our ASR system is trained with utterances with only single accent, which is beneficial to utterance-based accent recognition. For accent recognition, the goal is to classify speaker's accent according to their pronunciation peculiarity. One can perform the task using speaker or gender identification method, such as i-vector [10], x-vector [11] or End-to-End (E2E) neural network classifier based methods [12], [22].

Since E2E attention-based framework becomes popular, multilingual ASR is getting much simpler as the number of the grapheme-letters/characters for each language is quite limited (e.g. English has only 26 letters). One can simply merge the grapheme-letters/characters of different languages as recognition output. For instance, it is reported a 50-language

Students (\*) have joined SCSE MICL Lab, NTU, Singapore as exchange students. This work is supported by the National Key R&D Program of China (2020AAA0107902); Opening Project of Key Laboratory of Xinjiang Uyghur Autonomous Region, China (2020D04047); Natural Science Foundation of China (62137002, 61663044, 61761041). Hao Huang is correspondence author.

multilingual ASR system is built with E2E method in [7]. However, it is only focused on speech recognition task. Also using E2E framework, [23] proposed a method to improve multi-dialect speech recognition performance by using dialect information as an external input vector in the LAS ASR model. For joint systems, [9] proposed a joint language identification and speech recognition method on 10 different languages. Its ASR results are impressive but far from the state-of-the-art on individual languages. Besides, its language identification results are not compared with conventional i-vector, x-vector, or other E2E methods. Furthermore, [24] implemented the same structure as [9] and investigated its performance on many more languages.

Different from prior work, we employ DNN-HMM to realize a joint speech and accent recognition system. The framework is only a multilingual ASR system, by appending grapheme with different accent identifier (we are using position-dependent grapheme lexicon [3]). To be thorough, we not only compare our ASR performance with conventional monolingual DNN-HMM ASR system, we also compare our accent recognition performance with x-vector-based classifier. Besides, we also conduct a series of analysis on the proposed methods under different scenarios.

### III. JOINT SPEECH AND ACCENT RECOGNITION

Technically, we can think of each accent as different language, hence given a normal English word, we make the word and its phone labels different for each individual accent. We then merge the accented data to train a joint multilingual ASR system. From this perspective, not only can such an ASR system perform speech recognition, it can also recognize the accent.

During training, assuming word “hello” is from British accent, our ASR lexicon should have an item like “hello\_BRT h\_BRT\_WB e\_BRT l\_BRT l\_BRT o\_BRT\_WB”, realizing a British accent-based “hello”, that is represented as “hello\_BRT” to differentiate with “hello” from other accents. Here the “\_BRT” is the accent identifier, while “\_WB” is word boundary grapheme identifier as recommended in [3].

We have two methods to conduct joint recognition during decoding. One is to assume our ASR system as a monolingual system, but each word has different pronunciations, denoted as “**Mono-joint**”. Let us use word “hello” as example again, and one of its pronunciations in decoding lexicon would be “hello h\_BRT\_WB e\_BRT l\_BRT l\_BRT o\_BRT\_WB”. By this means, we demonstrate word-based accent encoding with the help of accent identifiers at phone level. When recognizing speech, we output phone, as well as word sequences. By simply counting the accent identifiers for phone/word, we get the accent of the recognized utterance. One of the drawbacks of the method lies in it cannot encode the accent on utterance level as each word in the utterance can be recognized as different accent. Therefore, we propose another method, that is, not only is the phone label different, the word label is also different, here we name

it as “**Multi-joint**”. In other words, the decoding lexicon is the same with the training lexicon, and the ASR is a multilingual ASR system. The benefit of this method is that it can approximately realize utterance-based accent encoding as the language model (LM) has no direct cross-accent n-gram ever happened.

### IV. DATA SPECIFICATION

The experimental data is from an accented English speech recognition workshop[25], sponsored by DataTang Company in China<sup>1</sup>. There are 2 challenge tracks, one is for accented English speech recognition, and the other is for accent recognition. The data contains two parts. One is in-domain accented English data, released by DataTang, including *train*, *dev* and *test* sets specified by the organizer. There are 8 accented English data sets among *train* and *dev* sets, each with  $\sim 20$  hours. The accents are American, British, Chinese, Indian, Japanese, Korean, Portuguese and Russian respectively. There are two extra accented English data sets on *test* set, which are Spanish and Canadian (they are ignored when we perform accent recognition). Table I reports the details of data specification. For clarification, Figure 1 also plots the utterance length distribution for the three data sets respectively. The other is out-domain data which is 960 hours of Librispeech<sup>2</sup>.

All data are read speech. As will be shown in Section VI, the WER is quite low for the ASR track. However, as is shown in Figure 1, majority of utterance length is less than 6 seconds (Table I also reports that the average utterance length is  $\sim 4$ s) for the three data sets. Therefore, one can imagine the accent recognition is rather challenging for such short utterances.

TABLE I  
DATA SPECIFICATION FOR THE ACCENTED ENGLISH DATA SETS  
INCLUDING TRAIN, DEV AND TEST RESPECTIVELY

	Train	Dev	Test
Total utts	124K	12k	18k
Length (Hrs)	148.51	14.50	20.95
Ave. word (per utt.)	9.72	9.66	9.00
Ave. second (per utt.)	4.29	4.35	4.15

### V. EXPERIMENTAL SETUP

All the experiments are conducted with Kaldi.<sup>3</sup> For DNN-HMM ASR framework, the acoustic models are trained with Lattice-free Maximum Mutual Information (LF-MMI) criterion [26] over the Factorized Time Delay Neural Network (TDNNf) [17]. The TDNNf is made up of 15 layers, and each layer is decomposed as 1536x512x1536, where 512 is the dimension of the bottleneck layer. To train the TDNNf, we also perform two kinds of data augmentation (DA) methods, one is speed perturbation (sp) (x3) [27], and the other is to add four kinds of noise, such as white noise, music noise, babble noise, as well as reverberant noise (x4). For baseline system, the lexicon is phonetic lexicon with  $\sim 21$ k word vocabulary.

<sup>1</sup> <https://www.datatang.com/INTERSPEECH2020>

<sup>2</sup> <http://www.openslr.org/12/> <sup>3</sup> <https://github.com/kaldi-asr/kaldi>

TABLE II  
SPEECH WER (%) AND ACCENT RECOGNITION ACC (%) RESULTS OF THE PROPOSED JOINT RECOGNITION METHODS ON DEV SET.

	Method	AVE (WER%)	US	UK	CHN	IND	JAP	KOR	POR	RUS
Speech Recognition	TDNNf	7.68	8.71	8.87	11.08	8.67	6.08	6.87	6.45	4.47
	Mono-joint	8.66	9.46	10.31	12.67	9.02	6.98	7.28	7.43	5.36
	Multi-joint	8.63	9.29	9.71	12.71	9.16	6.68	7.36	7.51	5.30
		AVE (ACC%)								
Accent Recognition	x-vector	64.86	55.05	88.24	55.44	75.25	55.28	56.04	79.58	52.91
	Mono-joint	61.70	42.64	88.43	63.76	86.82	42.54	56.65	58.97	54.95
	Multi-joint	63.19	44.53	88.48	66.31	87.66	43.88	58.71	61.51	55.63

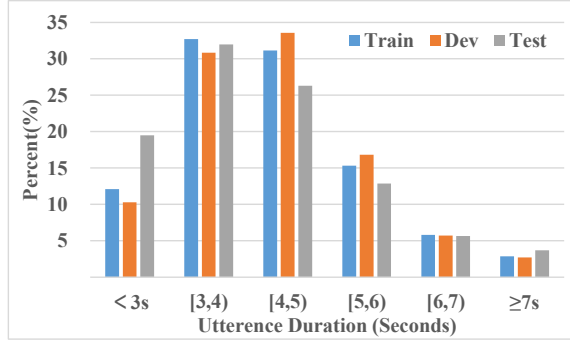


Fig. 1. Utterance length(s) distribution for accented English data sets that are train, dev and test respectively

For the multilingual ASR, the lexicon is position-dependent grapheme lexicon but with as many as eight times of the vocabulary for either two methods as mentioned. From our off-line experiments, the grapheme lexicon yields slightly worse results, compared with phonetic lexicon. We stick to use it because we want our TDNNf system to have the same advantage with E2E ASR system, that is, free of lexicon modeling, which is especially useful for multilingual ASR tasks. For the two methods, we build two sets of tri-gram language models (LMs) for decoding respectively. One is monolingual LM, and the other is a multilingual LM, labelling each word with different accent identifiers. Additionally, we also perform x-vector experiment for accent recognition as a contrast, while we choose logistic regression as the classifier. We find it consistently yields better results compared with PLDA [28] method in our case. To train the x-vector extractor, we follow the configuration of [11], and the resulting x-vector dimension is 512.

## VI. RESULTS

Table II reports the speech recognition and accent recognition results of the proposed methods, in contrast with conventional TDNNf, and x-vector systems respectively.

From Table II, we observe the proposed methods achieve worse ASR results compared with the baseline TDNNf system. This is understandable, since for the “Mono-joint” method, we have 8-times of pronunciations than the baseline system, yielding more confusion to the lexicon. Besides, the pronunciation difference for each specific accent is not fully considered. On the other hand, the TDNNf output of the “Multi-joint” method

has 8-times of senones to handle in theory, compared with the baseline TDNNf system. Such a big senone-number output is difficult to obtain by decision-tree-based clustering method, and it is also hard to learn.

For the accent recognition results in Table II, the proposed methods also yield worse accent recognition results, compared with x-vector method. However, the advantage of the proposed method lies in two aspects. First to perform accent recognition, it is easier for the proposed method to exploit out-domain data (Librispeech here) without accent label, compared with x-vector method. Such an advantage will be shown in Section VII-B. Secondly, it can perform both speech and accent recognition tasks simultaneously.

Besides, from Table II, we observe the best WERs are from Russian (RUS) and Japanese (JAP) accented speech respectively, while the worst of all is from Chinese (CHN) accented speech, which is followed by British (UK) accented speech. For the accent recognition, the best accuracy is from British accent (UK), and after that it is Indian accent (IND), while the worst is American accent (US) which is followed by the Japanese accent (JAP). One thing worth a note is that the proposed method outperforms the x-vector method with a big margin for the Chinese (CHN) and Indian (IND) accented speech respectively, while it performs much worse on the American (US), Japanese (JAP) and Portuguese (POR) accented speech. Furthermore, comparing speech recognition with accent recognition results in Table II, we note that better WER does not necessarily mean better accent recognition accuracy. For instance, for the Korean (KOR), Japanese (JAP), and Russian (RUS) accented speech, the ASR results are significantly better than the ones of remaining accents, their accent recognition results are much worse.

Finally, Table II shows that the proposed “Multi-joint” method outperforms the “Mono-joint” method for both speech and accent recognition results respectively. This suggests utterance-based accent encoding is superior than the word-based accent encoding at least for accent recognition.

## VII. ANALYSIS

### A. Joint accent acoustic models

For the above experiments, we merge 8 accented English data to train the joint TDNNf acoustic models, and each accent has its own phone set. The TDNNf models have  $\sim 8.8k$  outputs (tied-states/senones) in total. As a result, we only have about 1k outputs for each accent on average. We are curious that

TABLE III  
CROSS-ACCENT RECOGNITION CONFUSION ON DEV SET(%)

Recognition Ground Truth \	US	UK	CHN	IND	JAP	KOR	POR	RUS
US	37.45	27.77	0.07	0.00	0.35	1.82	8.06	24.47
UK	2.72	92.66	0.13	0.00	0.76	0.00	0.38	3.35
CHN	7.11	0.47	68.12	0.00	16.17	2.75	0.00	5.37
IND	2.89	0.00	0.30	82.79	0.30	0.00	13.63	0.08
JAP	0.00	0.00	0.67	0.00	49.46	27.96	0.00	21.91
KOR	0.21	0.27	3.36	0.00	24.42	65.43	3.09	3.22
POR	1.30	4.76	2.23	0.12	2.66	1.36	72.77	14.79
RUS	0.31	4.15	4.64	0.00	0.00	0.12	33.91	56.87

TABLE IV  
PERFORMANCE REPORT WITH DIFFERENT TDNNF OUTPUTS  
(TIED-STATES/SENONES), DIMENSION OF THE FINAL BOTTLENECK LAYER  
IS FIXED AT 512.

#PDF	Mono-joint		Multi-joint	
	WER%	ACC%	WER%	ACC%
4.3k	8.78	60.27	8.78	61.68
<b>8.8k</b>	<b>8.66</b>	<b>61.70</b>	<b>8.63</b>	<b>63.19</b>
14.6k	8.76	60.97	8.84	61.99
21.8k	8.76	60.97	8.80	62.28

if bigger outputs yield better results. Table IV reports our experimental results. From Table IV, bigger outputs of TDNNf don't bring performance improvement on both tasks. We also report the results of a smaller output of 4.3k where we only have less than 6 hundred output on average for each accent, and get a slightly worse performance compared with those of larger scales of outputs. Our conclusion is that TDNNf is not sensitive to the tied-state number (perhaps once it is satisfied with a minimal tied-state number). This even suggests we can use TDNNf to train the joint multilingual models with more languages in future.

### B. Transfer Learning

To employ out-domain data, we adopt transfer learning (TL) method [29], [30], [31], with Librispeech as the out-domain data. We start with a TDNNf model trained with overall out-domain and in-domain data. We then fine-tune the model with only the in-domain accented data. Figure 2 plots the curves of the recognition accuracy versus learning rate factor, with 3 epochs. From Figure 2, we observe consistent improvements on both tasks with bigger learning rate factor ( $\geq 0.5$ ).

Table V reports the results of transfer learning, and RNNLM rescoring. From Table V, we see that transfer learning is effective on both speech and accent recognition performance improvement, even outperforms x-vector (64.86%) on accent recognition. However, RNNLM rescoring only improves ASR WERs, while it yields no accuracy improvement on AR task at all. These suggest AR accuracy is closely related with the performance of the underlying acoustic model (AM), but less affected with LM. For AM, different acoustic senones are trained with different accented speech data, as a result it has capability to differentiate accents. However, for the LMs (RNNLM included), they are trained with overall transcripts,

and consequently have no accent discriminative capability on either word or utterance level.

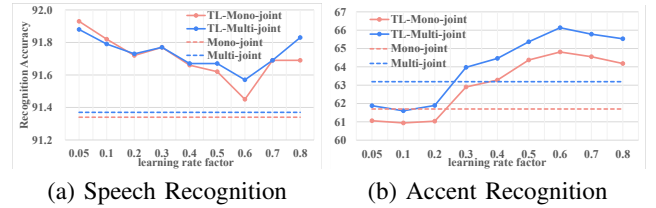


Fig. 2. TL performance versus learning rate factor

TABLE V  
RECOGNITION RESULTS WITH TL(LR=0.8) AND RNNLM RESCORING  
FOR DEV AND TEST SETS

Method	WER(%)		ACC(%)	
	dev	test	dev	test
Mono-joint	8.66	10.31	61.70	61.01
+Transfer Learning	8.31	9.83	64.18	63.70
+RNNLM Rescoring	6.93	8.28	64.29	62.66
Multi-joint	8.63	10.35	63.19	63.75
+Transfer Learning	8.17	9.74	65.53	66.68
+RNNLM Rescoring	7.05	8.40	65.35	66.40

### C. Cross-accent recognition confusion

We are also interested in how each specific accent is mistakenly recognized as another accent, that is, cross-accent recognition confusion pattern. Table III shows overall accent recognition confusion results. There are several patterns worth our notice. First, for British and Indian accents, there are minor in confusion with other accents. The most confusion counterparts for British and Indian accents are Russian and Portuguese ones, with 3.35% and 13.63% respectively. Second, Japanese and Korean accents, as well as Portuguese and Russian accents are mutually confused heavily. For instance, Japanese accent has 27.96% been miss recognized with Korean accents which in return has 24.42% been miss recognized with the Japanese accent. Finally, American accent is widely "overlapped" with other accents, particularly 27.77% with British, and followed by Russian (24.47%), and then Portuguese (8.06%) accents respectively.

### D. Accented-word characteristics

Except for accent confusion, we are also curious about if some words have stronger accented characteristics than

TABLE VI  
EXEMPLAR WORDS WITH HIGHEST AND LOWEST ACCENT RECOGNITION  
ACCURACY RESPECTIVELY

Top 10	Exemplar words	Ave. Length
most accurate	temperature, strong, feature, students, adapted, voyage, branches, value, plans, blind	7.3
least accurate	fake, key, finish minute, England, ocean, orders, sick, either, wash	5.4

others. In other words, if a word appears in some accent and it is correctly recognized in terms of accent recognition in majority of times, we think such a word has stronger accented characteristics. Table VI reveals some exemplar words in terms of highest and lowest accent recognition accuracy. We can see from Table VI, longer words or words with multiple syllables generally yield better accent recognition results. This seems to be agreed with our intuition.

### VIII. CONCLUSIONS

In this paper, we proposed a joint multilingual approach to realizing speech and accent recognition simultaneously using the DNN-HMM framework. On an 8-accent English speech recognition data set, we demonstrated the effectiveness of the proposed methods on both speech recognition and accent recognition. In addition, we adopted transfer learning to fully exploit out-domain data to boost the performance. We found that transfer learning is beneficial to the performance improvement for both tasks, particularly for the accent recognition task. We also tried RNNLM rescoring method, and achieved better WER as expected however no accent recognition improvement. This suggests accent recognition is more dependent on AM and less on LM for our data set. Besides, we also analysed cross-accent recognition confusion, as well as accented-word characteristics preliminarily, which are yet to be further studied in future.

### REFERENCES

- [1] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjali Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al., "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. of ICASSP 2018*, pp. 4774–4778.
- [2] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur, "End-to-end speech recognition using lattice-free mmi," in *Proc. of INTERSPEECH 2018*, pp. 12–16.
- [3] Duc Le, Xiaohui Zhang, Weiye Zheng, Christian Fügen, Geoffrey Zweig, and Michael L Seltzer, "From senones to chonones: Tied context-dependent graphemes for hybrid speech recognition," in *Proc. of ASRU*, pp. 457–464.
- [4] Shigeki Karita, Nelson Enrique Yalta Soplin, Shinji Watanabe, Marc Delcroix, Atsunori Ogawa, and Tomohiro Nakatani, "Improving Transformer-Based End-to-End Speech Recognition with Connectionist Temporal Classification and Language Model Integration," in *Proc. of INTERSPEECH 2019*, pp. 1408–1412.
- [5] Tian Tan, Yanmin Qian, Hu Hu, Ying Zhou, Wen Ding, and Kai Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM TASLP*, vol. 26, no. 8, pp. 1393–1405, 2018.
- [6] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP 2013*, pp. 7398–7402.

- [7] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.
- [8] Jaesin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *IEEE SLT 2018*, pp. 521–527.
- [9] Shinji Watanabe, Takaaki Hori, and John R Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *Proc. of ASRU 2017*, pp. 265–271.
- [10] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE TASLP*, vol. 19, no. 4, pp. 788–798, 2011.
- [11] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. of ICASSP 2018*, pp. 5329–5333.
- [12] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, "Self-attentive speaker embeddings for text-independent speaker verification," in *Proc. of INTERSPEECH 2018*, pp. 3573–3577.
- [13] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak, "Emotion identification from raw speech signals using dnns," in *Proc. of INTERSPEECH 2018*, pp. 3097–3101.
- [14] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Wahid Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *Proc. of ICASSP 2020*, pp. 6474–6478.
- [15] Zhong-Qiu Wang and Ivan Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. of ICASSP 2017*, pp. 5150–5154.
- [16] Maryam Najafian and Martin Russell, "Automatic accent identification as an analytical tool for accent robust automatic speech recognition," *Speech Communication*, 2020.
- [17] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of INTERSPEECH 2018*, pp. 3743–3747.
- [18] Jicheng Zhang, Yizhou Peng, Pham Van Tung, Haihua Xu, Hao Huang, and Eng Siong Chng, "E2e-based multi-task learning approach to joint speech and accent recognition," 2021.
- [19] Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, and Haizhou Li, "On the End-to-End Solution to Mandarin-English Code-Switching Speech Recognition," in *Proc. of INTERSPEECH 2019*, pp. 2165–2169.
- [20] Haihua Xu, Van Tung Pham, Zin Tun Kyaw, Zhi Hao Lim, Eng Siong Chng, and Haizhou Li, "Mandarin-english code-switching speech recognition," in *Proc. of INTERSPEECH 2018*, pp. 554–555.
- [21] Pengcheng Guo, Haihua Xu, Lei Xie, and Eng Siong Chng, "Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition," in *Proc. of INTERSPEECH 2018*, pp. 1928–1932.
- [22] Pooyan Safari, Miquel India, and Javier Hernando, "Self-Attention Encoding and Pooling for Speaker Recognition," in *Proc. of INTERSPEECH 2020*, pp. 941–945.
- [23] Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao, "Multi-dialect speech recognition with a single sequence-to-sequence model," in *Proc. of ICASSP 2018*, pp. 4749–4753.
- [24] Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki, "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," in *Proc. INTERSPEECH 2020*, 2020, pp. 1037–1041.
- [25] Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie, "The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods," *arXiv preprint arXiv:2102.10233*, 2021.
- [26] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Proc. of INTERSPEECH 2016*, pp. 2751–2755.
- [27] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Proc. of INTERSPEECH 2015*, 2015.

- [28] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y Chang, "Plda: Parallel latent dirichlet allocation for large-scale applications," in *International Conference on Algorithmic Applications in Management*. Springer, 2009, pp. 301–314.
- [29] Pegah Ghahremani, Vimal Manohar, Hossein Hadian, Daniel Povey, and Sanjeev Khudanpur, "Investigation of transfer learning for asr using lfmmi trained neural networks," in *Proc. of ASRU 2017*, pp. 279–286.
- [30] Dong Wang and Thomas Fang Zheng, "Transfer learning for speech and language processing," in *Proc. of APSIPA 2015*, 2015, pp. 1225–1237.
- [31] Zhen Huang, Sabato Marco Siniscalchi, and Chin-Hui Lee, "A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition," *Neurocomputing*, vol. 218, pp. 448–459, 2016.