Multimodal Articulation-Based Pronunciation Error Detection with Spectrogram and Acoustic Features

Sabrina Jenne, Ngoc Thang Vu

Institute for Natural Language Processing (IMS)
Universität Stuttgart, Germany
{sabrina.jenne, thang.vu}@ims.uni-stuttgart.de

Abstract

Articulation-based pronunciation error detection is concerned with the task of diagnosing mispronounced segments in non-native speech on the level of broad phonological properties, such as place of articulation or voicing. Using acoustic features and visual spectrograms extracted from native English utterances, we train several neural classifiers that deduce articulatory properties from segments extracted from non-native English utterances. Visual cues are thereby processed by convolutional neural networks, whereas acoustic cues are processed by recurrent neural networks.

We show that combining both modalities increases performance over using models in isolation, with important implications for user satisfaction. Furthermore, we test the impact of alignment quality on model performance by comparing results on manually corrected segments and force-aligned segments, showing that the proposed pipeline can dispense with manual correction.

Index Terms: pronunciation error detection, non-native speech, L1 German, L2 English

1. Introduction

Foreign language proficiency, in particular pronunciation skills, is found to contribute to a speaker's credibility and career chances [1], while poor pronunciation might cause the impression of unfriendliness and disinterest [2][3]. As a consequence, speakers who participate in exchange programs, complete international degree programs or vocational trainings, or start their career in a globally oriented company might have to improve their pronunciation skills autonomously. The demand for flexible, self-determined training underlines the attractiveness of automatic solutions. Accordingly, the task of pronunciation error detection concerns the automatic localization and diagnosis of incorrectly pronounced sounds in non-native speech, with the aim to provide valuable and interpretable feedback to the user. We hereby focus on native German speakers of English.

Existing tools make use of automatic speech recognition technology and offer feedback by highlighting the error, showing video material of facial movements, or providing written instructions [4][5]. The proposed approach focuses on errors on a segmental level and differs from these tools in the availability of phonologically grounded, class-based diagnoses. Specifically, the model learns to deduce broad phonological categories, such as place of articulation or voicing status, from individual segments, which can be transformed into direct, articulatory feedback. If a user fails to produce dental sounds, for instance, they could be asked to attempt a lisp-like sound in an effort to acquire the correct tongue position. Articulatory instructions are found to be helpful in areas such as speech therapy [6], suggesting applicability in non-native pronunciation training as well.

The intuition of our approach is to utilize two sources of information: visual cues present in spectrograms, such as noise bursts or formant bands, and acoustic features captured by MFCCs. Spectrograms, which are a valuable source of information to phonetic experts when doing manual segmentation [7], have been used for tasks such as speech emotion recognition [8][9][10], pronunciation pathology detection [11] and style transfer [12][13]. Furthermore, we suggested the viability of spectrogram-based representations for the purpose of pronunciation error detection in previous work [14]. MFCC features, likewise, are frequently used in automatic speech recognition systems [15], speaker recognition or verification systems [16][17], and other speech processing tasks. We show that the combination of human-interpretable representations and robust, yet compressed representations of speech improves pronunciation error detection. Integrating several sources of information is a popular approach in language processing tasks. In [18], MFCCs are combined with word embeddings for the purpose of speaker segmentation and diarization, [19] integrate textual features such as part-of-speech tags with spectral features for the purpose of emotion recognition, and MFCCs are combined with mouth movement parameters for the purpose of speech recognition in [20].

A second focal point of this paper is alignment quality. As manually prepared alignments are not available in an online setting, the user's speech must be aligned automatically, which might add distracting noise to the sound representations. We test the impact of alignment quality by conducting all experiments on manually and force-aligned test data. The paper is organized as follows: in the next section, we introduce both model types, as well as the data and the sounds of interest. In section 3, we discuss the error detection experiments we conducted with models in isolation and in combination. Section 4 then presents results as well as more detailed error analyses. We conclude in section 5.

2. Model

The entire pipeline is shown in Figure 1. As can be seen, we use different neural networks for spectrograms and MFCCs – CNNs and RNNs, respectively – to optimally account for the different representation types. CNNs, accordingly, are suited for pattern recognition [21], whereas RNNs are suited for sequential data [22].

For each data type, we train eight classifiers that recognize the following phonological properties from each segment, with the number of classes in parentheses: major class (3), tongue height (6), tongue fronting (5), vowel tenseness (3), lip rounding (3), voicing status (3), place of articulation (9), and manner of articulation (9). All categories feature a "none" class, which is used for properties that are only relevant for the other major

Table 1: *Sounds of interest and potential confusions.*

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

Table 2: *Architecture and parameters of both model types.*

MFCC	SPEC
three LSTM layers	two convolutional layers:
256 hidden units each	1) 20 filters of size 5×5
	2) 50 filters of size 5×5
tanh activations	ReLU activations
	two 2×2 max pooling layers
	fully-connected layer
output layer (softmax)	output layer (softmax)
batch size: 128	batch size: 32
Adam optimizer	Adam optimizer

class. If a segment is classified as consonant, for instance, the height, fronting, tenseness and rounding classifiers should ideally return the class "none". All models are trained five times with random initialization for 13 epochs. We report the average accuracy and standard deviation of each classifier, computed on 25% of the training data that is randomly set aside before training.

2.1. Data

We train all models on speech segments extracted from the TIMIT corpus [23], which provides the recordings of ten sentences read by 630 native English speakers. The underlying assumption is that native English segments do not exhibit pronunciation errors, enabling the models to learn appropriate target patterns. Furthermore, all word- and phone-level transcriptions have been manually verified, indicating relatively reliable time alignments. The native English training data comprises about 177,000 segments, spanning all eight dialect regions represented in the corpus.

We test the pipeline on segments extracted from 36 native German utterances that we obtain from the Speech Accent Archive [24], an online platform to which volunteers can upload a read speech sample of a pre-defined English paragraph. For the FORCE experiments, all utterances were force-aligned using a HMM-GMM trained on the TED-LIUM 3 corpus [25]. For the GOLD experiments, phone boundaries were manually corrected by a phonetic expert. The non-native utterances are expected to exhibit pronunciation errors, which is why we select a set sounds that have been found to be challenging for native German learners of English [26][3]. The sounds of interest and potential confusions are listed in Table 1; obstruents below the double line thereby refer to syllable-final instances only. In order to identify mispronounced sounds, all instances of the sounds of interest have been labeled with the actually spoken sound by the first author of this paper. The non-native test data comprises 1,723 hand-labeled sounds of interest.

2.2. Data representation types

As acoustic features, we extract 13 MFCCs using a 15 millisecond window and a 5 millisecond frame shift. The feature vectors corresponding to a single segment are then concatenated to ensure segment-based processing [27]. The classifiers' architecture and hyperparameters are defined in the first column of Table 2. Accuracy is given in the second column of Table 3.

As visual cues, we extract a spectrogram in Praat for each segment [28]. Except for the maximum frequency, which was increased to 8kHz, all default settings are used. Before being fed to the classifiers, all images are resized to 28×28 pixels. The classifiers' architecture and hyperparameters are defined in the second column of Table 2 [21][29], accuracy is given in the third column of Table 3 (SPEC). For each category, the data representation type that performs better is boldfaced. Spectrogram-based models outperform MFCC-based models by 3.17% on average, with superior performance in all categories except for voicing. The performance gap is most prominent for tenseness (+8.49%) and rounding (+5.71%), which leads to the assumption that spectrograms preserve valuable visible cues such as formant bands. A voice bar, as opposed to that, can be visually concealed by other patterns, potentially explaining the superiority of MFCC-based representations in this category.

Table 3: *Average accuracy of MFCC- and spectrogram-based classifiers on a randomly set aside test set.* ± *indicates standard deviation.*

Category	MFCC	SPEC
Class	89.15% ±0.2	**93.08%** ±0.2
Height	81.83% ±0.3	**84.08%** ±0.5
Fronting	83.19% ±0.4	**85.83%** ±0.6
Tenseness	82.80% ±0.4	**91.30%** ±0.2
Rounding	87.33% ±0.3	**93.04%** ±0.3
Voicing	**90.08%** ±0.1	88.92% ±1.0
Place	74.04% ±0.1	**76.87%** ±0.4
Manner	80.21% ±0.3	**80.88%** ±0.9

3. Experiments

For pronunciation error detection, we apply the best model of each category to the non-native data extracted from the Speech Accent Archive. For the GOLD experiments, the pre-trained models are applied to the manually corrected segments, whereas for the FORCE experiments, the pre-trained models are applied to the force-aligned segments. The combination of all predicted classes suffices to determine the predicted sound label. However, as the combination of the classes with the highest probability occasionally fails to map to a sound of English (<voiced dental plosive>, for instance, is not present in the phone set), we enforce the model to decide on a phone label. For this purpose, we scan all possible combinations of all classes that map to a sound of English, and compute the probability of each com-

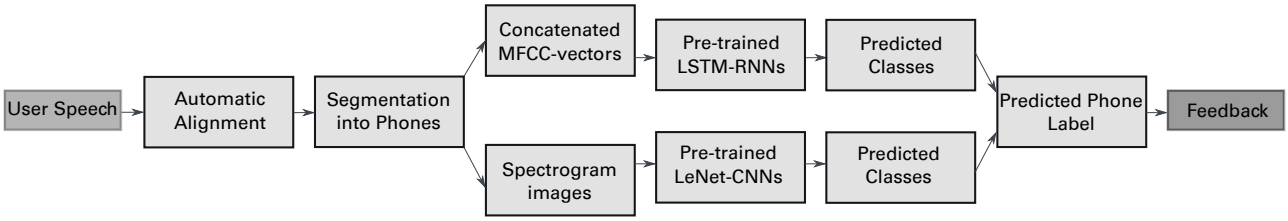

Figure 1: *Error detection pipeline in the FORCE setting.*

Table 4: *Results in the GOLD and FORCE setting, using both models separately.*

	GOLD		**FORCE**	
	MFCC	SPEC	MFCC	SPEC
Sensitivity	19.02%	**19.78%**	**19.00%**	16.14%
Specificity	**84.78%**	84.16%	**85.34%**	83.37%

Table 5: *Results in the GOLD and FORCE setting, using combined predictions.*

	GOLD	**FORCE**
Sensitivity	22.95%	21.41%
Specificity	81.56%	80.79%

bination by taking the average of its classes. As a result, we output the phone label with the highest average probability.

Evaluation is done on the basis of three labels: the canonical target sound as it would be expected from a native speaker, the predicted sound that corresponds to the combination with the highest average probability, and the human label provided by manual annotation, as explained in section 2.1. This allows us to compute sensitivity and specificity. The evaluation scheme thereby assumes the perspective of the user, meaning that correctly pronounced sounds are seen as *positives*, whereas pronunciation errors are seen as *negatives*. Consequently, sensitivity quantifies the proportion of correctly pronounced sounds (true positives + false negatives) classified as correct (true positives), whereas specificity quantifies the proportion of incorrectly pronounced sounds (true negatives + false positives) classified as errors (true negatives). A trade-off between these values is crucial to ensure that errors are detected reliably (high specificity), while correctly pronounced sounds are accredited (high sensitivity).

First, we test the performance of both modalities separately. For this purpose, we compute sensitivity and specificity for MFCC-based segments and spectrogram-based segments in isolation. To show the benefit of multimodal classification, we then combine these model types. This is done by averaging the predicted probabilities assigned to each class of a category by both model types prior to computing the highest-scoring combination. In other words, the predicted phone label corresponds to the best combination across both classifier types, thereby employing modality fusion at decision-level [19].

4. Results

Table 4 reports sensitivity and specificity for spectrogram- and MFCC-based classifiers in isolation. For each alignment setting, we boldface the modality that performs better. In the GOLD setting, spectrogram-based representations outperform MFCC-based representations with respect to sensitivity, whereas MFCC-based representations perform better with respect to specificity. In the FORCE setting, MFCC-based representations are superior with respect to both values, whereby the result for specificity even increases over the GOLD setting. We draw the following, preliminary conclusions: first, results for specificity are consistently high for both modalities in both settings. Much poorer results for sensitivity, in contrast, indicate that both models are overly strict in assessing sounds. Second, the performance gap resulting from the usage of force-aligned segments is surprisingly small overall, which is particularly encouraging in view of the envisioned user application. Third, MFCC-based representations outperform spectrogram-based representations in the FORCE setting, even though spectrogram-based classifiers are superior with respect to average accuracy, as shown in Table 3. We therefore assume that MFCC-based classifiers generalize well to non-native test samples, particularly in view of noisy alignments.

As can be seen in Table 4, spectrogram- and MFCC-based representations in isolation are capable of detecting a large proportion of pronunciation errors. This happens, however, at the expense of sensitivity, risking that users are demotivated by seemingly unjustified diagnoses. Consequently, further improvements should aim at increasing sensitivity, potentially at the expense of lowering specificity. In Table 5, we show results for both modalities combined, which indicates that combining both data representation types indeed improves sensitivity. Furthermore, differences between the GOLD and FORCE setting remain relatively small.

In order to better understand the strengths and weaknesses of the model, we also analyze phone confusions. Comparing each sound of interest to its most frequently predicted (incorrect) phone label sheds light on the phonological features that underlie the error, thereby revealing those categories that are particularly challenging for the model. When considering phone confusions of MFCC- and spectrogram-models in isolation, we find that frequent confusions relate to almost all categories, with the exception of lip rounding and tongue fronting. This leads to the impression that both models tend to guess phone labels when acting in isolation.

As opposed to that, when considering phone confusions of the combined model, we find that frequent confusions mainly relate to the categories of manner and place of articulation in the case of consonants, and tongue height in the case of the vowel. Voicing errors are constrained to occurrences of /z/ and /θ/, whereas we do not find frequent confusions due to wrong major class, tongue fronting, lip rounding, or vowel tenseness.

Table 6: *Error statistics for three example speakers with varying accent ratings.*

Accent	Sound	Count	#Errors	#Predicted errors
Weak	/ð/	6	5	5
	/d/	6	3	3
	/z/	12	8	5
Medium	/ð/	6	6	6
	/d/	5	3	3
	/z/	12	8	0
	/w/	5	5	5
Strong	/w/	5	3	3
	/ð/	6	3	3

This observation can be traced back to several factors. First, as shown in section 2, the height, place and manner classifiers have the largest number of classes to be distinguished, which increases difficulty. Second, these phonological categories might be harder to deduce from individual segments in general, exemplified by relatively low accuracy scores as shown in Table 3. In summary, multimodal error diagnosis does not only increase sensitivity, but also helps limiting prediction errors to a few, problematic categories.

4.1. Error analysis: speaker statistics

We suggest that in an actual pronunciation training tool, the model should prioritize systematic errors over individual mispronounced segments, as the latter could be traced back to random variations or personal factors such as fatigue. To single out how well the model recognizes systematic errors, we choose one speaker with a strong, medium and weak non-native accent, respectively. Accent ratings have been obtained by consulting a native English speaker on the comprehensibility, intelligibility and overall impression of a speaker's utterance. These scores are then averaged, with one being the lowest and five being the highest score. The weak-accented speaker, accordingly, received a mean score of 5, the medium-accented speaker received a mean score of 3.67, and the strong-accented speaker received a mean score of 2, which was the lowest score among all 36 speakers.

For each speaker, we count the total number of occurrences of each sound of interest (target sound) in the utterance. Sounds that occur less than five times are excluded. A pronunciation error is deemed to be systematic if the human label differs from the target sound in at least 50% of its total occurrences in the utterance. Lastly, we count the number of detected pronunciation errors for each sound of interest, using combined predictions. The assumption thereby is that the model should be able to detect systematic pronunciation errors over entire utterances, with less focus on the exact location of the error. Results for each speaker are presented in Table 6. We only show systematic errors, that is, target sounds that occur at least five times (Count) and that are mispronounced in at least 50% of their total occurrences in the utterance (#Errors). It can be seen that in the majority of cases, the model recognizes systematic errors almost perfectly (#Predicted errors). In other words, the model succeeds at reliably detecting the sounds that a speaker struggles with in a broader context.

Exceptions can be seen for the weak- and medium-accented speaker regarding the classification of syllable-final /z/. In the former, the model detects only five out of eight mispronunciations of /z/, whereas in the latter, the model fails to detect errors in the pronunciation of /z/ entirely. Thus, even though the model tends to be too strict, as indicated by low sensitivity and high specificity values, it is overly tolerant in the case of /z/. We trace this observation back to the distribution of training samples, whereby an abundance of samples in the classes `voiced`, `alveolar` and `fricative` could lead to the excessive generation of selfsame.

The presented speaker statistics offer several observations that are relevant to our core principles in pronunciation training. First, we would like to draw attention to the fact that the chosen speakers consistently struggle with the same set of sounds: all speakers seem to encounter difficulties in the pronunciation of /ð/, whereas two speakers exhibit errors concerning /d/, /z/ and /w/, respectively. This finding validates our assumption on the difficulty of these sounds for native German learners of English. Second, it can be seen that the medium-accented speaker exhibits the largest number of systematic pronunciation errors, whereas the strong-accented speaker seems to encounter difficulties with only two sounds of interest. On the one hand, this could be traced back to the selection of sounds, meaning that the strong-accented speaker might exhibit difficulties with sounds that are not considered here. On the other hand, this finding stresses the complexity of non-native accents and their perceptibility. Since we consider only phonemic errors, allophonic, intonational, rhythmic or durational factors are not accounted for here, while they might have played an important role in the accent ratings [30][31].

5. Discussion

In this paper, we present a pipeline that identifies pronunciation errors in non-native speech by deducing broad phonological categories from visual cues on the one hand, and acoustic features on the other hand. We train eight phonological classifiers for both modalities, which employ a CNN architecture using spectrogram images as input, and a LSTM-RNN architecture using MFCCs as input. Segments are extracted either from manually corrected or automatically aligned phone boundaries. While both models in isolation achieve satisfactory results with respect to specificity, performance is relatively poor with respect to sensitivity. Combining the prediction of both model types is shown to attenuate this disparity, while limiting model errors to a small number of categories that seem hard to classify in general. Furthermore, we show that the combined model succeeds at identifying a speaker's systematic pronunciation errors, which allows for prioritization of certain segments in training.

With respect to our claims in the introduction, we come to the following conclusions: first, the performance of MFCC- and spectrogram-based models is comparable in isolation, and yields impressive results for specificity. Combining predictions, however, is particularly useful to increase sensitivity, which is indispensable to avoid user frustration and loss of motivation. Second, performance gaps due to alignment quality are found to be relatively small, which is an encouraging revelation in the face of the envisioned, fully automatic user application.

Future research will focus on testing the pedagogical usefulness of our approach in an actual user study, which enables us to correlate evaluation scores, such as sensitivity and specificity, with user experience and progress. Furthermore, we will explore several feedback types, such as written articulatory instructions, animations, or synthetic accent removal [32].

6. References

[1] S. Lev-Ari and B. Keysar, "Why don't we believe non-native speakers? The influence of accent on credibility," *Journal of Experimental Social Psychology*, vol. 46, no. 6, pp. 1093–1096, 2010.

[2] S. Lindemann, "Who speaks "broken English"? US undergraduates' perceptions of non-native English," *International Journal of Applied Linguistics*, vol. 15, no. 2, pp. 187–212, 2005.

[3] H. Schmitt, *Teaching English Pronunciation*. Heidelberg: Universitätsverlag Winter, 2016.

[4] S. Korslund, "NativeAccent," Computer Assisted Language Instruction Consortium, Learning Technology Review 35.2, 2018. [Online]. Available: https://journals.equinoxpub.com/index.php/CALICO/article/view/33577/pdf

[5] K. Teske, "Duolingo," Computer Assisted Language Instruction Consortium, Learning Technology Review 34.3, 2017. [Online]. Available: https://journals.equinoxpub.com/index.php/CALICO/article/view/32509/pdf

[6] R. Duan, T. Kawahara, M. Dantsuji, and J. Zhang, "Effective Articulatory Modeling for Pronunciation Error Detection of L2 Learner without Non-Native Training Data," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, 2017, pp. 5815–5819.

[7] A. Khan, I. Steiner, Y. Sugano, A. Bulling, and R. Macdonald, "A Multimodal Corpus of Expert Gaze and Behavior during Phonetic Segmentation Tasks," in *11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, 2018, pp. 4277–4281.

[8] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in *Proc. Interspeech 2017*, Stockholm, 2017, pp. 1089–1093.

[9] Z. Yang and J. Hirschberg, "Predicting Arousal and Valence from Waveforms and Spectrograms using Deep Neural Networks," in *Proc. Interspeech 2018*, Hyderabad, 2018, pp. 3092–3096.

[10] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding," in *Proc. Interspeech 2018*, Hyderabad, 2018, pp. 3688–3692.

[11] W. Bodusz, Z. Miodońska, and P. Badura, "Approach for spectrogram analysis in detection of selected pronunciation pathologies," in *Innovations in Biomedical Engineering*, M. Gzik, E. Tkacz, Z. Paszenda, and E. Pietka, Eds. Cham: Springer International Publishing, 2018, pp. 3–11.

[12] P. Verma and J. O. Smith, "Neural Style Transfer for Audio Spectrograms," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, 2017.

[13] L. Wyse, "Audio spectrogram representations for processing with Convolutional Neural Networks," in *First International Workshop on Deep Learning and Music joint with IJCNN*, Anchorage, 2017, pp. 37–41.

[14] S. Jenne, A. Schweitzer, S. Zerbian, and N. T. Vu, "Phonological Error Detection for Pronunciation Training Using Neural Spectrogram Recognition," in *International Congress of Phonetic Sciences*, Melbourne, 2019.

[15] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. New Jersey: Pearson Education International, 2009.

[16] N. Singh, R. Khan, and R. Shree, "MFCC and Prosodic Feature Extraction Techniques: A Comparative Study," *International Journal of Computer Applications*, vol. 54, no. 1, pp. 9–13, 2012.

[17] M. A. Hossan, S. Memon, and M. A. Gregory, "A Novel Approach for MFCC Feature Extraction," in *International Conference on Signal Processing and Communication Systems*, Gold Coast, 2010.

[18] T. J. Park and P. Georgiou, "Multimodal Speaker Segmentation and Diarization using Lexical and Acoustic Cues via Sequence to Sequence Neural Networks," in *Proc. Interspeech 2018*, Hyderabad, 2018, pp. 1373–1377.

[19] Y. Gu, S. Chen, and I. Marsic, "Deep Multimodal Learning for Emotion Recognition in Spoken Language," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, 2018, pp. 5079–5083.

[20] S. Tamura, M. Ishikawa, T. Hashiba, S. Takeuchi, and S. Hayamizu, "A Robust Audio-visual Speech Recognition Using Audio-visual Voice Activity Detection," in *Proc. Interspeech 2010*, Makuhari, Chiba, 2010, pp. 2694–2697.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324.

[22] A. Graves, A. Mohamed, and G. E. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 2013, pp. 6645–6649.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1," Web Download, 1993.

[24] S. Weinberger, "Speech Accent Archive," Online, George Mason University, 2015. [Online]. Available: http://accent.gmu.edu

[25] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation," in *International Conference on Speech and Computer*, Leipzig, 2018, pp. 198–208.

[26] K. Knopf, *English Pronunciation Exercises: Sprachlaborkurs nach didaktischen Schwerpunkten der Ausgangssprache Deutsch*, 2nd ed. München: Wilhelm Fink Verlag, 1975.

[27] M. Ratajczak, S. Tschiatschek, and F. Pernkopf, "Frame and Segment Level Recurrent Neural Networks for Phone Classification," in *Proc. Interspeech 2017*, Stockholm, 2017, pp. 1318–1322.

[28] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," Online, 2018. [Online]. Available: http://www.fon.hum.uva.nl/praat/

[29] A. Rosebrock, "Image classification with Keras and deep learning," Online, 2017. [Online]. Available: https://www.pyimagesearch.com/2017/12/11/image-classification-with-keras-and-deep-learning/

[30] J. Anderson-Hsieh, R. Johnson, and K. Koehler, "The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentals, Prosody, and Syllable Structure," *Language Learning*, vol. 42, no. 4, pp. 529–555, 1992.

[31] H. Quené and L. E. van Delft, "Non-native durational patterns decrease speech intelligibility," *Speech Communication*, no. 52, pp. 911–918, 2010.

[32] S. Zhao, S. N. Koh, S. I. Yann, and K. K. Luke, "Feedback Utterances for Computer-Adied Language Learning Using Accent Reduction and Voice Conversion Method," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 2013, pp. 8208–8212.