AUTOMATIC ASSESSMENT OF SPOKEN LANGUAGE PROFICIENCY OF NON-NATIVE CHILDREN

Roberto Gretter[1], Marco Matassoni[1], Katharina Allgaier[1,2], Svetlana Tchistiakova[1,3,4], Daniele Falavigna[1]*

(1) Fondazione Bruno Kessler, Trento (Italy), (2) University of Heidelberg (Germany),

(3) University of Trento (Italy), (4) University of Saarland (Germany)

{gretter,matasso,falavi}@fbk.eu,katharina@allgaier-um.de,stchistiakova@gmail.com

ABSTRACT

This paper describes technology developed to automatically grade Italian students (ages 9-16) on their English and German spoken language proficiency. The students' spoken answers are first transcribed by an automatic speech recognition (ASR) system and then scored using a feedforward neural network (NN) that processes features extracted from the automatic transcriptions. In-domain acoustic models, employing deep neural networks (DNNs), are derived by adapting the parameters of an original out of domain DNN. Automatic scores are computed for low level proficiency indicators - such as: lexical richness, syntax correctness, quality of pronunciation, discourse fluency, semantic relevance to the prompt, etc - defined by human experts in language proficiency. A set of experiments was carried out on a large set of data collected during proficiency evaluation campaigns involving thousands of students, manually scored by human experts. Obtained results are presented and discussed.

Index Terms— language proficiency, non-native speech, code switching, multilingual speech recognition

1. INTRODUCTION

The problem of automatic scoring of second language (L2) learning proficiency has been largely investigated in the past in the framework of computer assisted language learning (CALL). Approaches have been proposed for two input modalities: written and spoken. In both cases, specific competencies of the human learners are processed by some suitable proficiency classifiers. The final goal is to measure L2 proficiency according to some standard scale. A well-known scale is the Common European Framework of Reference for Languages (Council of Europe, 2001). The CEFR defines 6 levels of proficiency: A1 (beginner), A2, B1, B2, C1 and C2.

This work[1] addresses automatic scoring of L2 learners, focusing on different linguistic competences, or "indicators," related both to the content (e.g. grammatical correctness, lexical richness, semantic coherence, etc) and to the speaking capabilities (e.g. pronunciation, fluency, etc). Refer to Section 2 for a description of the indicators adopted in this work. The learners are Italian students, between 9 and 16 years old, who study both English and German at school. The students took proficiency tests by answering question prompts provided in written form. Responses included typed answers and

---

spoken answers. The developed system is based on a set of features (see Section 4.2) extracted both from the input speech signal and from the automatic transcriptions of the spoken responses. The features are classified with feedforward neural networks, trained on labels provided by human raters, who manually scored the "indicators" of a set of about $29,000$ spoken utterances (about $2,550$ proficiency tests). Training and test data used in the experiments will be described in Section 2.

The task is very challenging and poses many problems which are only partially considered in the scientific literature. From the ASR perspective major difficulties are represented by: *a)* recognition of both child and non-native speech, i.e. Italian pupils speaking both English and German, *b)* presence of a large number of spontaneous speech phenomena (hesitations, false starts, fragments of words, etc.), *c)* presence of multiple languages (English, Italian and German words are frequently uttered in response to a single question), *d)* presence of a significant level of background noise due to the fact that the microphone remains opened for a fixed time interval (e.g. 20 seconds), and *e)* presence of non-collaborative speakers (students often joke, laugh, speak softly, etc.).

Relation to prior work. Scientific literature is rich in approaches for automated assessment of spoken language proficiency. Performance is directly dependent on ASR accuracy which, in turn, depends on the type of input, read or spontaneous, and on the speaker ages, adults or children (see [1] for an overview of spoken language technology for education).

Automatic assessment of reading capabilities of L2 children was widely investigated in the past at both sentence level [2] and word level [3]. More recently, the scientific community started addressing automatic assessment of more complex spoken tasks, requiring more general communication capabilities by L2 learners. The AZELLA data set [4], developed by Pearson, includes $1,500$ spoken tests, each double graded by human professionals, from a variety of tasks. The work in [5] describes a latent semantic analysis (LSA) based approach for scoring the proficiency of the AZELLA test set, while [6] describes a system designed to automatically evaluate the communication skills of young English students. Features proposed for evaluation of pronunciation are described for instance in [7].

Automatic scoring of L2 proficiency has also been investigated in recent shared tasks. One of these [8] addressed a prompt-response task, where Swiss students learning English had to answer to both written and spoken prompts. The goal is to label student spoken responses as "accept" or "reject". The winners of the shared task [9] use a deep neural network (DNN) model to accept or reject input utterances, while the work reported in [10] makes use of a support vector machine originally designed for scoring written texts.

Finally, it is worth mentioning that the recent end-to-end ap-

proach [11] (based on the usage of a bidirectional recurrent DNNs employing an attention model) performs better than the well known SpeechRater™ system [12], developed by ETS for automatically scoring non-native spontaneous speech in the context of an on-line practice test for prospective takers of the Test Of English as a Foreign Language (TOEFL)[2].

With respect to the previously mentioned works, the novelties proposed in this paper are as follows. Firstly, we introduce a unique multi-lingual DNN, for acoustic modeling (AM), trained on English, German and Italian speech from children and adults. This model copes with *multi-lingual* spoken answers, i.e. utterances where the student uses or confuses words belonging to the three languages at hand. A common phonetic lexicon, defined in terms of units of the International Phonetic Alphabet (IPA), is adopted for transcribing the words of all languages. Moreover, spontaneous "in-domain" speech data (details in Section 3) are included in the training material to model frequently occurring speech phenomena (e.g., laughs, hesitations, background noise). We also propose a novel method to compute acoustic features using the phonetic representation and likelihoods output by the ASR system. We employ both our best non-native ASR system, and ASR systems trained only on native English/German data to generate these features (see subsection 4.2).

Experimental results reported in the paper (see Section 5) show that: *a)* the usage of the multilingual DNN is effective for transcribing non-native children's speech, *b)* the usage of feedforward NNs allows us to classify each indicator with an average classification accuracy between around 60% and around 67%, and *c)* no large differences in classification performance have been observed among the different indicators (i.e. the set of adopted features performs pretty well for all indicators).

2. DESCRIPTION OF THE DATA

2.1. Evaluation campaigns on trilinguism

In Trentino (Northern Italy), a series of campaigns is underway for testing linguistic competencies of multilingual Italian students taking proficiency tests in English and German. A set of three evaluation campaigns were planned, taking place in 2016, 2017/2018, and 2020. Each one involves about 3000 students (ages 9-16), belonging to 4 different school grade levels, and three proficiency levels (A1, A2, B1). The 2017/2018 campaign was split into a group of 500 students in 2017, and 2500 students in 2018. Table 1 highlights some information about the 2018 campaign. Several tests aimed at assessing the language learning capabilities of the students were carried out by means of multiple-choice questions, which can be evaluated automatically. However, a detailed linguistic evaluation cannot be performed without allowing the students to express themselves in both written sentences and spoken utterances, which typically require the intervention of human experts to be scored. In this paper we will focus only on the spoken part of these proficiency tests.

Table 2 reports some statistics extracted from the spoken data collected in 2017/2018. We manually transcribed a part of the 2017 spoken data to train and evaluate the ASR system. 2018 spoken data were used to train and evaluate the grading system. Each spoken utterance received a total score from human experts, computed by summing up the scores related to the following 6 individual indicators: **answer relevance** (with respect to the question); **syntactical correctness** (formal competences, morpho-syntactical correctness); **lexical properties** (lexical richness and correctness); **pronunciation**; **fluency**; **communicative skills** (communicative efficacy and

---

Table 1. Evaluation of L2 linguistic competences in Trentino in 2018: level, grade, age and number of pupils participating in the English (ENG) and German (GER) tests.

CEFR	Grade, School	Age	#Pupils	#ENG	#GER
A1	5, primary	9-10	508	476	472
A2	8, secondary	12-13	593	522	547
B1	10, high school	14-15	1086	1023	984
B1	11, high school	15-16	364	310	54
	tot	9-16	2551	2331	2057

argumentative abilities).

Since every utterance was scored by only one expert, it was not possible to evaluate any kind of agreement among experts. However, according to [13] and [14], inter-rater human correlation varies between around 0.6 and 0.9, depending on the type of proficiency test. In this work, correlation between an automatic rater and an expert one is between 0.53 and 0.61, indicating a good performance of the proposed system. For future evaluations more experts are expected to provide independent scoring on the same data sets, so a more precise evaluation will be possible. At present it is not possible to publicly distribute the data.

Table 2. Spoken data collected during the 2017 and 2018 evaluation campaigns. Column "#Q" indicates the total number of different (written) questions presented to the pupils.

Year	Lang	#Pupils	#Utterances	Duration	#Q
2017	ENG	511	4112	16:25:45	24
2017	GER	478	3739	15:33:06	23
2018	ENG	2331	15770	93:14:53	24
2018	GER	2057	13658	95:54:56	23

2.2. Manual transcription of spoken data

In order to create adaptation and evaluation sets for ASR, we manually transcribed part of the 2017 data. Guidelines for the manual annotation required a trade-off between transcription accuracy and speed. We defined guidelines, where: *a)* only the main speaker has to be transcribed; presence of other voices (school-mates, teacher) should be reported only with the label "@voices", *b)* presence of whispered speech was found to be significant, so it should be explicitly marked with the label "()", *c)* badly pronounced words have to be marked by a "#" sign (without trying to phonetically transcribe the pronounced sounds), and *d)* code switched words (i.e. speech in a different language from the target language) has to be reported by means of an explicit marker, like in: *"I am 10 years old @it(io ho già risposto)"*.

Most of 2017 data was manually transcribed by students from two Italian linguistic high schools ("Curie" and "Scholl") and double-checked by researchers. Part of the data were independently transcribed by pairs of students in order to compute inter-annotator agreement, which is shown in Table 3 in terms of Word Accuracy (WA), using the first transcription as a reference (after removing hesitations and other labels related to background voices and noises, etc.). The low level of agreement reflects the difficulty of the task, although it should be noted that the transcribers themselves were non-native speakers of English/German. ASR results will also be affected by this uncertainty.

Table 3. Inter-annotator agreement between pairs of students, in terms of Word Accuracy. Students transcribed English utterances first and German ones later.

High school	Language	#Transcribed words	#Different words	Agreement (WA)
Curie	English	965	237	75.44%
Curie	German	822	139	83.09%
Scholl	English	1370	302	77.96%
Scholl	German	1290	226	82.48%

For both ASR and grading NN experiments, data from the student populations (2017/2018) were divided by speaker identity into training and evaluation sets, with proportions of $\frac{2}{3}$ and $\frac{1}{3}$, respectively (students across the training and evaluation sets do not overlap). Table 4 reports data about the spoken data set. The id *All* identifies the whole data set, while *Clean* defines the subset in which sentences containing background voices, incomprehensible speech and fragment words were excluded.

Table 4. Statistics about the spoken data sets (2017) used for ASR.

id	# of utt.	duration		tokens	
		total	avg	total	avg
Ger Train All	1448	04:47:45	11.92	9878	6.82
Ger Train Clean	589	01:37:59	9.98	2317	3.93
Eng Train All	2301	09:03:30	14.17	26090	11.34
Eng Train Clean	916	02:45:42	10.85	6249	6.82
Ger Test All	671	02:19:10	12.44	5334	7.95
Ger Test Clean	260	00:43:25	10.02	1163	4.47
Eng Test All	1142	04:29:43	14.17	13244	11.60
Eng Test Clean	423	01:17:02	10.93	3404	8.05

3. ASR SYSTEM

3.1. Acoustic model

The recognition of non-native speech, especially in the framework of multilingual speech recognition, is a well-investigated problem. Past research has tried to model the pronunciation errors of non-native speakers [15] both by using non-native pronunciation lexicons [16, 17, 18] or by adapting acoustic models with either native data and non native data [19, 20, 21, 22].

For the recognition of non-native speech, we demonstrated in [23] the effectiveness of adapting a multilingual deep neural network (DNN) trained on recordings of native speakers to children between 9 and 14 years old.

In this work we have adopted a more advanced neural network architecture for the multilingual acoustic model, using a time-delay neural network (TDNN) and the popular lattice-free maximum mutual information training (LF-MMI) [24]. The corresponding recipe features i-vector computation, data augmentation via speed perturbation, data clean up and the mentioned MMI training.

Acoustic model training is performed on the following datasets:

- **GerTrainAll** and **EngTrainAll** in-domain sets;
- **Child**, collected in the past in our labs, formed by speech of Italian children speaking: Italian (*ChildIt* subset [25]), English (*ChildEn* [26]) and German (*ChildDe*). The children were instructed

to read words or sentences in Italian, English or German, respectively; it contains 28,128 utterances in total, from 249 child speakers between 6 and 13 years old, comprising 44.5 hours of speech.

- **ISLE**, the Interactive Spoken Language Education corpus [27], consisting of 7,714 read utterances from 23 Italian and 23 German adult learners of English, comprising 9.5 hours of speech.

3.2. Language models for ASR

To train effective LMs in this particular domain, we needed sentences capable of representing the simple language spoken by the learners. For each language, we created three sets of text data. The first included simple texts, collected by grabbing data from Internet pages containing foreign language courses or sample texts (about 113K words for English, 12K words for German). The second included training data from the written responses on the written portion of the proficiency tests, acquired during the 2016, 2017 and 2018 evaluation campaigns (see Table 2) (about 393K words for English, 247K words for German). This data underwent a cleaning phase, in which we corrected the most common errors (i.e. *ai em → I am*, *becouse → because*, *seher → sehr*, *brüder → bruder*...) and removed unknown words. The third included the manual transcriptions of the 2017 spoken data set (see Section 2.2) (26K words for English, 10K words for German). In this case, we cleaned the data by deleting all markers indicating presence of extraneous phenomena, except for hesitations, which were retained.

This small amount of data (about 532K words for English and 269K words for German in total) was used to train two 3-gram LMs.

3.3. ASR performance

Table 5 reports word error rates (WERs) obtained on the 2017 test set (see Table 4) using acoustic models trained on both out-of-domain data and in-domain data, which contributes to modeling spontaneous speech and spurious speech phenomena (laughing, coughing, ...).

A common phonetic lexicon, defined in terms of the units of the International Phonetic Alphabet (IPA), is shared across the three languages (Italian, German and English) to allow the usage of a single acoustic model. The Table also reports results achieved on a clean subset of the test corpus, which was created by removing sentences with unreliable transcriptions, and spurious acoustic events.

Table 5. WER results on 2017 spoken test sets.

GerTestAll	GerTestClean	EngTestAll	EngTestClean
42.6	37.5	35.9	32.6

4. PROFICIENCY ESTIMATION

4.1. Scoring system

The classification task consists of predicting the scores assigned by human experts to the spoken answers. For each utterance, 6 scores are given according to the proficiency indicators (described in section 2). The scores are in the set $\{0, 1, 2\}$ where 0 indicates not-passed, 1 almost-passed and 2 passed.

For estimating each individual score, we employed feed forward NNs, using the corresponding scores assigned to each sentence by the experts as targets. In all cases the score provided by the system corresponds to the index of the output node with the maximum value. All NNs are trained using the features described below, and are characterized by three layers of dimension equal to the feature size; they

use $ReLU$ as activation function, and SGD and $AdaGrad$ for the optimizer. The learning rate is set to 0.05.

Questions associated with the answers are hierarchically clustered according to: *a)* the language, *b)* the proficiency level (i.e. A1, A2 and B1) and *c)* two sessions, S_1 containing common questions of proficiency tests (e.g. *how old are you?*, *where do you come from?*, etc) and S_2 containing specific questions (e.g. *describe your visit to Venice*, etc), and *d)* a question identifier. NNs used for classification follow this subdivision.

4.2. Classification features

The features used to score the indicators are derived both from the automatic transcriptions of the spoken answers of the students and from the speech signal.

To extract features from the transcriptions we use a set of LMs trained over different types of text data, namely, out-of-domain general texts (for English, transcriptions of TED talks [28] - around 3 million words; for German, news - around 1 million words), and in-domain texts containing the "best" written/spoken data, collected during the previously mentioned evaluation campaigns carried out in the years 2016, 2017 and 2018. Best data are selected among those that, in the training part, got the highest scores by the experts. To compute feature vectors we use 20 language models, obtained by computing *four* n-gram LMs (estimated by varying the the size of n-grams, from 1 to 4) over *five* different training data sets of decreasing size, as follows: *a)* out-of-domain text data, *b)* all in-domain data, *c)* in-domain data that share the same CEFR proficiency level, *d)* in-domain data that share the same session identifier, and *e)* in-domain data that share the same question identifier. In this way, we assume we know, for each test sentence to score, the language, the intended proficiency level, and the session/question identifiers.

For each test sentence formed by N_W words, of which N_{OOV} are out-of-vocabulary (OOV), we compute the following 5 features using each LM (taking inspiration for features from the works of [29, 30, 12, 9, 10]):

a) $\frac{\log(P)}{N_W}$, that is, the average log-probability of the sentence,

b) $\frac{\log(P_{OOV})}{N_{OOV}}$, that is, the average contribution of OOV words to the log-probability of the sentence,

c) $\frac{\log(P) - log(P_{OOV})}{N_W}$, that is, the average log-difference between the two above probabilities,

d) $N_W - N_{bo}$, where N_{bo} is the number of back-offs applied by the LM to the input sentence (this difference is related to the frequency of n-grams in the sentence that have also been observed in the training set),

e) N_{OOV}, the number of OOVs in the sentence.

Note that if word counts N_W or N_{OOV} are equal to zero (i.e. both P and P_{OOV} are not defined), the corresponding average log-probabilities are replaced by -1.

In this way we compute $5 \times 4 \times 5 = 100$ features for each input sentence (five data sets × four n-gram levels × five features). To this set we add 11 more transcription-based features, i.e.: the total number N_W of words in the sentence, the number of content words, the number of OOVs and the percentage of OOVs wrt a reference lexicon, the numbers of words used in Italian, in English and in German, the number of words that had to be corrected by our in-house spelling corrector adapted to this task, the number BoW (Bag of Words) of content words that match the most frequent ones of the "best" answers in the training data, BoW divided by all words in the sentence and BoW divided by all the content words in the sentence. This results in a vector of 111 features.

Finally, we use the acoustic model outputs to generate 5 additional pronunciation-based features. For this we used our best non-native model, and two additional native-language acoustic models (one trained on adult English speech from TED talks, and one trained on adult German speech from the BAS corpus). The acoustic model outputs include an alignment of the acoustic frames to phone states and a likelihood of being in that phone state given the acoustic features. We removed frames aligned to silence/background noise, and then generated the following features: *a)* the length of the utterance in number of acoustic frames, *b)* the number of silence frames in the utterance output by our best ASR system, *c)* a confidence score based on the sum of the likelihoods of each context independent phone class, similar to the work by [31], but averaged over all states in the utterance for that phone class and normalized over the unique phones in the utterance, *d)* the edit distance between the phonetic outputs from the native ASR system and the non-native ASR, similar to [32], *e)* the difference between the confidence scores from the native and non-native ASR system, similar to [33]. Therefore, in total, when making use of all features, we represent the student's answer with a vector of dimensionality 116.

5. CLASSIFICATION RESULTS AND CONCLUSIONS

For measuring the performance, we consider three metrics: Correct Classification (*CC*), linear Weighted Kappa (*WK*), Correlation (*Corr*) between the expert's scores and the predicted ones. For all the three metrics, the value of 1.0 corresponds to perfect classification; completely wrong classification is 0.0 for CC and WK, -1.0 for Corr. For each indicator, we used the training data to train a feedforward NN by grouping sentences sharing language, proficiency level and session. Average classification results are reported in Table 6.

Table 6. Average classification results on 2018 data, spoken, reported both on training and test data, given in terms of Correct Classification (*CC*), Weighted Kappa (*WK*), Correlation (*Corr*).

Dataset	English			German		
	CC	WK	Corr	CC	WK	Corr
2018 train	0.712	0.840	0.684	0.763	0.866	0.763
2018 test	0.596	0.775	0.532	0.667	0.822	0.613

Looking at the results in Table 6, the performance in terms of all reported metrics (CC, WK and $Corr$) is good, showing that the automatically assigned scores are not far from the manual ones assigned by human experts. The low difference between the performance on training and corresponding test sets indicate that the models do not overfit the data. More importantly, the values of the achieved correlation coefficients resemble those reported in [13], related to human rater correlation, on a conversational task which is, in terms of difficulty for L2 learners, similar to some of the tasks analyzed in this paper.

Future work will address the usage of both features selection and regression models for proficiency classification. Also the investigation of an extended set of features, partially inspired by the hybrid set of features proposed in [28] for ASR quality estimation will be carried out.

6. REFERENCES

[1] M. Eskenazi, "An overview of spoken language technology for education. speech communication," *Speech Communication*, vol. 51, no. 10, pp. 2862–2873, 2009.

[2] K. Zechner, J. Sabatini, and L. Chen, "Automatic scoring of childrens read-aloud text passages and word lists," in *Proc. of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, 2009.

[3] J. Tepperman, M. Black, P. Price, S. Lee, A. Kazemzadeh, M. Gerosa, M. Heritage, A. Alwan, and S. Narayanan, "A bayesian network classifier for word-level reading assessment," in *Proc. of ICSLP*, 2007.

[4] J. Cheng, Y. Zhao-DAntilio, X. Chen, , and J. Bernstein, "Automatic spoken assessment of young english language learners," in *Proc. of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014.

[5] M. Angeliki and Cheng J., "Using Deep Neural Networks to improve proficiency assessment for children English language learners," in *Proc. of Interspeech*, 2014, pp. 1468–1472.

[6] K. Evanini and X. Wang, "Automated speech scoring for nonnative middle school students with multiple task types," in *Proc. of Interspeech*, 2013, pp. 2435–2439.

[7] H. Kibishi, K. Hirabayashi, and S. Nakagawa, "A statistical method of evaluating the pronunciation proficiency/intelligibility of english presentations by japanese speakers," *ReCALL*, vol. 27, no. 1, pp. 58–83, 2015.

[8] C. Baur, J. Gerlach, M. Rayner, M. Russell, and H. Strik, "A Shared Task for Spoken CALL?," in *Proc. of SlaTe*, Stockholm, Sweden, 2017, pp. 237–244.

[9] Y. R. Oh, H-B. Jeon, H. J. Song, B. O. Kang, Y-K. Lee, J.G. Park, and Y-K. Lee, "Deep-learning based Automatic Spontaneous Speech Assessment in a Data-Driven Approach for the 2017 SLaTE CALL Shared Challenge," in *Proc. of SlaTE*, Stockholm, Sweden, 2017, pp. 103–108.

[10] K. Evanini, M. Mulholland, E. Tsuprun, and Y. Qian, "Using an Automated Content Scoring System for Spoken CALL Responses: The ETS submission for the Spoken CALL Challenge," in *Proc. of SlaTe*, Stockholm, Sweden, 2017, pp. 97–102.

[11] 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.

[12] K. Zechner, D. Higgins, X. Xi, and D. Williamson, "Automatic scoring of nonnative spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[13] V. Ramanarayanan, P Lange, K. Evanini, H. Molloy, and D. Suendermann-Oeft, "Human and automated scoring of fluency, pronunciation and intonation during HumanMachine Spoken Dialog Interactions," in *Proc. of Interspeech*, Stockholm, Sweden, 2017, pp. 1711–1715.

[14] M. Nicolao, A. V. Beeston, and T. Hain, "Automatic assessment of English learner pronunciation using discriminative classifiers," in *Proc. of ICASSP*, 2015, pp. 5351–5355.

[15] G. Bouselmi, D. Fohr, I. Illina, and J. P. Haton, "Multilingual non-native speech recognition using phonetic confusion-based acoustic model modification and graphemic constraints," in *Proc. of ICSLP*, 2006, pp. 109–112.

[16] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Proc. of ICASSP*, 2003, pp. 540–543.

[17] Y. R. Oh, J. S. Yoon, and H. K. Kim, "Adaptation based on pronunciation variability analysis for non native speech recognition," in *Proc. of ICASSP*, 2006, pp. 137–140.

[18] H. Strik, K.P. Truong, F. de Wet, and C. Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[19] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Articulatory modeling for pronunciation error detection without non-native training data based on dnn transfer learning," *IEICE Transactions on Information and Systems*, vol. E100.D, no. 9, pp. 2174–2182, 2017.

[20] W. Li, S. M. Siniscalchi, N. F. Chen, and C. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *Proc. of ICASSP*, 2016, pp. 6135–6139.

[21] A. Lee and J. Glass, "Mispronunciation detection without non-native training data," in *Proc. of Interspeech*, 2015, pp. 643–647.

[22] A. Das and M. Hasegawa-Johnson, "Cross-lingual transfer learning during supervised training in low resource scenarios," *Proc. of Interspeech*, pp. 3531–3535, 2015.

[23] 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.

[24] 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.

[25] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, vol. 49, no. 10, pp. 847 – 860, 2007.

[26] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong, "The PF-STAR children's speech corpus," in *Proc. of Eurospeech*, 2005, pp. 2761–2764.

[27] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proceedings of LREC*, 2000, pp. 957–964.

[28] S. Jalalvand, M. Negri, D. Falavigna, M. Matassoni, and M. Turchi, "Automatic quality estimation for ASR system combination," *Computer Speech and Language*, vol. 47, pp. 214–239, 2017.

[29] K. Sakaguchi, M. Heilman, and N. Madnani, "Effective feature integration for automated short answer scoring," in *Proc. of NAACL*, Denver (CO), USA, 2015, pp. 1049–1054.

[30] S. Srihari, R. Srihari, P. Babu, and H. Srinivasan, "On the automatic scoring of handwritten essays," in *Proc. of IJCAI*, Hyderabad, India, 2007, pp. 2880–2884.

[31] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Proc. of ICASSP*, 1997, vol. 2, pp. 1471–1474.

[32] K. Zechner and I. Bejar, "Towards automatic scoring of non-native spontaneous speech," in *Proc. of HLT-NAACL*, 2006, pp. 216–223.

[33] F. Landini, *A Pronunciation Scoring System for Second Language Learners*, Ph.D. thesis, Universidad de Buenos Aires, 2017.