Overview of the Interspeech TLT2020 Shared Task on ASR for Non-Native Children's Speech

Roberto Gretter[1], *Marco Matassoni*[1], *Daniele Falavigna*[1], *Keelan Evanini*[2], *Chee Wee Leong*[2]

[1]Fondazione Bruno Kessler (FBK), Trento, Italy
[2]Educational Testing Service, Princeton, USA

(gretter,matasso,falavi)@fbk.eu, (kevanini,cleong)@ets.org

Abstract

We present an overview of the ASR challenge for non-native children's speech organized for a special session at Interspeech 2020. The data for the challenge was obtained in the context of a spoken language proficiency assessment administered at Italian schools for students between the ages of 9 and 16 who were studying English and German as a foreign language. The corpus distributed for the challenge was a subset of the English recordings. Participating teams competed either in a closed track, in which they could use only the training data released by the organizers of the challenge, or in an open track, in which they were allowed to use additional training data. The closed track received 9 entries and the open track received 7 entries, with the best scoring systems achieving substantial improvements over a state-of-the-art baseline system. This paper describes the corpus of non-native children's speech that was used for the challenge, analyzes the results, and discusses some points that should be considered for subsequent challenges in this domain in the future.

Index Terms: non-native children's speech, speech recognition, ASR, deep learning

1. Introduction

The availability of large amounts of training data and large computational resources have made automatic speech recognition (ASR) technology usable in many application domains, and recent research has demonstrated that ASR systems can achieve performance levels that match human transcribers for some tasks [1]. However, ASR systems still present deficiencies when applied to speech produced by specific types of speakers, in particular, non-native speakers [2, 3] and children [4, 5, 6].

Several phenomena that regularly occur in non-native speech can greatly reduce ASR performance, including mispronounced words, ungrammatical utterances, disfluencies (including false starts, partial words, and filled pauses), and code-switched words [7, 8, 9, 10]. ASR for children's speech can be challenging due to linguistic differences from adult speech at many levels (acoustic, prosodic, lexical, morphosyntactic, and pragmatic) caused by physiological differences (e.g., shorter vocal tract lengths), cognitive differences (e.g., different stages of language acquisition), and behavioral differences (e.g., whispered speech). Developing ASR systems for both of these domains is made more challenging due to the lack of publicly available databases for both non-native speech and children's speech. While recent studies have demonstrated that state-of-the-art ASR systems for spontaneous speech produced by adult native speakers of English can achieve quite low error rates, these difficulties result in substantially worse performance for the few prior studies that have specifically investigated non-native children's speech, typically in the context of automated language proficiency assessments[11, 12]. In one study, [13] report a word error rate (WER) value of 33.0% for open-ended spoken responses produced by K-12 English learners who took a standardized speaking assessment in the USA. In another study of adolescent English learners representing a range of first language backgrounds, [14] report a WER of 23.2% for responses to an academic summarization task.

Despite these difficulties, a significant portion of the speech transcribed by ASR systems in practical applications may come from both non-native speakers, (e.g., newscasts, movies, internet videos, human-machine interactions, human-human conversations in telephone call centers, etc.) and children (e.g., educational applications, smart speakers, speech-enabled gaming devices, etc.). Therefore, it is necessary to continue to improve ASR systems to be able to accurately process speech from these populations. An additional important application area is the automatic assessment of second language speaking proficiency, where the ASR difficulties can be increased by the low proficiency levels of the speakers, especially if they are children. The lack of training data is especially pronounced for this population (non-native children's speech).

To help address these gaps and stimulate research that can advance the present state-of-the-art in ASR for non-native children's speech we have freely distributed a new data set containing non-native children's speech and have organized a challenge whose results will be presented and discussed in a special session at Interspeech 2020. The data set consists of spoken responses collected in Italian schools from students between the ages of 9 and 16 in the context of English speaking proficiency assessments[1]. The data included in the release contains development and an evaluation sets (ca. 2 hours each) and an adaptation set (ca. 9 hours), all of which were carefully transcribed by human listeners. An additional set of around 40 hours of spoken responses that were transcribed using a less strict protocol was also distributed. A baseline system, based on the Kaldi toolkit [15], was released together with the data, and a challenge web site was developed for collecting and scoring submissions.

Our intention is that the release of this data set will allow researchers to establish benchmarks in the area of non native children's speech as well as to address research topics in several ASR sub-fields, including the following:

- Language models: how to handle grammatically incorrect sentences, false starts and partial words, code-switched words, etc.
- Lexicon: generation of multiple pronunciations for non-native accents, training of pronunciation models, etc.

[1]We acknowledge IPRASE (https://www.iprase.tn.it) the Italian Institution in the Trentino region that organized the evaluation campaigns, for giving the permit to distribute the data.

- Acoustic models: multilingual model training, transfer learning approaches, model adaptation for non-native children (supervised, unsupervised, lightly supervised), modeling of spontaneous speech phenomena, acoustic models for non-native children, etc.
- Evaluation: database acquisition and annotation of non-native speech [16], performance evaluation for non-native children's speech.
- Handling low resource training/adaptation data for less commonly studied populations (non-native speech, children's speech).
- Establishing a common data set for additional future annotations for applications beyond ASR (e.g., computer assisted language learning).

We selected CodaLab as the platform for organizing the challenge due to its ease of use, availability of communication tools such as mass-emailing, online forum for clarification of task issues, and tracking of submissions in real time. Submissions were anonymized on an individual basis and identified only by the team name. The statistics displayed were the lowest word error rate (WER) score of all submissions from a given team to-date along with the total number of submissions for the team since the beginning of the challenge. The metric used for evaluation and ranking of participants is the WER value, and participants were also able to see more detailed evaluation results (including # of insertions, # of deletions and # of substitutions) for each submission. The challenge included both a **closed tack** 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 to train the models. The submission window was open for a total of 7 days and teams were allowed to provide at most one submission per day to each track, with a maximum total of 7 submissions per track per team. More details about the challenge can be found at the CodaLab site.[2]

A total of 9 teams submitted results for the closed track and 7 teams submitted results for the open track. In both tracks, the best systems substantially outperformed a strong baseline system built using state-of-the art models and algorithms.

2. Audio and language resources

In Trentino, an autonomous 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. Three evaluation campaigns were foreseen, two having been completed in 2016 and 2018, and a final one scheduled in 2020; due to the Covid19 emergency, no one knows if and when it will take place. The 2018 campaign was split into two parts: the 2017 try-out data set and the actual 2018 data. Table 1 highlights some information about the pupils that took part to the campaigns. More information can be found in [17]; here we just mention that, together with multiple-choice questions, pupils were asked to produce both written and spoken material. Spoken data represent the core of the TLT2020 challenge; written data were made available for language model purposes.

The prompts for the English spoken data consist of the same 24 prompts in 2017 and 2018; 85 different prompts were used in 2016. A1 prompts correspond to simple questions, while A2 and especially B1 prompts give rise to more open-ended utterances.

[2]https://competitions.codalab.org/competitions/23672

Table 1: *L2 linguistic competences in Trentino: level, grade, age and number of pupils participating in the evaluation campaigns. Most pupils did both English and German tests.*

CEFR	Grade, School	Age	Number of pupils		
			2016	2017	2018
A1	5, primary	9-10	1074	320	517
A2	8, secondary	12-13	1521	111	614
B1	10, high school	14-15	378	124	1112
B1	11, high school	15-16	141	0	467
tot	5-11	9-16	3114	555	2710

Table 2: *English spoken data collected during different evaluation campaigns. "#Q" indicates the total number of different prompts presented to the pupils. German data are similar.*

Year	Lang	#Pupils	#Utterances	Duration	#Q
2016	ENG	2748	17462	69:03:37	85
2017	ENG	511	4112	16:25:45	24
2018	ENG	2332	15770	93:14:53	24

2.1. Spoken Data

Table 2 reports some statistics extracted from the spoken data collected so far in all the campaigns. Normally, around 20 students of the same class took the test together, so it is quite common that speech of classmates or teachers overlaps with the speech of the student speaking in her/his microphone. On average, the audio signal quality is nearly good, while the main problem is caused by a high percentage of extraneous speech. In fact, recordings have a fixed duration - depending on the question - so at the end of the response some extra speech is often captured. In addition, background noise is often present due to several sources (doors, steps, keyboard typing, background voices, street noises if the windows are open, etc). Finally, many answers are whispered and difficult to understand.

The audio recordings distributed with the challenge belong to two sets: 2017 recordings, manually transcribed by FBK (TLT2017train, TLT2017dev, TLT2017eval); and a selection of 2016 and 2018 recordings, manually transcribed by ETS (TLT1618train). Table 3 reports some more information about these datasets. Note that all the recordings uttered by a given speaker are in the same dataset, i.e. there is no overlapping of speakers among datasets. Every single utterance in the datasets is an audio file whose name contains two IDs: *speaker id* and *question id*. In this way it is possible to exploit this information to perform some fine-tuning on the data.

Table 3: *Some statistics on the speech data distributed with the challenge; number of utterances, pupils, questions, running words, total duration.*

id	#Utt	#Pup	#Q	#Words	Duration
TLT1618train	11711	3112	109	*136578*	40:29:37
TLT2017train	2299	338	24	22450	08:59:30
TLT2017dev	562	84	24	5287	02:05:18
TLT2017eval	578	84	24	6206	02:20:48

2.1.1. Manual Transcriptions of 2017 data

In order to create an ASR benchmark, most utterances in the 2017 data sets were manually transcribed at FBK. The whole process is described in more detail in [17], here we just briefly report the most important guidelines: • only the main speaker has to be transcribed; the presence of other voices (schoolmates, teacher) is reported with "@voices"; • whispered speech is explicitly marked with the label "()", • badly pronounced words have to be marked by a "#" sign; "#*" marks incomprehensible speech; • speech in a different language from the target language is marked with an explicit label *"I am 10 years old @it(io ho già risposto)"*.

2.1.2. Selection of 2016/2018 data and their manual transcriptions

To enlarge the ASR benchmark, we decided to select approximately 40 hours of speech from the 162 hours of recordings belonging to the 2016 and 2018 English data. We decided to keep utterances corresponding to every Question ID (QID) (max. 200 utterances or 30 minutes for every QID), to favour longer utterances and to discard similar phrases (by looking at the ASR output). A selection procedure, described in the documentation of the challenge, was implemented following these criteria, and resulted in about 40 hours of speech.

The initial idea was to distribute this data set untranscribed, to enable participants to explore unsupervised training approaches. Then, ETS performed a manual transcription of this data set in February, 2020, shortly before the start of the challenge. This manual transcriptions suffers some lack of knowledge about Italian language, in particular for Italian names and geographical Trentino toponyms that are sometimes transcribed incorrectly.

Table 4: *Some statistics about the text data distributed with the challenge: number of running words, lexicon size, number of different and running OOV words, OOV rate, perplexity computed with a 4-gram Maximum Entropy model.*

id	#Run Words	Lex Size	#Diff - #Run OOV	OOV rate	Perpl
TLT2016Wtrain	185777	3385	0 - 0	0.00%	7.1
TLT2017train	22450	1493	0 - 0	0.00%	8.2
TLT2017dev	5287	708	103 - 120	2.27%	49.8
TLT2017eval	6206	788	108 - 119	1.92%	52.6

2.2. Text Data

To build and evaluate language models, two sources of in-domain text data were provided: (1) manual transcriptions of the 2017 audio data divided into train, dev, eval according to the audio data; and (2) written data, extracted from the written sentences provided by the pupils in 2016. Lowercase texts are obtained after a cleaning phase which consists of: • spoken data: replace every foreign word sequence with some *unk* label; remove all the @ phenomena, truncated and incomprehensible words; • written data: keep only true English words, by looking at some English, Italian and German lexicons. Some statistics about these data are reported in Table 4.

Table 5: *Baseline WER results*

	Dev	Eval
baseline	37.92	**35.09**
+ 40h	23.79	**22.54**

2.3. Baseline system

The baseline acoustic model is based on a traditional Kaldi recipe that features a factorized TDNN [18] trained with LF-MMI [19]. The language model is based on a 4-gram maximum entropy model trained on the distributed text (about 200k words).

The adopted phoneset is derived from the standard CMU dictionary[3] adding additional units for specific acoustic phenomena appearing in the manual transcriptions (e.g., laughs, background noise) and code-switching words (*unk-it* and *unk-de*). Missing pronounciations are generated using a grapheme-to-phoneme conversion tool[4] included in the recipe.

Table 5 reports the resulting WERs obtained on *Dev* and *Eval* sets with the released baseline and the improved acoustic model exploiting the additional transcribed 40 hours. The results demonstrate the importance of the amount of training data as well as having in-domain data. In this scenario, all of the acoustic data comes from students with Italian as their first language and, despite some errors in the outsourced transcriptions, the boost in the resulting acoustic model is evident.

3. Results

The results in terms of WER, for both the closed and open tracks are provided in Table 6. The top-ranking systems greatly improved upon the baseline system provided by the organisers with the WER of 15.67% for the best system at less than half of the baseline WER of 35.09%. In addition, the best submissions also outperformed the baseline results achieved using all of the training data including the additional 40 hours of transcribed data, (i.e., the WER of 22.54% mentioned in Section 2.3). Since this result was obtained with a strong state-of-the-art system, the results obtained by several of the teams participating in the challenge are quite impressive. In the next section, we will summarize and compare the main features of the submitted systems; it is envisioned that further details about these systems will be disclosed in future publications by the various teams that participated in the challenge.

Table 6 demonstrates that the best overall performance in the challenge was achieved by a submission to the closed track; the top submission to the open track was just a duplicate of the top submission to the closed track. This indicates that the best performance was obtained by only using the in-domain data distributed for the challenge and that the addition of larger amounts of out-of-domain data did not lead to a substantial improvement. This suggests that the best performing systems are highly tuned to the specific characteristics of the data released for the challenge and may not generalize to other data sets, for example containing responses from children with different first language backgrounds or to different speaking prompts that are not included in the training data. Nevertheless, the best result of 15.67% on this data set is still quite impressive compared to previously published results on this corpus [20] and on other corpora of non-native children's speech [14, 13].

4. Discussion

4.1. Closed track

For comparison purposes Tables 6 summarizes some of the features of the submitted systems of which we are aware (this in-

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[4] https://github.com/mozilla/g2p

Table 6: *Results achieved by the participants in the challenge. The main features of the submitted systems are also listed.*

Closed track

Rank	%WER	ASR engine	acoustic model	language model	system combination
1	15.67	kaldi,HTK	graphemic TDNNs	n-grams, RNNLM	ROVER,CNC
2	17.59	kaldi	VTLN, TDNN, CNN-TDNNF	n-gram, LM rescoring	model ensemble
3	18.71	kaldi	TDNN-BLSTM	n-gram, LM rescoring	lattice MBR combination
4	18.80	kaldi	TDNN	n-grams, RNNLM rescoring	-
5	19.64	unknown	unknown	unknown	unknown
6	21.63	kaldi	CNN-TDNN	n-grams	-
7	22.24	unknown	unknown	unknown	unknown
8	26.38	kaldi	TDNN-F	n-grams	-
9	26.61	kaldi	TDNN	n-grams	-
baseline	35.09	kaldi	TDNN	n-grams, LM rescoring	-

Open track

Rank	%WER	ASR engine	acoustic-model, language-model, system combination	additional data
1	15.67	kaldi,HTK	same as closed track - rank 1	-
2	17.06	kaldi	CNN+TDNN-F, 4-gram lattice rescoring, MBR lattice combination	≈325h from corpora of children's speech
3	17.79	kaldi	very similar to closed track - rank 2	-
4	18.71	kaldi	same as closed track - rank 3	-
5	22.34	unknown	unknown	unknown
6	23.24	kaldi	CNN+TDNN	transfer learning from Zamia (≈1500h)
baseline	35.09	kaldi	same as closed track	-
7	36.73	unknown	unknown	unknown

formation 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).

In summary, we notice that: *a)* all participants use hybrid ASR systems based on kaldi (only the winning system used HTK in combination with kaldi); *b)* all participants use long temporal contexts (in the form of time delayed nodes and/or recurrent nodes) in neural networks to model acoustic observations; *c)* some participants try to exploit the additional acoustic data, applying schemes for selecting reliable transcriptions; *d)* most participants apply LM rescoring; *e)* all participants use the provided lexicon and phonetic transcriptions derived with the phonetic transcriber (G2P) provided with the kaldi toolkit (the winning system makes also use of language ID tags for foreign words in order to enrich the phonetic trees); *f)* the best performing systems use system combination (ROVER and/or CNC).

4.2. Open track

As mentioned above, Table 6 shows that only two teams used additional data for acoustic model training in their open track submissions (the other teams submitted results based on the same or very similar systems that were developed for the closed track), and that the use of additional external corpora did not enable those two teams to beat the best performing system submitted to the closed track challenge. However, it is worth noting that the system that used approximately 325 hours of additional data from corpora containing children's speech (the OGI, CMU Kids, MyST, and CU Kids' corpora) achieved the 2nd place result in the open track with a WER of 17.06% whereas the system that used a much larger amount of adult speech (1500 hours from the Zamia corpus) had a much higher WER. This seems to indicate that the use of additional external data that is a closer match to the acoustic characteristics of the children's speech in the challenge corpus (even if it wasn't a perfect match since the corpora contain speech from native speakers) is much more beneficial than mismatched adult data. A further confirmation of this result can be found in [21], where the use of a small amount of carefully selected spoken utterances was shown to be effective for the recognition of children's speech.

5. Conclusions and further directions

This paper has described a corpus that was released for an Interspeech 2020 challenge on the task of ASR for non-native children's speech and has presented the results of the systems that were submitted to the challenge. The results indicate that substantial progress has been made in the state-of-the-art for this difficult task. The corpus will be released at no cost for non-commercial research purposes outside the scope of the Interspeech 2020 challenge and it is envisioned that future research will continue to improve on the promising results that have already been obtained. As discussed above, it is possible that some of the systems developed for this challenge may have over-fitted the characteristics of this particular data set. Therefore, we have plans to release additional data sets in the future that would probe the robustness of the systems through the inclusion of responses to assessment prompts that were not seen in the training data set as well as responses from speakers from diverse native language backgrounds other than Italian [22, 23].

Future directions for additional shared tasks to investigate the performance of speech processing technology for non-native children's speech include an ASR for the Italian students who were learning German (drawn from the larger corpus that contains the English responses used in this challenge) as well as a modelling speaking proficiency scores (such as fluency, pronunciation, etc.) for young language learners [7, 8, 9, 24, 25, 26].

6. 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] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[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] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional lstm-rnn for improving automated assessment of non-native children's speech," in *Proc. of Interspeech*, 2017, pp. 1417–1421.

[15] 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.

[16] 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.

[17] R. Gretter, M. Matassoni, S. Bannò, and D. Falavigna, "TLT-school: a Corpus of Non Native Children Speech," in *Proc. of LREC*, 2020.

[18] 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.

[19] 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.

[20] 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.

[21] M. Matassoni, D. Falavigna, and D. Giuliani, "DNN adaptation for recognition of children speech through automatic utterance selection," in *Proc. of IEEE SLT*, Dec 2016, pp. 644–651.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.