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

ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech

R. Gretter[1], *M. Matassoni*[1], *D. Falavigna*[1], *A. Misra*[2], *C.W. Leong*[2], *K. Knill*[3], *L. Wang*[3]

[1] Fondazione Bruno Kessler, Trento, Italy
[2] Educational Testing Service, Princeton, USA
[3] Cambridge University, Cambridge, UK

{gretter,matasso,falavi}@fbk.eu, {amisra001,cleong}@ets.org, {kmk1001,lw519}@cam.ac.uk

Abstract

The paper presents the Second ASR Challenge for Non-native Children's Speech proposed as a Special Session at Interspeech 2021, following the successful first challenge at Interspeech 2020. The goal of the challenge is to advance research on non native children's speech recognition technology, as speech technology still struggles when applied to both children and non-native speakers. The audio data consists of spoken responses provided by L2 students in the context of both English and German speaking proficiency examinations, the latter language added for 2021. Additional training data and a new evaluation set was released for L2 English recorded by speakers of different native languages. Participants could build systems for one or both languages. Each had a closed track where a predetermined set of audio and linguistic resources were selected, and an open track where additional data was allowed. After a description of the released corpora, the paper analyzes the results achieved by the participating systems. Some issues suggested from these results are discussed.

Index Terms: speech recognition, shared task, non native children's speech

1. Introduction

Automatic speech recognition (ASR) technology has become popular and recent systems can achieve human performance in some specific cases [1]. However, ASR still struggles when applied to speech produced by specific types of speakers, including non-native speakers [2, 3] and children [4, 5, 6]. Although ASR applied to adult native speakers usually achieves very low error rates, the case of non-native children's speech, typically in the context of automated language proficiency assessments, still represents a particularly challenging scenario. Typically, L2 speakers produce phenomena such as disfluencies or code-switching [7, 8, 9, 10, 11, 12] that become more critical in the case of young speakers in a test-taking setting; e.g. [13, 14, 15] report significant error rates for open-ended spoken responses produced by K-12 English learners who took a standardized speaking assessment in the USA.

Moreover, research is more difficult due to the lack of public data-sets, even though many speech applications are used by an increasing number of non-native speakers. Therefore it is vital to improve performance of ASR systems also for these particular users.

To stimulate research that can advance the present state-of-the-art in ASR for non-native children's speech we have freely distributed two new data sets containing non-native children's speech and have organized the Special Session at Interspeech 2021 (the web page of the challenge can be found at the URL: https://sites.google.com/fbk.eu/ss-is2021-nonnativechildren/).

The data sets consist of spoken responses collected in Italian schools from students between the ages of 9 and 16 in the context of both English and German speaking proficiency assessments, the latter is new for 2021. Additionally, this year we included around 50 hours of English data from students belonging to different parts of the world and taking a global test of English proficiency. Two baseline systems (one for English and one for German), based on the Kaldi toolkit [16], were released together with the data, and a challenge web site was developed for collecting and scoring submissions.

Researchers can use the corpora to establish benchmarks in the area of non native children's speech and, more generally, in scenarios characterized by low resource data for less commonly studied populations as well as to address research topics in these sub-fields:

- Acoustic models: multilingual models exploiting L1 and L2 data, adaptation techniques using unsupervised or lightly supervised approaches, models for specific L2 speech phenomena such as disfluencies or hesitations;
- Lexica: multiple pronunciations for non-native accents and training of pronunciation models [17];
- Language models: ad-hoc models for grammatically incorrect sentences, false starts and partial words, code-switched words;
- Evaluation: acquisition and annotation of non-native speech [18].

In addition, a well-studied dataset can be useful for applications beyond ASR such as computer assisted language learning [19] or computer assisted pronunciation training [20, 21].

The metric used for evaluation and ranking of participants is the word error rate (WER). For each language, the challenge included both a **closed track** and an **open track**. In the closed track, only the training data distributed as part of the shared task could be used to train the models; in the open track, any additional data could also be used. The submission window was open for a total of 11 days and teams were allowed to provide at most one submission per day to each track, with a maximum total of 11 submissions per track per team. Participants were also able to see more detailed evaluation results (including # of insertions, # of deletions and # of substitutions) for each submission. More details about the challenge can be found at the CodaLab site[1].

[1] https://competitions.codalab.org/competitions/28890

http://dx.doi.org/10.21437/Interspeech.2021-1237

2. Audio and language resources

This section describes the data and the linguistic resources provided for the Interspeech 2021 non-native children speech recognition challenge [22].

ETLT stands for Extended Trentino Language Testing: in fact ETLT2021 is an extension of last year challenge, TLT2020. The extension is twofold: we enlarge the English dataset with new training, dev and eval data coming from another source, ETS, and we introduce a second language, German, for which the amount of available data is quite limited. Table 1 highlights this fact: while for English there are more than 100 hours of transcribed speech, coming from two sources (FBK and ETS), for German the amount of transcribed speech is about 7 hours overall, so the challenge is more focused on how to exploit untranscribed audio.

Table 1: *Some statistics, in hours, about the ETLT2021 challenge.*

	English	German
transcribed train	54 h (ETS) + 49 h (FBK)	5 h (FBK)
untranscribed train	–	64 h (FBK)
dev	3 h (ETS)	1 h (FBK)
eval	3 h (ETS)	1 h (FBK)

2.1. English data from ETS

ETS conducts a global test of English speaking proficiency for students of ages 11 and above, generally belonging to middle school or lower levels of high school. The test provides a reliable measure of students' English communication skills that can help in understanding their strengths and challenges to further improve their proficiency levels. All the students are non-native speakers of English and hence, the challenge is not only to recognise children's speech but also non-native speech at varying proficiency levels. The amount of babble or other types of background noise is less in the data as the test is conducted in a controlled environment that is administered by a proctor.

The **audio recordings** in this package are: *1)* audio/train: around 53.43 hours of data coming from 800 speakers. Each speaker has 4 recordings leading to a total of 3200 audio files. *2)* audio/dev: about 3.3 hours of data belonging to 50 speakers. There are a total of 200 audio files corresponding to 4 recordings per speaker. *3)* audio/eval: approximately 3.3 hours of audio data from 50 speakers. Similar to train and dev, there are 4 recordings available per speaker leading to a total of 200 audio files.

Table 2 provides some more information about the data. Every test taker has to answer four different types of questions leading to four separate recordings per speaker. These recordings involve both read as well as spontaneous speech. The test takers come mainly from South America, Korea, Japan and Turkey.

2.2. English data from FBK

In Trentino, a small region in northern Italy, there is a series of evaluation campaigns underway for testing L2 linguistic competence of Italian students taking proficiency tests in both English and German [23, 24]. Data were collected in 2016, 2017 and 2018, involving about 6000 students ranging from 9 to 16 years, belonging to four different school grade levels and three proficiency levels (A1, A2, B1). Part of this material is included in the challenge.

Table 2: *Some statistics on the speech data in the package; number of utterances, pupils, different questions, running words, total duration. For ETS data (*), we report only the number of different questions for each speaker (4).*

id	#Utt	#Pup	#Q	#Words	Dur
		English			
ETStrain	3200	800	4(*)	173770	53:26
TLT1618train	11700	3111	109	136482	40:11
TLT2017train	2299	338	24	22882	09:00
ETSdev	200	50	4(*)	11671	03:20
ETSeval	200	50	4(*)	11864	03:20
		German			
deTLT2017train	1445	296	23	8536	04:42
deTLT1618train	10047	2658	124	–	63:55
deTLT2017dev	339	72	23	2244	01:07
deTLT2017eval	329	72	23	2180	01:07

In the following we report an example of manual transcription of one English sentence in the corpus where it can be noticed the presence of switched words, mispronounced words (marked with #), hesitations and words fragments.

@bkg @de(hallo) @breath @e i'm from # my #hobby it is @e #football and tennis @e @ns (@it(come si dice nel tempo libero)) @ns in the eve- in the evening i play football and @em and play tennis @breath @voice @breath in my family there are my dad my mom and my sister @e i have two #friends @em @ns @e @ns*

The training **audio recordings** in this package are the same as TLT2020 challenge: • TLT1618train: about 40 hours, manually transcribed by ETS, coming from 2016 and 2018 recordings; • TLT2017train: about 9 hours, manually transcribed, coming from 2017 recordings. To avoid confusions, FBK's dev and eval data of the last challenge (TLT2020) were not provided. Some statistics about the audio data are in Table 2.

To build and evaluate language models, two **text sources** are provided: (1) manual transcriptions of the 2017 audio data and (2) written data, extracted from the written sentences provided by the pupils in 2016. Some statistics about the text data in the package are reported in Table 3.

Table 3: *Some statistics on the text data provided in the package: running words, lexicon size, input modality.*

id	Running Words	Lex Size	mode
	English		
2016Wtrain	185777	3385	written
2017train	22450	1493	spoken
	German		
de16W18Wtrain	234842	3211	written
deTLT2017train	7590	749	spoken

Phonetic lexica, based on the CMU dictionary, are provided for the training part of this package. In addition, automatic procedures provide phonetic transcriptions of unknown words. Note that in the corpus also Italian as well as German words are present, for which the English transcriptions are not accurate. Therefore, two additional lexica are provided, for Italian and German words, using SAMPA units.

2.3. German data

German data are similar to the English ones. In fact, most of the pupils took part both to the English and to the German TLT evaluation campaigns; the German task is more difficult just because the amount of transcribed material is much less: about 7 hours which include training, dev and eval. In addition, looking at the automatic transcriptions of the untranscribed training set, we observed that about half of the German data consists of background noise, thus significantly reducing the amount of signal useful for training. Finally, note that we decided to keep the same speaker in the same group for both languages: so English training speakers overlap with German training speakers. This fact could be exploited, in the **open condition**, to enlarge the training amount of data in a language independent fashion. Some statistics about the audio data are in Table 2.

As for English, to build and evaluate language models, two **text sources** are provided: (1) manual transcriptions of the 2017 audio data and (2) written data, extracted from the written sentences provided by the pupils in 2016 and 2018. Lowercase texts are obtained after a cleaning phase. Statistics are reported in Table 3. To exploit language dependent lexicons, a DEI (Deutsch, English, Italiano) version of the data was also provided, for which each word is preceded by the (automatically assigned) language identifier (en_ and it_ - de_ is omitted).

Concerning **phonetic lexica**, we provided 2 of them: in the first, each word is assigned a language (it_, en_, or German) and the corresponding phonetic transcriber is applied. In this way the lexicon includes 3 phonetic sets, partially overlapping. In the second lexicon, all Italian and English phones are arbitrarily mapped to some German phone, resulting in a lexicon which uses only German phones.

3. Released baselines

3.1. English baseline

The English baseline is provided as a Kaldi recipe [16]. This system used 48 hours of the acoustic model (AM) training data from ETLT_2021_ETS_EN. The AM is a chain model trained with lattice-free maximum mutual information (LF MMI) [25], and consists of 6 initial Convolutional Neural Network (CNN) layers followed by 9 factorised TDNN layers (TDNN-F) [26] of sizes 1024.

The input features are 40-d high-resolution MFCC, and 100-d online extracted i-vectors provide speaker adaptation. In addition to 3-way speed perturbation, SpecAugment [27] is also applied, where time and frequency bands of the spectrogram are randomly masked out in training, with the proportion of frequency and time frames bands zeroed out set to 0.5 and 0.2, respectively.

A four-gram language model (LM) is built using the AM training data transcriptions. Phonetic pronunciations are used from the CMU lexicon, with a G2P system trained on Phonetisaurus [28]. WERs of 26.08% and 33.21% were achieved on the provided English *dev* and *eval* datasets, respectively. In the baseline directory a script is also supplied (*local/data_prep_all.sh*) to prepare all of the supplied data for AM and LM training in a consistent fashion.

3.2. German baseline

The German baseline is provided as a Kaldi recipe; a seed acoustic model is initially trained on a limited amount of manually transcribed spoken responses (less than 5 hours), using a standard *chain* model using LF-MMI while a n-gram language model is trained on the written answers collected in 2016 and 2018 campaigns along with the transcriptions of the released audio material.

A multi-lingual lexicon is used to model also words not belonging to the expected language (e.g. in Italian or English) in order to exploit as much as possible the available speech. The preliminary model is then used to create alignments for the unsupervised material: the final TDNN model is trained with a weighted combination of supervised and unsupervised data; the procedure follows the recipe proposed in [29]. Table 4 reports the baseline results on *dev* and *eval* datasets.

Table 4: *Baseline WER results for (closed) German track.*

German track	Dev	Eval
baseline	50.51	48.08
+ 64h untranscribed	47.03	**45.21**

4. Challenge conditions and evaluation

The shared task consists of two tracks for each one of the two languages, English and German: a closed track and an open track. In the closed track, only the training data distributed as part of the shared task can be used to train the models; in the open track, any additional data can be used to train the models. The evaluation of the challenge is performed in terms of Word Error Rate (WER), after removing • fragment words, • spontaneous speech phenomena, • extraneous speech, • speech not in the target language.

5. Results and discussion

The results in terms of WER, for all of the tracks (both the closed and open) are provided in Table 5. For comparison purposes the Table summarizes some of the features of the submitted systems of which we are aware (this information was provided by the participants in response to an informal survey after the completion of the competition; not all participants chose to respond to the survey). We refer readers to the papers accepted at Interspeech 2021, or to contact directly the authors (see the web page [2] of the challenge for the list of the labs that submitted a system to the challenge), for getting more details.

What appears at a first glance is the significant improvements in the performance of some of the systems w.r.t. the baseline systems. These results are particularly impressive for the German tracks, where the WER of the best system (23.50%) is approximately half of the baseline one (45.21%), which was in turn achieved with a strong ASR system. It is worth noticing also the fact that the performance on German tracks was reached with a very small amount ($\approx$ 5 hours) of transcribed speech. Furthermore this year some of the systems make use of end-to-end approaches, that demonstrated effective in quite all of the tracks (best ranking systems in German open/closed track and English open track).

In summary, we notice that: *a)* all participants use long temporal contexts to represent audio in the form of TDNNs/CNNs or sequence-to-sequence models; *b)* all participants apply data augmentation techniques in the form of pitch, speed, volume perturbation and/or spectral augmentation; *c)* some participants apply either rescoring with RNNLM or word lattice combination or both; *d)* all participants, except one, use the pro-

[2] https://sites.google.com/fbk.eu/ss-is2021-nonnativechildren/

Table 5: *Results achieved by the participants in all tracks of the challenge. The main features of the submitted systems are also listed.*

English Closed Track

Track-Rank	%WER	ASR engine	Acoustic Model	Language Model	System Comb.
EC-1	25.69	Kaldi	TDNNs, CNN-TDNNs	4-grams	YES
EC-2	29.27	Kaldi	CNN-TDNN	word pronunciation	YES
EC-3	29.74	Kaldi	CNN-TDNNs	4-grams + RNNLM resc.	NO
EC-4	31.22	-	-	-	-
EC-5	31.36	-	-	-	-
EC-6	32.21	Kaldi	TDNN + CycleGan augm.	4-grams	NO
EC-7	33.18	Kaldi	-	4-grams	NO
EC-8 baseline	33.21	baseline	baseline	baseline	NO
EC-9	37.05	Kaldi,FAIRSEQ	same as GC-1	n-grams, LM resc.	NO

English Open Track

Track-Rank	%WER	ASR engine	Acoustic Model, Language Model, System comb.	Additional Data
EO-1	23.98	E2E	transformer, no LM, CTC dec(w=0.5)	native adult
EO-2	29.08	Kaldi	same as EC-2	NO
EO-3	29.58	-	-	-
EO-4	29.63	Kaldi	same as EC-3	NO
EO-5	30.61	FAIRSEQ	WAV2VECT, 4-grams	$\approx$2000h conversational
EO-6 baseline	33.21	baseline	baseline	NO
EO-7	37.05	Kaldi,FAIRSEQ	same as GC-1 (not used all train set)	NO

German Closed Track

Track-Rank	%WER	ASR engine	Acoustic Model	Language Model	System Comb.
GC-1	23.50	Kaldi,FAIRSEQ	WAV2VECT + transformer	n-grams, LM resc.	NO
GC-2	38.55	Kaldi	CNN+TDNN	4-grams, RNNLM resc.	NO
GC-3	39.98	Kaldi	DNN-BLSTM	4-grams, RNNLM resc.	NO
GC-4	40.04	Kaldi	-	4-grams + graphemic lex.	NO
GC-5	40.51	-	-	-	-
GC-6	40.63	-	-	-	-
GC-7	43.13	Kaldi	same as EC-2	same as EC-2	
GC-8 baseline	45.21	baseline	baseline	baseline	NO

German Open Track

Track-Rank	%WER	ASR engine	Acoustic Model, Language Model, System Combination	Additional Data
GO-1	23.50	Kaldi,FAIRSEQ	same as GC-1	NO
GO-2	39.98	Kaldi	same as GC-3	NO
GO-3	40.27	-	-	-
GO-4	40.69	-	-	-
GO-5	40.87	E2E	conformer, no LM, CTC dec(w=0.5)	NO
GO-6 baseline	45.21	baseline	baseline	NO

vided lexicon and phonetic transcriptions obtained with an automatic phonetic transcriber (basically G2P); one of the systems performed better on the German closed track using a graphemic lexicon; some participants added transcriptions coming from additional corpora (e.g. switchboard); *e)* some participants haven't used all the provided training data. The best performance in the open tracks challenges is slightly better (for English) or even the same (for German) of the related closed tracks. These results are in line with those achieved in the challenge of last year (see [22]) where it was noticed that the best performing systems didn't use additional (e.g. adult) speech for training. Also this year the results seems indicating that the usage of large amount of adult speech for training is again not much effective for recognizing children speech.

In the last year challenge it was noticed that the submitted systems could have been highly tuned to the specific characteristics of the data released (Italian students answering to a small set of prompts), instead the results achieved this year are more general, since the English data contains also speech uttered by children from different L1 language, i.e. exhibiting higher variability with respect to last year data.

6. Conclusions

This paper provides an overview of the Interspeech 2021 ASR for Non-native Children's Speech Challenge. The corpora released for the challenge are described and the results of the systems that were submitted to the challenge. These results indicate that substantial progress has been made in the state-of-the-art for this difficult task. We have released additional data sets in order to evaluate the robustness of the systems in diverse settings: speakers from diverse native language backgrounds other than Italian [30, 31] and two languages (English and German), characterized by different audio resources. Future directions to investigate can include the speaking proficiency scores (such as fluency, pronunciation, etc.) for young language learners [8, 9, 10, 32, 33, 20].

7. References

[1] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[2] S. Park and J. Culnan, "A comparison between native and non-native speech for automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 145, pp. 1827–1827, 03 2019.

[3] A. Rajpal, A. R. MV, C. Yarra, R. Aggarwal, and P. K. Ghosh, "Pseudo likelihood correction technique for low resource accented ASR," in *Proc. of ICASSP*, 2020, pp. 7434–7438.

[4] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," *Computer Speech and Language*, vol. 63, 2020.

[5] M. Matassoni, R. Gretter, D. Falavigna, and D. Giuliani, "Non-native children speech recognition through transfer learning," in *Proc. of ICASSP*, Calgary, Canada, 2018, pp. 6229–6233.

[6] S. P. Dubagunta, S. Hande Kabil, and M. Magimai.-Doss, "Improving children speech recognition through feature learning from raw speech signal," in *Proc. of ICASSP*, 2019, pp. 5736–5740.

[7] C. Bergmann, S. A. Sprenger, and M. S. Schmid, "The impact of language co-activation on L1 and L2 speech fluency," *Acta Psychologica*, vol. 161, pp. 25–35, 2015.

[8] Y. Gao, B. M. Lal Srivastava, and J. Salsman, "Spoken english intelligibility remediation with PocketSphinx alignment and feature extraction improves substantially over the state of the art," in *Proc. of 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2018, pp. 924–927.

[9] M. O'Brien, T. Derwing, C. Cucchiarini, D. Hardison, H. Mixdorff, R. Thomson, H. Strik, J. Levis, M. Munro, J. Foote, and G. Levis, "Directions for the future of technology in pronunciation research and teaching," *Journal of Second Language Pronunciation*, vol. 4, pp. 182–207, 01 2018.

[10] L. Chen, J. Tao, S. Ghaffarzadegan, and Y. Qian, "End-to-end neural network based automated speech scoring," in *Proc. of ICASSP*, Calgary, Canada, 2018, pp. 6234–6238.

[11] X. Xie and T. F. Jaeger, "Comparing non-native and native speech: Are L2 productions more variable?" *The Journal of the Acoustical Society of America*, vol. 147, pp. 3322–3347, 05 2020.

[12] N. Jose, B. R. Chakravarthi, S. Suryawanshi, E. Sherly, and J. P. McCrae, "A survey of current datasets for code-switching research," in *Proc. of International Conference on Advanced Computing and Communication Systems (ICACCS)*, 2020, pp. 136–141.

[13] J. Cheng, X. Chen, and A. Metallinou, "Deep neural network acoustic models for spoken assessment applications," *Speech Communication*, vol. 73, pp. 14–27, 2015.

[14] J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," in *Proc. of IEEE SLT*, 2014, pp. 294–299.

[15] M. Mulholland, M. Lopez, K. Evanini, A. Loukina, and Y. Qian, "A comparison of ASR and human errors for transcription of non-native spontaneous speech," in *Proc. of ICASSP*, 2016, pp. 5855–5859.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselỳ, "The Kaldi Speech Recognition Toolkit," in *Proc. of IEEE ASRU*, Hawaii (US), December 2011.

[17] K. Kyriakopoulos, K. M. Knill, and M. J. Gales, "Automatic Detection of Accent and Lexical Pronunciation Errors in Spontaneous Non-Native English Speech," in *Proc. Interspeech*, 2020, pp. 3052–3056.

[18] W. Wang, W. Wei, Y. Xie, M. Guo, and J. Zhang, "Improve the accuracy of non-native speech annotation with a semi-automatic approach," in *Proc. of International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2018, pp. 116–120.

[19] C. Baur, A. Caines, C. Chua, J. Gerlach, M. Qian, M. Rayner, M. Russell, H. Strik, and X. Wei, "Overview of the 2019 spoken call shared task," in *Proc. of SLATE*, Graz, Austria, 2019, pp. 1–5.

[20] L. Zhang, Z. Zhao, C. Ma, L. Shan, H. Sun, L. Jiang, S. Deng, and C. Gao, "End-to-end automatic pronunciation error detection based on improved hybrid CTC/attention architecture," *Sensors*, vol. 20, p. 1809, 03 2020.

[21] N. Bogach, E. Boitsova, S. Chernonog, A. Lamtev, M. Lesnichaya, I. Lezhenin, A. Novopashenny, R. Svechnikov, D. Tsikach, K. Vasiliev, E. Pyshkin, and J. Blake, "Speech processing for language learning: A practical approach to computer-assisted pronunciation teaching," *Electronics*, vol. 10, no. 3, 2021.

[22] R. Gretter, M. Matassoni, D. Falavigna, K. Evanini, and C. Leong, "Overview of the INTERSPEECH TLT2020 Shared Task on ASR for Non-Native Children's Speech," in *Proc. of Interspeech*, 2020.

[23] R. Gretter, M. Matassoni, K. Allgaier, S. Tchistiakova, and D. Falavigna, "Automatic assessment of spoken language proficiency of non-native children," in *Proc. of ICASSP*, 2019.

[24] R. Gretter, M. Matassoni, S. Bannò, and F. Daniele, "TLT-school: a corpus of non native children speech," in *Proceedings of the 12th Language Resources and Evaluation Conference*, May 2020, pp. 378–385.

[25] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. of Interspeech*, 2016, pp. 2751–2755.

[26] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. of Interspeech*, September 2018, pp. 3743–3747.

[27] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Proc. of Interspeech)*, 2019, pp. 2613–2617.

[28] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Nat. Lang. Eng.*, vol. 22, no. 6, pp. 907–938, 2016.

[29] V. Manohar, H. Hadian, D. Povey, and S. Khudanpur, "Semi-supervised training of acoustic models using lattice-free MMI," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4844–4848, 2018.

[30] S. Ghorbani, A. E. Bulut, and J. H. L. Hansen, "Advancing multi-accented LSTM-CTC speech recognition using a domain specific student-teacher learning paradigm," in *Proc. of IEEE SLT*, 2018, pp. 29–35.

[31] R. Ubale, V. Ramanarayanan, Y. Qian, K. Evanini, C. W. Leong, and C. M. Lee, "Native language identification from raw waveforms using deep convolutional neural networks with attentive pooling," in *Proc. of IEEE ASRU*, 2019, pp. 403–410.

[32] K. M. Knill, M. J. F. Gales, P. P. Manakul, and A. P. Caines, "Automatic grammatical error detection of non-native spoken learner English," in *Proc. of ICASSP*, 2019, pp. 8127–8131.

[33] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, "Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 391–401, 2020.