Automatic Pronunciation Assessment of Non-native English based on Phonological Analysis

C. D. Rios-Urrego* D. Escobar-Grisales* S. A. Moreno-Acevedo*
P. A. Perez-Toro*† E. Nöth† J. R. Orozco-Arroyave*†
Corresponding author: cdavid.rios@udea.edu.co

* Faculty of Engineering, University of Antioquia UdeA, Medellín, Colombia
† Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg

Abstract. The rapid development of speech recognition systems has motivated the community to work on accent classification, considerably improving the performance of these systems. However, only a few works or tools have focused on evaluating and analyzing in depth not only the accent but also the pronunciation level of a person when learning a non-native language. Our study aims to evaluate the pronunciation skills of non-native English speakers whose first language is Arabic, Chinese, Spanish, or French. We considered training a system to compute posterior probabilities of phonological classes from English native speakers and then evaluating whether it is possible to discriminate between native English speakers vs. non-native English speakers. Posteriors of each phonological class separately and also their combination are considered. Phonemes with low posterior results are used to give feedback to the speaker regarding which phonemes should be improved. The results suggest that it is possible to distinguish between each of the non-native languages and native English with accuracies between 67.6% and 80.6%. According to our observations, the most discriminant phonological classes are alveolar, lateral, velar, and front. Finally, the paper introduces a graphical way to interpret the results phoneme-by-phoneme, such that the speaker receives feedback about his/her pronunciation performance.

Keywords: Pronunciation assessment, Speech, English, Phonological Analysis

1 Introduction

English is the official language in over 50 countries and is widely used as a second language in many others. It is considered the language of international communication in business, academia, politics, and others [4]. Thus, there is a broad interest in learning this second language for speakers with a different native language. Typically, the English level is evaluated by a human, which is not always accurate due to subjective biases; for instance, evaluators may have different expectations and standards, leading to inconsistent and unreliable assessments [1]. Computer-based assessments can give a more objective and effective assessment of the English level by analyzing specific aspects of speech to provide feedback to users, which helps them identify their strengths and weaknesses. There are multiple tools for automatic assessment of English level, where grammatical skills, vocabulary knowledge, and others are evaluated using Automatic Speech Recognition (ASR) systems based on metrics such as word accuracy

rate [8]. However, few tools evaluate or analyze deeply aspects of the English level, such as fluency, naturalness, or phonological precision, where it is possible to identify specific phonemes that can be more difficult to pronounce according to the native language in order to emphasize them in the learning process.

Automatic accent classification in speech recognition plays an important role in adapting systems to linguistic variations, improving recognition accuracy and robustness to different regions and contexts [21]. Therefore, many works have been addressed in the scientific community to classify accents. For example, in [13], a system based on Convolutional Neural Networks (CNNs) was trained and evaluated for classifying nine accents; the authors achieved an accuracy of up to 98.6%. Similar work was performed in [7], where the classification problem consisted of determining whether English speech samples are spoken by native speakers of English, Japanese, Dutch, French, or Polish. Again, this work using CNNs reported accuracies of up to 90% for discriminating the five accents. In [2], the authors used classical techniques and CNNs to recognize five accents (English, Arabic, French, German, and Hindi). They showed that the classical methods are not sufficiently efficient to solve this problem, and they obtained the best results with a deep learning approach with a mean accuracy of 90.2%. For the same corpus, in [17], five accents were evaluated (Arabic, English, French, Mandarin, and Spanish) using classical and deep approaches; in this case, the Mel-Frequency Cepstral Coefficients (MFCCs) obtained the best performance with an accuracy of 71.4%.

However, only some studies have investigated the level of pronunciation of each participant in addition to accent classification. A first approach to this can be found in [5], where the authors propose a model based on random forests and MFCCs to detect and correct automatic pronunciation errors in English classes. This work performed a biclass classification (correct pronunciation vs. mispronunciations), obtaining accuracies of up to 74.7%. In [12], the authors propose an automatic pronunciation evaluation for non-native speakers based on robust models such as Wav2Vec 2.0 and HuBERT + bidirectional long short-term memory with the layer-wise contextual representations and the corresponding text. The authors achieved correlations of up to 0.82 when comparing model performance against human-labeled annotations. Following the same line of automatically evaluating the accent, in [16], a bidirectional long short-term memory layer in a neural network was proposed to predict human ratings of the accentedness of recorded speech. When the model prediction was compared with the human ratings, correlations of up to 0.57 were reported. Finally, in [10], a work that identifies pronunciation errors in non-native speech using spectrogram and MFCCs was presented. The authors evaluated each modality's performance and included their fusion for classifying some phonological classes, in addition to the error per phoneme. They observed that the fusion of both modalities achieved the best performance, and the erroneous phonemes found automatically are similar to those labeled manually.

Motivated by this, our study seeks to provide insights into the challenges faced by non-native English speakers in mastering English pronunciation and improving language learning and teaching strategies. Initially, we trained and evaluated Phonet[1], which computes the posterior probabilities of phonological classes from speech sig-

[1] `https://phonet.readthedocs.io/en/latest/?badge=latest`

nals. Moreover, it considers several phoneme groups according to the place and manner of articulation.

Thus, we obtained the posterior probabilities for each audio from the target database to perform a classification between native English speakers vs. non-native English speakers for each phonological class and considered the fusion of these phonological classes. Finally, in each non-native English language, we obtained the most discriminative phonological class; then, we assessed weak phonemes in pronunciation to give feedback to each participant on which phonemes they had difficulty pronouncing compared to native speakers as a strategy to improve their pronunciation performance.

The rest of the paper is as follows: section 2 describes the corpora considered for this study. Section 3, presents the methods used in the study. Section 4 shows the results and analysis of the study; and finally, section 5 contains the conclusions and future work.

2 Data

2.1 TIMIT Corpus

In this work, the architecture used was trained and evaluated with the TIMIT database, which consists of 2342 sentences read by 630 speakers with different dialects of American English [6]. This corpus was developed mainly to train and evaluate automatic speech recognition systems. The TIMIT corpus includes time-aligned orthographic, phonetic, and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. In addition, the TIMIT corpus transcriptions have been hand-verified. Test and training subsets, balanced for phonetic and dialectal coverage, are specified.

2.2 Speech Accent Archive

We used the Speech Accent Archive as the target corpus [20]. This dataset contains 2140 speech samples, each from a different talker reading the same reading passage in English (69-word paragraph). Talkers come from 177 countries and have 214 different native languages. Due to the large imbalance that exists in the database (English: 27%, Spanish: 7.5%, Arabic: 4.7%, etc). We only considered the native speakers of the corpus (English), and the first 4 groups of non-native English speakers with the largest number of participants: Spanish, Arabic, Mandarin, and French. In addition, due to the idea of assessing the pronunciation level of each non-native speakers vs. native speakers, we chose a subset of English to assess each set of non-native speakers that will guarantee age and gender balance from the t-test and Chi-squared test, respectively. Therefore, each language was paired with the same number of English participants as follows: Spanish (162 participants), Arabic (102 participants), Mandarin (65 participants), and French (63 participants).

3 Methods

Figure 1 summarizes the architecture proposed in this work. Initially, we prepared the TIMIT corpus audios with their respective transcriptions to train Phonet. Then, we take

the recordings of native and non-native speakers of English from the Speech Accent Archive and compute the phonological posteriors associated with each phonological class. Finally, we performed 2 approaches: (i) we classified between each set of non-native speakers, i.e., Spanish, Arabic, Mandarin, and French vs. their corresponding group of native speakers (English); this classification was performed using a Support Vector Machine (SVM), for each phonological class and considering the fusion of all of them. (ii) After finding the most discriminative phonological class for each set of non-native speakers, we performed a phoneme-level analysis to give feedback per phoneme on the pronunciation level of a specific speaker compared to a native speaker. Details of each stage are presented below.

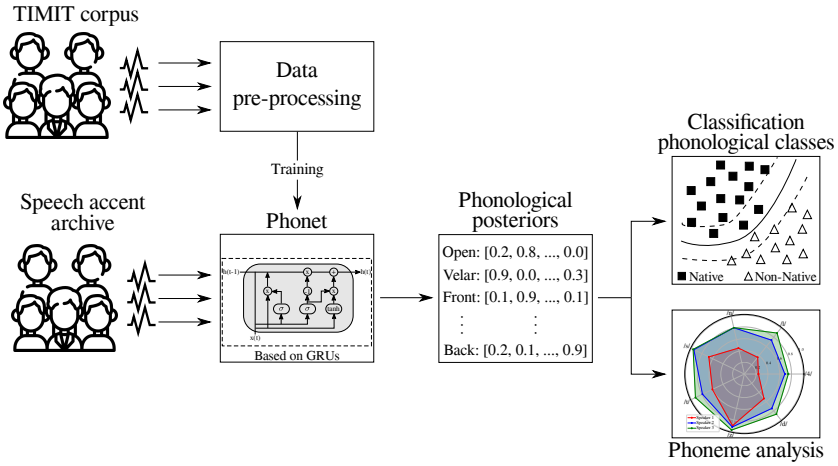

Fig. 1. Architecture proposed in this work.

3.1 Phonological analysis

Phonological features are used to model the information about the place and manner of articulation of a speaker. These features are more understandable for clinicians than the standard high-dimensional features used in speech processing. Therefore, these features are typically used to model pathological speech, such as dysarthria, apraxia, and others [11, 3]. Models of phonological analysis aim to detect the phonological class of a speech frame, where a phonological class is composed of a set of phonemes that share certain features, such as voicing, place of articulation, or manner of articulation. For instance, the phonological class “Alveolar” is a phonological class that groups the phonemes that are articulated with the tongue tip touching the alveolar ridge, which is the bony ridge behind the upper teeth. In this study, we used a toolkit called Phonet to estimate the phoneme articulation precision of different speakers and used these posteriors to classify native and non-native English speakers.

3.2 Phonet

This toolkit was proposed in [18], and it is designed to estimate phonological posteriors using bidirectional Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs). A speech segment of 400ms is defined as sequence size, and each element in the sequence is a frame of 25ms with a time-shift of 10ms. The model's input corresponds to the log-energy of the speech frame distributed into 33 triangular filters separated according to the Mel scale. This input is used to feed two bidirectional GRU layers with 128 cells. The output of the second bidirectional GRU is processed using N_c time-distributed dense layers, where N_c is the number of phonological classes. The model was trained following a multitask learning strategy to detect different phonological classes, and a Softmax activation function was used to get posterior probabilities. In [18], the model was trained with Spanish language utterances using the CIEMPIESS corpus to predict 21 phonemes distributed into 18 phonological classes. In this study, we trained the same model to predict phoneme articulation precision in English; therefore, we used the TIMIT corpus and considered 22 phonological classes: diphthong, back, closed, rounded, vowel, voiceless, postalveolar, open, velar, nasal, alveolar, bilabial, front, glottal, voiced, fricative, approximant, labiodental, dental, plosive, trill, and lateral. The notation of the phonemes is based on the International Phonetic Alphabet (IPA).

3.3 Classification and Analysis Stage

For the classification stage, we obtained a static representation for each phonological class, for which we calculated six different functionals: mean, standard deviation, skewness, kurtosis, maximum, and minimum. For this experiment, we considered classifying each set of non-native speakers vs. its corresponding group of native speakers using an SVM. This method allows discriminating N samples by finding a separating hyperplane that maximizes the margin between classes. We used a radial basis function as the kernel for the SVM, and its parameters were optimized upon a grid-search. The complexity parameter was varied as $C \in \{0.001, 0.005, 0.01, \cdots, 100, 500, 1000\}$ and the bandwidth of the kernel was varied as $\gamma \in \{0.0001, 0.001, \cdots, 1000\}$. We train, optimize and evaluate each phonological class individually and consider the fusion of all phonological classes forming a final vector of 132 features per participant (22 phonological classes $\times$ 6 statistics). All experiments are performed following a 5-fold cross-validation strategy. The results are reported in terms of mean and standard deviation computed along the folds. In the analysis stage, we consider it important to give feedback to the user on which phonemes are the most difficult to recognize in the system. For this, we consider a radar figure where we show for the most discriminative phonological class every mean posterior of each phoneme and compare it with the same phonemes of a native speaker (considered their target).

4 Experiments and Results

4.1 Training Phonet

Twenty-two phonological classes were trained and classified during the development of this work. In addition, it was guaranteed that each extracted phoneme had at least one or more phonological classes. The results show that the system's mean accuracy is 92.46% with a deviation of 3.16%. The lowest-performing phonological class is "Back" with an accuracy of 86.9%. In addition, the model for phoneme recognition proposed in [18] was trained with the TIMIT corpus in order to obtain a model that can recognize 51 phonemes of the English language. The system manages to predict the 51 phonemes with an accuracy of 67.7%.

4.2 Classification of Phonological Classes

The purpose of our study is to evaluate the pronunciation skills of non-native English speakers from Arabic, Chinese, Spanish, and French backgrounds. To achieve this goal, we apply a phonological approach to measure the accuracy of their pronunciation using Phonet to differentiate between native and non-native English speakers. We measure the confidence level of the classification to determine the degree of proficiency in English pronunciation. A higher score indicates a higher level of accuracy in differentiating between native and non-native English speakers. For instance, a high confidence score suggests that the speaker struggles with proper pronunciation.

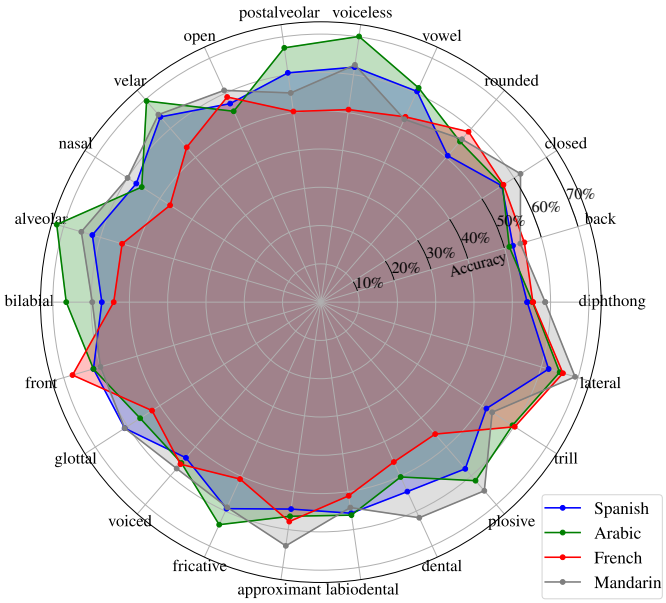

Fig. 2. Accuracy (%) of native vs non-native English speakers for all phonological classes.

The radar chart presented in Figure 2 demonstrates the performance of the classification between native vs. non-native English speakers for all phonological classes. We could see that French speakers (red color), in general, have superior pronunciation skills compared to the other language groups, as their coverage area on the chart is relatively small.

Additionally, the performance of French speakers in the nasal class is relatively poor compared to the other languages, which shows that it is easier to identify a non-native speaker of Spanish, Arabic or Chinese than a French speaker. This result can be attributed to the presence of nasal phonemes (primarily vocal) that are specific to English and French and not present in the other languages [9, 15, 14, 19]; thenceforth, French and English speakers pronounce the nasal class better than the Spanish, Arabic or Chinese speakers.

The findings of Figure 2 led to conduct a detailed analysis of the most distinguishing phonological class for each language tested. For this analysis, we selected the class with the highest score for each language as the most discriminatory class. Specifically, we identified the Alveolar class as the most distinguishing class for Arabic, the Lateral class for Chinese, the Velar class for Spanish, and the Front class for French.

Table 1. Native vs non-native English speakers for all classes and the most discriminant class.

Native Language	**Phon. Classes**	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-score (%)**
Arabic	All	80.6 ± 2.5	85.9 ± 2.7	75.3 ± 2.9	80.5 ± 2.5
	Alveolar	71.9 ± 3.1	68.4 ± 1.0	75.3 ± 5.7	71.8 ± 3.1
Mandarin	All	74.2 ± 2.8	78.5 ± 4.5	69.9 ± 1.9	74.1 ± 2.8
	Lateral	69.2 ± 3.2	71.7 ± 3.0	66.8 ± 3.5	69.2 ± 3.2
Spanish	All	72.0 ± 1.8	71.9 ± 2.0	72.1 ± 2.9	72.0 ± 1.8
	Velar	64.0 ± 0.9	73.6 ± 2.9	54.4 ± 2.4	63.7 ± 0.9
French	All	67.6 ± 2.0	74.3 ± 3.4	61.0 ± 5.0	67.4 ± 2.1
	Front	67.6 ± 1.5	79.7 ± 5.4	55.6 ± 5.0	67.1 ± 1.5

Table 1 presents the results of our analysis on the discriminant power of each language tested, which includes the averages for all classes as well as the most distinguishing class for each language. Arabic stands out as the most distinguishable language with an accuracy of 80.6% and 71.9%, for all classes and the Alveolar class, respectively, making it the easiest to differentiate between native and non-native English speakers. Chinese is the second most discriminant language, with an overall accuracy of 74.2% and 69.2% for the Lateral class. In Spanish, we obtained an accuracy score of 72% and 64% for all classes and the Velar class, respectively. In contrast, French, as shown in Figure 2, is the least discriminant language with an accuracy score of 67.6% for both all and front classes. Our findings suggest that for Arabic, Chinese, and Spanish, all classes perform better in identifying non-native English speakers than relying on a single phonological class.

The alveolar class in Arabic may be more discriminant because it contains emphatic consonants that are not present in English, as reported in a previous study [14]. This

difference in phonemes could explain why Arabic speakers can be more easily differentiated from native English speakers based on their pronunciation. On the other hand, the phonemes in the lateral class of Chinese and English are quite distinct, with Chinese phonemes being dental and English ones being alveolar, according to Wang [19]. Additionally, Spanish has a higher number of velar phonemes than English [15], which, like in Arabic, could contribute to its better discrimination. Finally, some elongated vowels that are common in English but not in the other languages fall into the front class [9], which may be why this class is more important for distinguishing native from non-native speakers.

4.3 Phoneme Analysis

To continue the analysis on the identification of the weakest phonological classes in each native language, we would like to perform an example of how the Phonet system can automatically generate feedback for each phonological class on the phoneme-by-phoneme pronunciation level, compared to a target (native speaker). Figure 3 shows the distribution from a radar plot of the mean posterior for three different speakers for the Alveolar phonological class. In particular, the Non-native 1 and Non-native 2 participants are Arabic native speakers of male gender and 55 and 43 years old, respectively. The Native participant is a native speaker from the USA, female, and 29 years old.

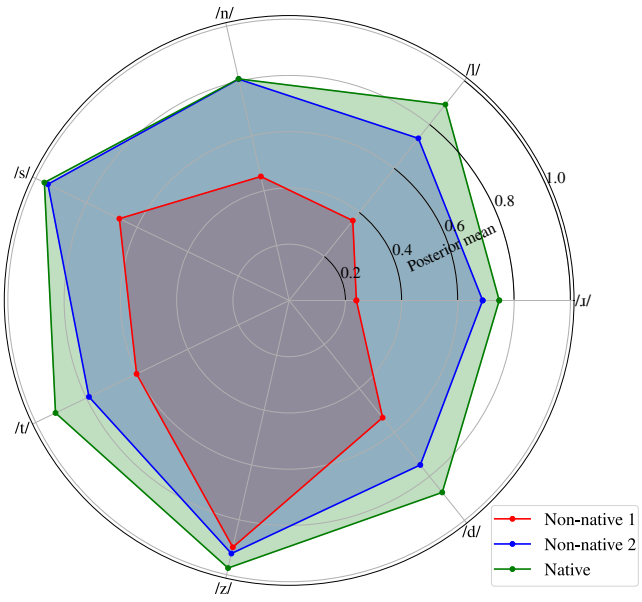

Fig. 3. Comparison of posterior means for the alveolar phonological class of 2 non-native and a native English speaker.

In Figure 3, we can observe that the speaker Non-native 2 (blue color) has similar posterior means to the Native speaker (green color), even equaling in some phonemes such as /s/ and /n/; therefore, we could conclude that this participant has a high level of pronunciation in comparison with a native speaker of English. However, the opposite is the case when we compare the Non-native 1 participant (red color) with the native speaker; in this case, the difference in most of the posterior means of the native speaker vs. non-native speaker is evident. From this figure, we can conclude that this non-native speaker should focus on improving the pronunciation of all phonemes of the Alveolar phonological class, focusing on the phonemes /ɹ/, /l/, and /n/, which is where he shows lower performance compared to a native speaker and even to another person of the same native language.

5 Conclusions

The purpose of our study is to evaluate the pronunciation skills of non-native English speakers from Arabic, Chinese, Spanish, and French backgrounds; we consider training in English a tool called Phonet that allows calculating of posterior probabilities of phonological classes from speech for several groups of phonemes according to the place and manner of articulation. We consider evaluating each non-native speaker from each phonological class and also considering the combination of all of them. In general, the results suggest that Arabic and Mandarin speakers have greater difficulty pronouncing English than Spanish and French speakers. Particularly, when we performed an analysis at the phonological class level, it was possible to identify the Alveolar class as the most distinguishing class for Arabic, the Lateral class for Chinese, the Velar class for Spanish, and the Front class for French. In addition, it was possible to discuss, from previous work, the possible reason why these phonological classes allow to discriminate in a better way each non-native speaker from native speakers of English. Additionally, it was possible to observe that our system can automatically generate feedback for each phonological class on the phoneme-by-phoneme pronunciation level, compared to a target (native speaker) as a strategy to improve their pronunciation performance.

In future work, we will consider training a multilingual system that allows the automatic evaluation of pronunciation not only of English but of different languages. In addition, we will implement multi-class classification of the different non-native speakers involved in this work, including a variety of accents and dialects of each language.

Acknowledgment

This work received funding from UdeA grant # ES92210001 and CODI grant No. PI2023-58010, and PRG2017-15530.

References

1. Bachman, L., Palmer, A.: Language Assessment in Practice: Developing language Assessments and Justifying Their Use in the Real World. Oxford University Press (2022)

2. Berjon, P., et al.: Analysis of french phonetic idiosyncrasies for accent recognition. Soft Computing Letters p. 100018 (2021)
3. Cernak, M., et al.: Characterisation of voice quality of parkinson's disease using differential phonological posterior features. Computer Speech & Language pp. 196–208 (2017)
4. Crystal, D., et al.: English as a Global Language. Cambridge university press (2003)
5. Dai, Y.: An automatic pronunciation error detection and correction mechanism in english teaching based on an improved random forest model. Journal of Electrical and Computer Engineering (2022)
6. Garofolo, J.S.: Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium (1993)
7. Graham, C.: L1 identification from l2 speech using neural spectrogram analysis. In: Proceedings of Interspeech. pp. 3959–3963 (2021)
8. Huang, C., et al.: Accent issues in large vocabulary continuous speech recognition. International Journal of Speech Technology pp. 141–153 (2004)
9. Huerta, E.: Fonética comparada, español-francés, francés como segunda lengua para hispanohablantes, los fonemas complicados (2013)
10. Jenne, S., Vu, N.T.: Multimodal articulation-based pronunciation error detection with spectrogram and acoustic features. In: Proceedings of Interspeech. pp. 3549–3553 (2019)
11. Jiao, Y., et al.: Interpretable phonological features for clinical applications. In: Proceedings of ICASSP. pp. 5045–5049. IEEE (2017)
12. Kim, E., et al.: Automatic pronunciation assessment using self-supervised speech representation learning. arXiv preprint arXiv:2204.03863 (2022)
13. Lesnichaia, M., et al.: Classification of accented english using cnn model trained on amplitude mel-spectrograms. In: Proceedings of Interspeech. pp. 3669–3673 (2022)
14. Millar Cerda, M.A.: Los arabismos en la lengua española (1998)
15. Ocal, A.P.: Fonética contrastiva español/alemán, español/inglés, aspañol/francés y su aplicación a la enseñanza de la pronunciación española (1997)
16. Schnoor, T.T., et al.: Automatic accentedness rating using deep neural networks. In: Proceedings of Meetings on Acoustics ASA. vol. 45, p. 060013. Acoustical Society of America (2021)
17. Singh, Y., et al.: Features of speech audio for accent recognition. pp. 1–6 (2020)
18. Vásquez-Correa, J.C., et al.: Phonet: A tool based on gated recurrent neural networks to extract phonological posteriors from speech. In: Proceedings of Interspeech. pp. 549–553 (2019)
19. Wang, H.Y.: Estudio fónico del chino mandarín y del español (2001)
20. Weinberger, S.: Speech accent archive. George Mason University. Retrieved (2015)
21. Weninger, F., et al.: Deep learning based mandarin accent identification for accent robust asr. In: Proceedings of Interspeech. pp. 510–514 (2019)