

PHONOLOGICAL ERROR DETECTION FOR PRONUNCIATION TRAINING USING NEURAL SPECTROGRAM RECOGNITION

Sabrina Jenne¹, Antje Schweitzer², Sabine Zerbian³, Ngoc Thang Vu⁴

^{1,2,4}Institute for Natural Language Processing, University of Stuttgart, Germany

³Department of Linguistics: English, University of Stuttgart, Germany
{sabrina.jenne, antje.schweitzer, thang.vu}@ims.uni-stuttgart.de, sabine.zerbian@ifla.uni-stuttgart.de

ABSTRACT

We present a neural-based approach to the detection of pronunciation errors in non-native speech, which enables feature-based user feedback in a computer-assisted pronunciation training scenario. Error diagnoses that make reference to phonological classes provide the user with detailed articulatory information, rather than just pointing out mispronounced segments or words.

Several phonological classifiers are trained on raw spectrograms of sounds in isolation and in local phonetic context that are extracted from native English utterances. The models are then used to classify non-native speech segments along distinctive phonological categories purely on the basis of visible spectrogram patterns.

Keywords: language learning, pronunciation error detection, non-native speech

1. MOTIVATION

The popularity of English as *lingua franca* makes foreign language competence indispensable. In secondary schools in Germany, English is an integral part of the foreign language classroom, with approximately 87% of students receiving formal instruction in English in 2014 [23]. However, the success of second language acquisition depends on developmental, environmental and individual factors, leading to varying levels of proficiency throughout and following school education [4]. German native speakers who spend time abroad in exchange programs, complete an international degree or start their career in a globally oriented company might face the challenge of having to improve their English skills autonomously. On top of lexical diversity and grammatical adequacy, appropriate pronunciation contributes to a speaker's credibility [13], career chances [16], self-confidence and sense of belonging in a foreign language environment [22]. As opposed to that, strong German

accents can be seen as unpleasant and unfriendly [14] or even as unattractive [9].

Existing tools such as `NativeAccent` or `Duolingo` make use of speech recognition and provide feedback in terms of instructions, video material of facial movements, or pointers to the erroneous item [10][26]. `Transparent Language`, as opposed to that, 'compares' the recorded user voice to a native speaker's recording and shows the waveforms for comparison [25]. A different approach is taken by `ReLANpro BYOLL`, in which users can upload their recordings to a cloud learning server and receive feedback from instructors [5]. Our system differs from these approaches mainly in its use of spectrogram-based classification on the one hand and the output of detailed phonological diagnoses on the other hand, providing more specific feedback.

This paper presents the use of neural classification models, which are powerful in pattern recognition, to detect pronunciation errors in non-native speech. Speech spectrograms have previously been used in neural speech emotion recognition [20][29][30], as well as language identification [17] and pronunciation pathology detection [2] using Support Vector Machines. We show that this representation can also be used in the detection of erroneous segments in non-native speech. At its core, the model presented here learns to classify pronunciation errors on a segmental level, based solely on information visible in the segment's spectrogram. A convolutional neural network (CNN) is trained to recognize phonological categories from spectral representations, thereby revealing not only erroneous segments, but also the erroneous characteristic, as for example the place of articulation. As a result, discrepancies between the model output and the intended sound can be used to provide direct, applicable feedback to the user.

While neural networks work exceedingly well, they lack explanatory power, in the sense that it is not known which visual patterns are attended to. In this paper, we show that the model is capable

Table 1: Sounds of interest and potential confusions.

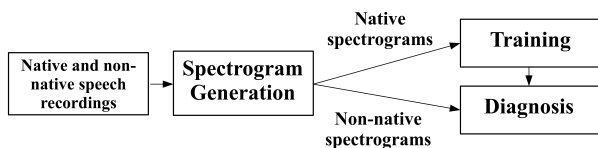
Sound	Confusions	Sound	Confusions
/æ/	[ɛ], ...	/b/	[p], ...
/ð/	[z], [d], ...	/d/	[t], ...
/θ/	[s], [t], ...	/g/	[k], ...
/w/	[v], ...	/v/	[f], ...
		/z/	[s], ...

of making pronunciation adequacy judgments that are comprehensible and useful from a linguistic perspective, underlining the attractiveness of neural networks despite the black-box problem.

2. ARCHITECTURE

The entire pipeline is shown in Figure 1. The approach takes native and non-native spectrograms on two contextual levels as input. The native English segment spectrograms serve as training data for several phonological classifiers, which can then be used to detect potentially erroneous features in non-native spectrograms.

Figure 1: Pronunciation error detection pipeline.



2.1. Data and sounds of interest

To evaluate the usefulness of the model’s predictions, they must be compared against a gold standard, indicating whether the sound was pronounced correctly or, if it was mispronounced, what was said instead. As it is not feasible to hand-label the entire non-native data set on a phone level, Table 1 specifies a set of sounds that are found to be particularly challenging for native German learners of English, as well as frequent confusions [1][9][21][22]. Due to the contrastive nature of sounds, errors can lead to deficits in comprehensibility, as for example in the lack of a distinction between <think> and <sink> or <bad> and <bed>.

We use speech data from the Speech Accent Archive [27], which is a platform to which volunteers can upload their reading of a standardized English paragraph. The text is designed to contain all sounds of English, as well as a few challenging sound combinations. The archive features record-

ings from over 300 native language backgrounds. As non-native speech corpus, we use 36 recordings provided by native speakers of German (22 female, 14 male). Since the transcriptions provided by the Speech Accent Archive lack time alignments and use narrow annotations that mismatch the phonemic output of our system, we decided to gold label the non-native utterances ourselves. In the process, the first author of the paper listened to each interval that was identified as a sound of interest in the alignment, and labeled it with the perceived sound. This allows us to evaluate how well the model detects actual pronunciation errors. To train the classifiers, the recordings of 102 native English speakers are used (58 female, 44 male).

2.2. Spectrogram generation

Our classification system relies on the availability of phone-based speech segments. To that end, the data from the Speech Accent Archive must be time-aligned with their transcription. This task can be accomplished with the acoustic model of an Automatic Speech Recognition (ASR) System, which, in the scenario presented here, where the spoken text is known, can be used to find phone boundaries. ASR performance suffers when dealing with non-native speech [6][24][28][31], which also means that segment boundaries are not necessarily perfect or exact. However, the envisioned application must segment the user’s speech *ad hoc*, meaning that noisy segment boundaries represent an authentic, real-world scenario. We show that the model performs well despite this potential weakness.

The ASR toolkit of choice is Kaldi [18]. The speech recognizer is trained on the TED-LIUM 3 corpus, which contains 452 hours of transcribed speech data from 2028 unique speakers [8]. It features a triphone HMM-GMM acoustic model with speaker-adaptive training. The alignment process results in Praat TextGrids that contain phone-based segment boundaries [3]. Praat is then used to create a visible spectrogram for each segment. We use a maximum frequency of 8kHz, a Gaussian window of 5 milliseconds length, a dynamic range of 70dB, and autoscaling to ensure optimal visibility of spectral patterns. We further explore two segmentation variants: one in which the sound is segmented in isolation and one in which the sound is segmented along with the preceding and the following interval. The latter allows the model to access local phonetic context, which can be helpful in the face of co-articulation effects and prominent acoustic transitions, such as the lowering of the third formant in rhotacized vowels [11][15]. Additionally, the inclu-

sion of the preceding and following interval is expected to attenuate the noisy alignment problem.

2.3. Training

The classification of phonological features based on spectrograms is done by a Convolutional Neural Network (CNN), which is particularly suitable for image recognition. We construct a LeNet which features two convolutional layers in Keras [7][12][19]. The first layer learns 20 filters of size 5×5 , followed by a ReLU activation function and 2×2 max-pooling. The second layer learns 50 filters of size 5×5 , again followed by ReLU activation and 2×2 max-pooling. The output is flattened and fed into a fully-connected layer with 500 nodes and ReLU activation, followed by a last fully-connected layer with softmax activation. Each model is trained with the Adam optimization algorithm for 25 epochs. The learning rate is 0.001 and the batch size is 32. Before being fed to the model, all spectrograms are resized to 28×28 pixels. Since we find that the usage of input images of size 64×64 and 128×128 pixels does not increase accuracy while drastically increasing training time, we assume that the relatively small size of 28×28 pixels is sufficient for the model.

The model outputs a probability distribution over the respective classes of a category, meaning that one model is trained for each phonological category. If a segment is classified as vowel in the first step (classifier Class), it is further classified along the parameters of height, fronting, rounding and tenseness. Conversely, if a segment is classified as consonant, it is further classified along its place of articulation, manner of articulation, and voicing status. The number of classes thereby depends on the category. In total, eight models are trained five times with random initialization. 25% of the native English spectrograms thereby serve as test data. The accuracy of each model is averaged over all five runs and given in Table 2. Categories are listed with the respective number of classes in parentheses.

The best and worst result for each context level are boldfaced. When considering sounds in isolation, apparently the easiest parameter to detect is the major class. This observation coincides with the fact that the prominent formant structure of vowels is visibly different from the spectral signature of most consonants, which display patterns such as noise or bursts. The lower performance on major class when considering sounds in context could be explained by the fact that the model is confronted with the acoustic signature of several segments, with vowels and consonants interspersed. In contrast, the increased performance in detecting lip rounding when consid-

Table 2: Accuracy of classifiers on the test set, using sound spectrograms in isolation and in context. \pm indicates standard deviation.

	ISOLATION	CONTEXT
CATEGORY	Acc %	Acc %
Class (2)	89.34 ± 0.5	86.04 ± 0.7
Fronting (4)	70.35 ± 2.9	77.28 ± 1.0
Height (5)	64.19 ± 0.5	71.59 ± 1.3
Rounding (2)	86.00 ± 0.4	86.12 ± 1.6
Tenseness (2)	76.33 ± 1.0	74.81 ± 1.7
Place (7)	68.78 ± 0.4	72.25 ± 1.6
Manner (6)	74.06 ± 0.9	75.11 ± 0.9
Voicing (2)	83.18 ± 0.5	81.07 ± 1.4

ering sounds in context coincides with the fact that the second and the third formant tend to be lowered in a rounded vowel [11], which might be enhanced by the visibility of local transitions. On both context levels, the vowel height classifier performs worst. One potential explanation might be that the first formant, which correlates with vowel height, could be displayed more or less conflated with the fundamental frequency, causing the classifier to mistake the second formant as first formant in such cases.

The performance of the various models in Table 2 provides important insight into the functionality of the system. Even though neural models suffer from the black-box problem, meaning that users do not know which patterns in the image are attended to, the model outlined in this paper is capable of making judgments that are linguistically justifiable.

In order to evaluate whether the spectrogram-based approach offers an advantage over more carefully extracted features, we compute 13 MFCC features for each sound instance, paint the corresponding values over time in Praat, and use the resulting images to train the model. Interestingly, the spectrogram-based models outperform MFCC-based models by 3.4% in the isolated scenario and 5.6% in the context scenario on average. To test the impact of the number of features, we re-train the model on images based on 24 MFCC features. The difference in performance to the spectrogram-based models even increases to 6.9% in the isolated scenario and 9.4% in the context scenario on average, which supports the hypothesis that the system benefits from patterns present in raw spectrograms without explicit feature extraction. The accuracy of MFCC-based models for each category, using 13 features, is given in Table 3, averaged over five runs.

Table 3: Accuracy of classifiers on the test set, using MFCC-based representations of sounds in isolation and in context. \pm indicates standard deviation.

	ISOLATION	CONTEXT
CATEGORY	Acc %	Acc %
Class	83.34 \pm 1.8	83.14 \pm 1.2
Fronting	69.26 \pm 0.5	72.91 \pm 1.4
Height	65.27 \pm 1.1	65.75 \pm 2.0
Rounding	87.85 \pm 0.4	85.47 \pm 0.6
Tenseness	75.10 \pm 1.4	69.96 \pm 1.2
Place	61.00 \pm 0.7	62.04 \pm 0.9
Manner	64.36 \pm 1.2	63.29 \pm 0.7
Voicing	78.76 \pm 0.7	76.93 \pm 1.7

2.4. Diagnosis

Classification is done by feeding the spectrograms extracted from 36 native German speakers of English to the phonological classifiers described in section 2.3. For each category, the best model is used. The combined output of all models allows us to uniquely identify the sound that is recognized. The recognized sound is then compared to the gold label on the one hand, which specifies the actually realized sound, and the target sound on the other hand, which conforms to the canonical pronunciation of a word. This three-way comparison leads to five scenarios, exemplified on the sound /ð/ in Table 4. In all cases, /ð/ is the target realization, as for example in the word <brother>.

Table 4: Evaluation scenarios.

	Recognition	Gold	Target
True positive (TP)	[ð]	[ð]	/ð/
True negative (TN)	[d]	[d]	/ð/
False positive (FP)	[ð]	[d]	/ð/
False negative (FN)	[d]	[ð]	/ð/
Ambiguous (A)	[z]	[d]	/ð/

True positives, true negatives and ambiguous scenarios are particularly interesting. While true positives can be seen as correctly pronounced sounds, true negatives can be seen as mispronunciations. In both cases, the model is capable of predicting the sound exactly. Ambiguous cases are interesting in so far that they indicate a mispronunciation as well, however, as opposed to true negatives, there is no agreement on what was said instead. Since false positives and false negatives denote a disagreement between the annotator and the model with respect to pronunciation accuracy itself, no adequate response to such cases is currently possible.

3. RESULTS

Diagnosis on non-native speech segments is done separately for segment spectrograms in isolation and in context. The proportion of the five evaluation scenarios is shown in Table 5.

Table 5: Results for each evaluation scenario, using the best model for each category.

Isolation	%	Context	%
TP	18.40	TP	22.63
TN	10.50	TN	6.44
FP	14.97	FP	18.11
FN	26.93	FN	21.65
A	29.19	A	31.17

The model is right in detecting whether a sound is pronounced correctly or not in 58.09% of cases in the isolated scenario (sensitivity: 40.59%, specificity: 72.61%), and in 60.24% of cases in the context scenario (sensitivity: 51.11%, specificity: 67.50%), counting ambiguous cases as correctly localized pronunciation errors.

4. DISCUSSION AND OUTLOOK

We present a pipeline that is capable of judging the pronunciation accuracy of German native speakers of English, based on neural image classification on spectrograms that display sounds in isolation or in local context. The model's judgment on the correct or incorrect pronunciation of a sound is appropriate in roughly 60% of cases, enabling relatively reliable, feature-oriented diagnoses. Future studies will further assess the pedagogical value of this type of feedback. Furthermore, even though it is not known which characteristics or patterns in the image are learned exactly, the model's decisions are comprehensible and useful from a linguistic viewpoint. Improved alignment quality and the consultation of a second annotator, as well as a comparison to the transcriptions provided by the Speech Accent Archive are expected to increase the confidence of pronunciation accuracy judgments even more.

5. REFERENCES

- [1] Avery, P., Ehrlich, S. 1992. *Teaching American English Pronunciation*. Oxford: Oxford University Press.
- [2] Bodusz, W., Miodońska, Z., Badura, P. 2018. Approach for spectrogram analysis in detection of selected pronunciation pathologies. In: Gzik, M., Tkacz, E., Paszenda, Z., Piętka, E., (eds), *Innova-*

- tions in *Biomedical Engineering*. Cham: Springer International Publishing 3–11.
- [3] Boersma, P., Weenink, D. 2018. Praat: Doing Phonetics by Computer. <http://www.fon.hum.uva.nl/praat/>.
- [4] Brice, A. E. 2015. Multilingual Language Development. In: *International Encyclopedia of the Social & Behavioral Sciences*. 57–64.
- [5] Burston, J. 2017. ReLANpro BYOLL (Bring Your Own Language Lab). Learning Technology Review 34.1. Computer Assisted Language Instruction Consortium.
- [6] Byrne, W., Knodt, E., Khudanpur, S., Bernstein, J. 1998. Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English. *Proceedings of Speech Technology in Language Learning* Marholmen. 37–40.
- [7] Chollet, F. 2015. Keras. <https://keras.io>.
- [8] Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., Estève, Y. Sept. 2018. TEDLIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. *International Conference on Speech and Computer* Leipzig. 198–208.
- [9] Knopf, K. 1975. *English Pronunciation Exercises: Sprachlaborkurs nach didaktischen Schwerpunkten der Ausgangssprache Deutsch*. München: Wilhelm Fink Verlag.
- [10] Korslund, S. 2018. NativeAccent. Learning Technology Review 35.2. Computer Assisted Language Instruction Consortium.
- [11] Ladefoged, P., Johnson, K. 2011. *A Course in Phonetics*. Boston: Wadsworth Cengage Learning.
- [12] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. Nov. 1998. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE* 86(11), 2278–2324.
- [13] Lev-Ari, S., Keysar, B. Nov. 2010. Why Don't We Believe Non-Native Speakers? The Influence of Accent on Credibility. *Journal of Experimental Social Psychology* 46(6), 1093–1096.
- [14] Lindemann, S. 2005. Who speaks “broken English”? US undergraduates' perceptions of non-native English. *International Journal of Applied Linguistics* 15(2), 187–212.
- [15] Machač, P., Skarnitzl, R. 2009. *Principles of Phonetic Segmentation*. Prague: Epocha Publishing House.
- [16] Ministerium für Kultus, Jugend und Sport, Nov. 2012. Fremdsprachen in den weiterführenden Schulen. Baden-Württemberg, Postfach 10 34 42, 70029 Stuttgart.
- [17] Montalvo, A., Costa, Y. M. G., de Lara, J. R. C. 2015. Language Identification Using Spectrogram Texture. *CIARP: Iberoamerican Congress on Pattern Recognition* Montevideo. 543–550.
- [18] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. Dec. 2011. The Kaldi Speech Recognition Toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding Hawaii*.
- [19] Rosebrock, A. Dec. 2017. Image classification with Keras and deep learning. <https://www.pyimagesearch.com/2017/12/11/image-classification-with-keras-and-deep-learning/>.
- [20] Satt, A., Rozenberg, S., Hoory, R. 2017. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. *Proc. Interspeech 2017* Stockholm. 1089–1093.
- [21] Sauer, W. 2001. *American English Pronunciation: A Drillbook*. Heidelberg: Winter.
- [22] Schmitt, H. 2016. *Teaching English Pronunciation*. Heidelberg: Winter.
- [23] Statistisches Bundesamt Wiesbaden, Mar. 2016. Schulen auf einen Blick. <https://www.destatis.de/>.
- [24] Tan, T. P., Besacier, L. 2006. A French Non-Native Corpus for Automatic Speech Recognition. *International Conference on Language Resources and Evaluation* Genoa. 1610–1613.
- [25] Tang, X. 2018. Transparent Language for Learning Chinese. Learning Technology Review 35.3. Computer Assisted Language Instruction Consortium.
- [26] Teske, K. 2017. Duolingo. Learning Technology Review 34.3. Computer Assisted Language Instruction Consortium.
- [27] Weinberger, S. 2015. Speech Accent Archive. <http://accent.gmu.edu/>.
- [28] Winebarger, J., Stüker, S., Waibel, A. 2014. Adapting Automatic Speech Recognition for Foreign Language Learners in a Serious Game. *Games and Natural Language Processing: Papers from the AI-IDE Workshop* Raleigh. 38–40.
- [29] Yang, Z., Hirschberg, J. 2018. Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks. *Proc. Interspeech 2018* Hyderabad. 3092–3096.
- [30] Yenigalla, P., Kumar, A., Tripathi, S., Singh, C., Kar, S., Vepa, J. 2018. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. *Proc. Interspeech 2018* Hyderabad. 3688–3692.
- [31] Zapata, J., Kirkedal, A. S. 2015. Assessing the Performance of Automatic Speech Recognition Systems When Used by Native and Non-Native Speakers of Three Major Languages in Dictation Workflows. *Proceedings of the 20th Nordic Conference of Computational Linguistics* Vilnius. 201–210.