Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-024-09344-4

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Automatic Accent Identification Using Less Data: a Shift from Global to Segmental Accent

Justina Grigaliūnaitė[1] · **Gerda Ana Melnik-Leroy**[1]

Received: 27 March 2024 / Accepted: 8 July 2024

Abstract
Accentedness is a prominent feature of foreign language learning. While humans have a remarkable capacity to adapt their perception to accents, they remain a hard challenge to the robustness of automatic speech recognition (ASR). In particular, the necessity to use large non-native annotated datasets for model training remains a crucial issue. This paper investigates the possibility of reducing the data need of these systems by using targeted training datasets, focusing on the most challenging rather than on all non-native phonemes. Specifically, the study examines whether training data for ASR, and accent identification systems in particular, could focus not on global but on segmental accent. Segmental accent refers to the deviations in pronouncing specific phonemes, while global accent captures the extent to which a non-native speaker is perceived to differ from a native one. An accent identification problem was formulated, where models were trained on two types of data: full words (i.e., global accent) versus isolated difficult vowels (i.e., segmental accent), both uttered by natives and non-natives. Two novel highly controlled and professionally annotated datasets were used for that purpose. Throughout experiments, a transfer learning approach using pretrained deep residual neural networks was applied, with subsequent comparison to a baseline support vector machine. Results showed that although word-based classification yielded better accuracy, the dataset consisting of isolated vowels could account for much of the accent, when used with both methods (up to 80%). Applications of this approach and the possibility of using smaller, but more representative datasets are discussed.

Keywords Machine learning · Accent identification · Foreign accent · ResNet · ASR · Targeted datasets

1 Introduction

Accentedness is the most striking particularity of non-native speech that occurs due to the mismatch between the properties of the native language and the foreign one. These difficulties in foreign or second language (L2) pronunciation might stem from inaccurate perception and/or from articulatory constraints (for a review, see [1]). In both cases, foreign accent strongly impacts speech perception at a physiological level [2, 3] and this results in strong social consequences [4] and even changes in moral judgment [5]. In recent years, accentedness started impacting not only social interactions, but more and more human–computer interactions [6–8].

However, while humans have a remarkably fast capacity to perceptually adapt their processing to foreign-accented speech [3, 9], it remains a hard challenge to the robustness of ASR systems [7, 10, 11]. Specifically, automatic speech recognition (ASR) systems still demonstrate reduced effectiveness when assessing speech with foreign accents [12], even though they use large amounts of data for training. Moreover, as training such systems with native data only is ineffective, they require non-native annotated datasets, which are scarce, small and often difficult to access [13]. This issue becomes even more problematic when dealing with under-resourced languages [14].

In order to overcome the need for large training samples several solutions have been proposed. For instance, ASR systems can be complemented with preprocessing algorithms that identify the accent, which helps to tailor the recognition algorithm to the specific accent [12]. However, this solution does not alleviate the issue of training data scarcity. To mitigate this problem, data augmentation, such as speed modifications or noise addition, has been proposed [15, 16]. Multilingual training is another possible way of tackling the

✉ Gerda Ana Melnik-Leroy
gerda.melnik@mif.vu.lt

[1] Institute of Data Science and Digital Technologies, Vilnius University, Akademijos str. 4, LT-08412 Vilnius, Lithuania

Published online: 13 August 2024

issue by relying on universal phonetic structures across languages. In this case, the hidden layers in deep neural networks are jointly trained using data from multiple languages [17]. In addition to this, solutions based on transfer learning have been proposed [13], as well as methods using multi-task training and accent embedding [8].

In the current paper, we further investigate the issue of ASR for accented speech and the possibility to reduce the data need of these systems. We explore the possibility to use datasets that would be smaller, but more representative of the accent in question. Specifically, foreign accent refers not only to the global accent, but first and foremost to accent at the segmental level [18]. Global accent in this context refers to the degree to which the speech pronounced by a non-native speaker is perceived to differ from that of a native speaker, while segmental accent describes the deviations in pronouncing specific sounds [19]. Importantly, a global foreign accent does not imply that all phonemes are mispronounced. Rather, some of them are more difficult and bear most of the accent, which makes the non-native speech sound globally accented. The physiological and cognitive reasons for this phenomenon, as well as its importance will be discussed in the following section.

Based on this, we hypothesize that training data for ASR systems, and accent identification systems in particular should focus not on global, but on segmental accent, and thus, not on all non-native phonemes, but especially on the difficult accented phonemes. A similar proposal has recently been expressed in the neighboring field of ASR for dialectal variations [20]. The authors conclude that creating targeted training datasets focusing on linguistic features that emphasize the phonemic differences between a language and its dialect(s) might allow the use of smaller training sets.

In order to test to what extent the difficult sounds can be representative of a foreign accent, we formulated an accent identification problem, where the models were trained on two types of data: a dataset consisting of full words (dataset I) versus a dataset of isolated difficult vowels (dataset II). For this we used novel highly controlled and professionally annotated datasets containing both native and non-native pronunciations of full words (global accent) and isolated vowels (segmental accent). Importantly, we ensured that the items in the datasets were pronounced naturally, and thus with realistic accent, by using a picture naming task to elicit them [21]. We focused on a well-studied example of non-native vowel mispronunciation, namely the French contrast /u/ (as in "boule", *ball*) versus /y/ (as in "bulle", *bubble)*, which has been shown to be extremely difficult to perceive and produce for English speakers even at high levels of proficiency [22–24]. The items were produced by native and non-native speakers of French. The non-native group consisted of English natives proficient in French.

First, the study aimed at investigating whether the foreign accent can be detected when the algorithm is trained on whole words containing difficult vowels (example of global accent—dataset I) vs. on isolated difficult vowels (example of segmental accent—dataset II). In particular, in this 2-class classification problem, the model had to correctly classify accented and unaccented speech, either based on dataset I or dataset II.

The second objective of the study was to test the effectiveness of a deep learning method ResNet (which is a special case of a convolutional neural network (CNN) [25], compared to a baseline system built with a support vector machine (SVM). It is well known that deep convolutional neural networks (CNNs) perform well on image processing tasks due to their ability to capture various features in the image [26]. CNNs have already been shown to perform well on 2D speech data representations such as spectrograms [27, 28] in language identification problems [29, 30] and speech emotion recognition [31]. A recent study [32] investigated the efficiency of CNNs on the task of regional accent recognition, achieving high accuracy, compared to traditional methods such as i-vector-based linear discriminant analysis and support vector machines.

However, there have been only limited applications of CNNs to foreign-accented speech. The few existing examples used CNN for accent identification [33, 34], mispronunciation detection [35] or allophone classification [36]. In the current study we test for the first time the accuracy of a convolutional neural network (ResNet) versus the baseline SVM on the identification of a foreign accent with isolated words versus isolated difficult phonemes as input data.

ResNet has been chosen instead of a classical deep convolutional neural network in order to overcome the "vanishing gradient" problem [25]. Specifically, after a certain number of layers the classical CNN's performance often starts to decrease. This occurs as repeated multiplication of gradients during backpropagation can cause them to diminish significantly as they propagate backward through the layers. Thus, the network's performance can get saturated or even start degrading rapidly as it goes deeper. Residual neural networks have been proposed as a solution to this problem, as in this model layers are restructured to learn residual functions relative to the input of each layer, rather than learning unreferenced functions. He et al. [25] provide comprehensive empirical evidence showing that these networks have lower complexity, are easier to optimize and benefit considerably from the significantly increased depth. As a result, ResNet is considered as one of the top performing networks for computer vision tasks [29].

Based on the literature review provided earlier, the primary contributions of this study to the field of automatic speech recognition and accent identification can be summarized as follows:

- Unlike previous approaches that use extensive non-native datasets, this research demonstrates the efficacy of using smaller, targeted datasets focused on the most challenging and representative phonemes for training accent identification systems within ASR.
- We shift the focus from global to segmental accents, providing a detailed examination of how segment-specific training can effectively enhance model accuracy. The novel datasets used for the purpose were generated through a picture naming task, providing more naturalistic and spontaneous pronunciations compared to traditional reading tasks, thus minimizing orthographic interference.
- By utilizing ResNet, our study advances accent identification by overcoming the limitations of standard convolutional neural networks (CNNs), such as the vanishing gradient problem.
- The findings suggest practical applications in designing more efficient ASR systems that require less data, thereby lowering barriers to development, especially for under-resourced languages.

The remainder of this paper is organized as follows: Sect. 2 briefly introduces the physiological and cognitive causes of foreign accent; Sect. 3.1 introduces the proposed system, while Sect. 3.2 presents the datasets and describes their creation. Section 3.3 presents the audio data preprocessing that was carried out. The methodology of automatic accent identification (classification) is discussed in Sects. 3.3 and 3.4. Section 4 assesses the effectiveness of the proposed methods by analyzing experimental outcomes on both datasets. A discussion and concluding remarks are presented in Sects. 5 and 6, respectively.

2 What is a Foreign Accent and Where It Comes From

A great number of studies have shown that adult speakers face major difficulty when producing and perceiving speech sounds that are not used in their native language (for reviews, see [1, 37, 38]. This difficulty stems from the particularities of the language acquisition and processing mechanisms. In one of the early studies on foreign language acquisition, Trubetzkoy [39] hypothesized that our native language (L1) phonology acts as a "phonological sieve" and filters out those properties of the foreign language speech signal which are not relevant to the phonological system of the native language. This has been confirmed by evidence from L1 acquisition studies, showing that although newborns are sensitive to phonological contrasts of any language [40], speech perception becomes attuned to the contrastive sounds of the native language very early in life [41]. This attunement to one's maternal language might result in distorted perception of foreign sounds that differ from the L1 sounds in some phonological characteristics [42–45]. In foreign language pronunciation, the problems caused by the attunement to one's native language are even more evident than in perception. Namely, foreign accentedness is one of the most salient features that accompanies foreign language learning in adulthood.

The mismatches between the properties of the native language and the foreign one are "repaired" by the perceptual system. Three types of "repair" strategies have been attested: sound change, deletion and insertion [37, 46]. The difficulty in foreign language pronunciation can arise in two, not mutually exclusive, ways: first, some sounds can be perceived inaccurately, which can lead to wrong pronunciation. Second, the pronunciation of certain sounds requires using some articulators that are not used in the pronunciation of one's native sounds, thus resulting in motor difficulty [47–49]. The level of accentedness can also depend on the amounts of exposure and experience with the foreign language. For instance, the age of learning (AOL) refers to the age at which the learner was first exposed to the foreign language. The general assumption is that the earlier the AOL, the better the outcomes of learning are [50]. Finally, the degree of perceived foreign accent depends on a variety of other factors, such as formal instruction, motivation, language learning aptitude, amount of native language use and communicative pressure [1].

The example we focus on in the current paper, namely the French contrast /u/-/y/, is difficult for English learners, as in English only a single category /u/ exists. Moreover, acoustically it is not identical to the French /u/ (it is somewhere in between the two French sounds). Thus, both French /u/-/y/ are mapped to a single category in English natives, leading to poor perceptual discrimination and inaccurate pronunciation.

3 Methods

3.1 The Proposed System

The proposed system is designed to address the challenges of accent identification within automatic speech recognition systems, employing machine learning techniques and optimized data preprocessing methods. Figure 1 represents the proposed system and outlines the structure of our approach, beginning with the initial audio data acquisition, followed by systematic preprocessing steps that prepare the data for effective machine learning analysis. The subsequent descriptions will delve into each component in detail, explaining the functionalities and the integration within the overall system.

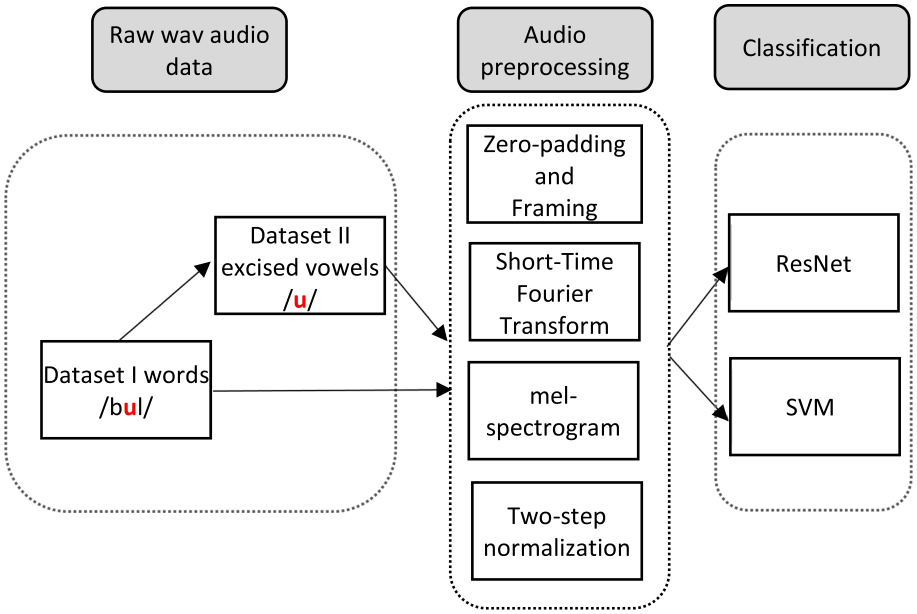

Fig. 1 Block diagram of the proposed approach

Table 1 Summary table describing both datasets

	Non-natives	Natives
Dataset I	1602	1136
Whole word samples containing /u/	870	588
Whole word samples containing /y/	732	548
Whole word max length	6.63 (s)	3.35 (s)
Whole word average length	1.13 (s)	0.84 (s)
Dataset II	1602	1136
/u/ samples	870	588
/y/ samples	732	548
Single vowel max length	0.65 (s)	0.38 (s)
Single vowel average length	0.12 (s)	0.09 (s)
Total count of samples	3204	2272

3.2 Materials

The experiments were carried out on two new datasets that have not been previously assessed using machine learning methods. A part of these datasets was recorded for a study in psycholinguistics by Melnik-Leroy et al. [24] The main characteristics of both datasets are presented in Table 1.

3.2.1 Dataset I

The dataset consisted of 2738 high-quality recordings of French words, pronounced by 39 non-native speakers of French (English natives) and 20 French native speakers. The participants were aged from 21 to 45 years. More participants were required in the non-native group, as their pronunciations were expected to be much more variable than those in the group of native speakers. All participants were living in Paris at the time of recording. The non-native speaker group consisted of native speakers of American or British English who had started to learn French at school. Speakers of only these two dialects were chosen, as some vowels vary across dialects and it is important to avoid this variability. The participants were students staying in France for at least one year and they all were medium-to proficient speakers of French. Native French speakers were recruited as control participants. None of the participants had hearing or language problems.

The data were obtained by using a picture naming task. The naming task is ecologically more suitable to obtain naturalistic pronunciations than reading tasks, as it is more spontaneous and prevents a possible interference of orthography in the performance [51]. 30 images that represent various nouns containing /u/ and 30 containing /y/ were chosen. All nouns were likely to be familiar to all participants. Pictures representing the nouns were presented one by one on the screen in a pseudo-random order, such that no more than three objects with the same target vowel in their name appeared in a row. Participants were asked to name the object they saw and to press a button to proceed to the next picture. All items were recorded in a controlled acoustic environment using a soundproof booth at 16 bits mono with a sampling rate of 44.1 kHz. The waveform and the wideband spectrogram of the pronunciation data were visualized by professional phoneticians using the software *Praat* (version 6.0.17) and an annotation text was added for each audio file.

As some of the recordings were not suitable for the experiment due to technical problems (they were too short or

contained noise), the final dataset consisted of 1458 words containing the vowel /u/ and 1280 containing the vowel /y/. The ratio between French and American speakers' recordings was approximately 2:3 (1136 French and 1602 American); therefore, the final dataset was slightly imbalanced.

3.2.2 Dataset II

This dataset contained recordings of vowels /u/ and /y/, excised from all the recordings of French words from dataset I. Thus, dataset II contained the same number of items as dataset I. The waveform and the wideband spectrogram of the pronunciation data were visualized by professional phoneticians using the software *Praat* (version 6.0.17), and an annotation text was added for each audio file to segment and label the target vowels. To manually segment the vowels, boundaries were set at zero crossings.

3.3 Audio Data Preprocessing

The recordings were in waveform audio file format (WAV) and needed to be preprocessed and converted to a format suitable for the selected models. Since ResNet is based on CNN architecture which is fitted for image data, the speech signals needed to be represented as two-dimensional images. For that purpose, mel-spectrograms were used.

3.3.1 Fourier Transform and Mel-spectrograms: Theoretical Background

The Fourier transform allows to decompose a signal into its individual frequencies and the frequency's amplitude. It converts the signal from the time domain into the frequency domain. The result is called a spectrum. For digital (and discrete) applications it is possible to use fast Fourier transform (FFT) that can efficiently compute the discrete Fourier transform (DFT). To capture the dynamic nature of speech signals over time, short-time Fourier transform (STFT) is commonly employed [52]. This involves computing the Fourier transform on multiple overlapping windowed segments of the signal, thereby generating a spectrum representation that evolves with time, called a spectrogram (Fig. 2). It, essentially, is a heat-map where the X-axis represents the time, Y-axis represents the frequency, and the color/shade determines the amplitude of the sound. It is a way to visually represent a signal's loudness or amplitude, as it varies over time at different frequencies.

As auditory perception in humans is rather logarithmic than linear (we are more accurate at detecting differences in lower frequencies than in higher ones), the use of a psychoacoustical scale such as the mel scale is preferable [53]. Hence, a mel-spectrogram is a type of spectrogram in which the frequencies are transformed into the mel-scale (Fig. 3).

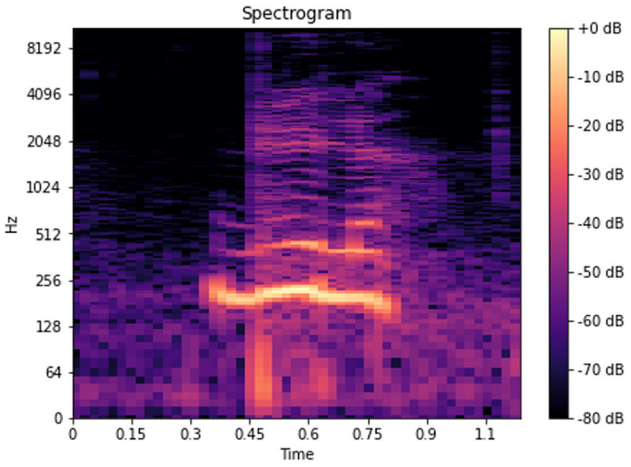

Fig. 2 Example of a spectrogram (the French word "bulle")

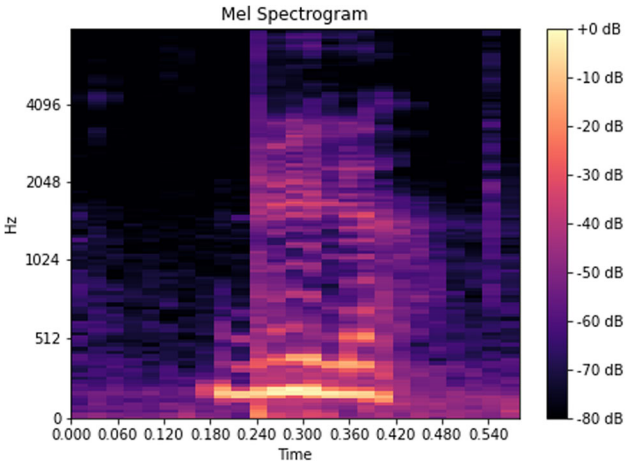

Fig. 3 Example of a mel-spectrogram (the French word "bulle")

3.3.2 Practical Application

To convert WAV files to 2D mel-spectrograms, the Python library *librosa* was used [54]. The first step involved loading the audio files into variables each of which was an array of floating-point time series, which also contained the sampling rate information (in this case, 44,100 Hz). Since all of the recordings had different duration, the audio data had to be processed to have the same length. This was achieved by padding the arrays from both sides with a constant value of 0 (i.e., zero-padding). The fixed length of 2.4 (s) for whole words (dataset I) and 0.3 (s) for isolated vowels (dataset II) were chosen, based on the approximated value at 99th quantile. The recordings that were longer than the specified fixed value were trimmed (from the right). The resulting arrays were used to calculate spectrograms (which operate on a power spectrum) using STFT.

When computing an STFT, a number of hyperparameters had to be selected. First, various types of framing, indicating

the length of the segments on which FFT is calculated, were checked and 2048 was selected. Usually, these frames overlap (to avoid information loss) and the number of samples in between successive frames is represented by the parameter hop length which should be less than frame length (set to default 512 in our case). The number of mel bands to generate was set to 128. After obtaining spectrograms (amplitude squared), we converted them to decibel (dB) units to obtain a logarithmic scale, which corresponds more closely to human auditory perception.

After the spectrogram calculation, two-step normalization was applied: Z-normalization and value scaling. Z-normalization normalizes every value in a dataset such that the mean of all values is 0 and the standard deviation is 1 by applying the formula: $x_{\text{new}} = (x - \mu)/\sigma$. Then, the spectrogram scaling maps the values to the range between 0 and 1. It is known that CNN models perform well on data in this range and the performance is improved by applying the neural network's calculations on small numbers. Scaling was done on each column of the 2D arrays by aggregating over the rows. This highlighted the energy peaks over time.

After normalization, an additional dimension, which corresponds to a color channel, was appended to the data. Then, arrays were converted to tensors and the audio data were combined with the appropriate label (0—native and 1—non-native).

Finally, the data in each dataset were split into 80% train, 10% validation and 10% test, maintaining class ratio in each split. The validation dataset was used to track and evaluate the model's performance during training and to notice if the model starts to overfit the training data. Test data were used at the end of the training cycle to obtain the final performance metrics. Data were split in this way 5 different times and re-shuffled. Such fivefold split was used to apply model cross-validation and to better evaluate the performance of the suggested approach.

3.4 Deep Learning Method: Residual Neural Network

In this study, we employed a cross-domain transfer learning approach to address the challenge of phoneme-based accent identification, by taking advantage of the architectural strengths of a pretrained residual networks (ResNet) model.

3.4.1 Original ResNet

ResNet was developed by Kaiming He et al. in 2016 to facilitate the training of much deeper neural networks than was previously feasible. The core concept behind ResNet is the introduction of "residual learning" blocks to train deeper networks effectively. It addresses the vanishing gradient problem by introducing skip connections that allow layers to learn residual functions with respect to the input.

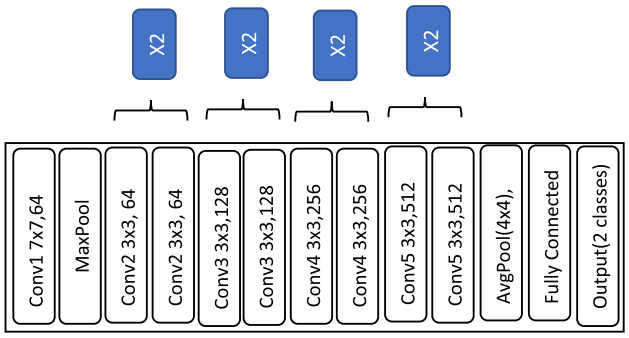

Fig. 4 The proposed ResNet architecture

In ResNet, the layers are designed to learn residual functions with respect to the input, rather than learning the functions directly. This is mathematically represented as:

$$H(x) = F(x) + x$$

where $H(x)$ is the desired output, x is the input to the residual block, and $F(x)$ is the residual mapping that the layers need to learn. The residual function is specifically given by:

$$F(x) = H(x) - x$$

This formulation enables identity mapping as the default when $F(x)$ approaches zero, facilitating easier optimization and deeper network training without degradation in performance.

3.4.2 The Revised Model

We used an 18-layer ResNet model, pretrained on the ImageNet dataset, provided by the *torchvision* package, a component of the *PyTorch* framework [55]. In order to fine-tune the original model to the task of accent identification, it was adapted to process spectrogram images of audio signals (the proposed architecture is presented in Fig. 4). This involved adjusting the network's first convolutional layer to accept single-channel input, representing mel-spectrograms, instead of its original design for three-channel RGB images. A 7×7 kernel, a stride of 2, and 3×3 padding were used to align with the original implementation by He et al. [25]. The operations included batch normalization and max pooling.

Then, it was followed by a sequence of residual blocks, each containing two convolutional layers with 3×3 kernels, followed by batch normalization. ReLU (rectified linear units) were used as activations. These layers did not differ from the architecture of the original ResNet-18. These blocks are designed to facilitate the learning of residual

Table 2 Model performance with different batch sizes

Batch size	Epoch	Train loss	Train accuracy	Val loss	Val accuracy
4	10	55	72	48	77
8	10	47	78	56	77
16	10	40	81	48	75
32	10	42	80	72	71
64	10	31	86	58	78

Table 3 Model performance with different learning rates

Learning rate	Epoch	Train loss	Train accuracy	Val loss	Val accuracy
0.0001	10	19	92	57	81
0.001	10	41	81	52	75
0.01	10	48	76	56	74
0.1	10	55	71	60	74

functions with reference to the layer inputs, incorporating shortcut connections that bypass one or more layers. Following the convolutional stages, the network ended with a global average pooling layer and a fully connected layer that outputs to a classification mechanism. For the purpose of binary classification—distinguishing between native and non-native accents—the final layer's output was configured to two neurons, corresponding to the two accent categories (native vs. non-native). The network's output was then passed through a softmax function as part of the cross-entropy loss calculation during training, which implicitly handles the conversion of logits to probabilities for class prediction. Note that cross-entropy loss function is widely used for classification problems and is particularly useful for unbalanced training sets [56]. The Adam optimizer was utilized to adaptively update the weights of the ResNet model, enhancing the learning rate adjustment process.

To improve the model's performance, several hyperparameters were evaluated: different batch sizes, learning rates and number of epochs. In each iteration, only one hyperparameter was changed to obtain comparable results.

Batch size: we experimented with various batch sizes to understand their impact on model performance. As summarized in Table 2, the evaluation criteria included training loss, training accuracy, validation loss, and validation accuracy across 10 epochs for each batch size setting. Higher numbers than 64 were not considered due to the relatively small dataset sizes. Our analysis indicated that a batch size of 16 offered the most favorable trade-off, achieving comparatively lower loss and higher accuracy without significant discrepancies between training and validation metrics.

Learning rate: the learning rate's influence was assessed by fixing the batch size at 16 and varying the learning rate across several magnitudes (Table 3). The chosen rates were evaluated based on the same performance metrics as the batch size experiment. Table 3 shows the performance results with different learning rates (with batch size 16). The lowest loss in both sets was produced when the learning rate was equal to 0.001. It also produced more stable training and validation convergence; therefore, this learning rate was selected for the final model.

Epochs: An evaluation up to 30 epochs (with batch size 16 and learning rate 0.001) revealed that the model began to overfit past the 15th epoch, as evidenced by diverging loss and accuracy curves in both the training and validation sets (see Fig. 5). To mitigate overfitting, we implemented an early stopping mechanism, selecting 15 epochs as the optimal training duration to balance learning and generalizability.

To ensure a rigorous evaluation and generalizability of the proposed approach, a fivefold cross-validation technique was used. This process involved dividing the dataset into five unique subsets for training and evaluation and iteratively training and validating the model on these subsets. Importantly, in every fold, a distinct test data subset was reserved, ensuring no overlap with the training data. After each fold, metrics for the obtained model were calculated.

3.5 Baseline Method: Support Vector Machine

The support vector machine (SVM) is a supervised learning technique which is a classical, relatively simple and popular method capable of solving problems in classification, regression and outlier detection tasks [26]. The principle behind SVM classification is as follows: the algorithm constructs a hyperplane or a set of hyperplanes in a high or infinite dimensional space which separates the data into classes [57].

In SVM, the goal is to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data point from each class. By maximizing

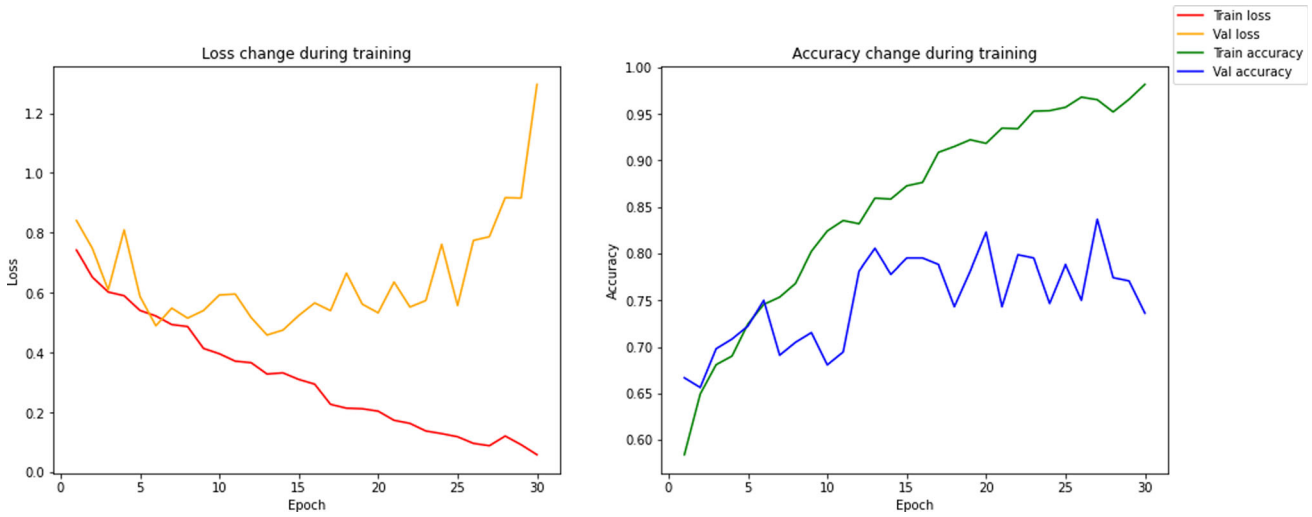

Fig. 5 Loss and accuracy change over epochs

		Actual values	
		Positive	Negative
Predicted value	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Fig. 6 The confusion matrix

this margin, SVM aims to achieve the best possible separation between the classes in the feature space. The larger the margin—the lower the generalization error of the classifier. Data samples on the margin boundaries are called "support vectors".

3.6 Performance Metrics

In this paper, we employed classical performance measures commonly utilized in machine learning classification tasks, namely accuracy, recall, precision and F1. Moreover, we used the ROC-AUC as an additional metric for imbalanced datasets, in order to mitigate potential biases that could arise from relying solely on accuracy as a performance metric. These metrics are calculated based on the counts of true positive, true negative, false positive, and false negative predictions, which are represented in the confusion matrix below (Fig. 6).

Accuracy describes the proportion of correct predictions, both TP and TN, out of all observations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Recall (or sensitivity or true positive rate) describes the proportion of true positive cases that are correctly identified out of all actual positive cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision, or confidence, describes the proportion of true positive cases that are correctly identified out of all predicted positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The F1-score is the harmonic average of the precision and recall. It is especially suitable for imbalanced datasets, as it equally considers false positives (which impact precision) and false negatives (which impact recall), making it a robust choice for assessing model performance in skewed class distributions.

$$F1\text{Score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Finally, the ROC (receiver operating characteristic) analysis gives geometric insights into the nature of the measures and their sensitivity to skew. ROC is created by plotting the fraction of true positive rate (TPR—the proportion of true positive cases correctly identified out of all actual positive cases) on the Y-axis and the true negative rate (TNR—the proportion of true negative cases correctly identified out of

all actual negative cases) on the X-axis. It is a probability curve that allows to separate the "signal" from the "noise". The ROC-AUC metric represents the area under the ROC curve, and it serves as a measure of a classifier's ability to distinguish between classes. This metric provides a summary of the ROC curve's performance, with higher values indicating better discrimination between positive and negative classes.

4 Experiments

4.1 Experiment 1: Word-Based Classification with ResNet

The goal of the word-based classification was to classify words containing sounds /u/ and /y/ as pertaining to non-native accented (pronounced by non-native speakers) versus native (pronounced by French natives) speech. This experiment was thus assessing classification based on global accent. Table 4 shows accuracy and other metrics' changes over epochs.

From Table 4, we can see that all 5 models reached an average of 90% accuracy on test (unseen) data at the end of 15th epoch. As the dataset was slightly imbalanced, the F1 and ROC-AUC score were also calculated to ensure the precision of the evaluation. The results showed that these metrics provide similar values to accuracy (F1 = 89%; ROC-AUC = 90%). For the visual representation of the ROC curve, see Fig. 7a.

4.2 Experiment 2: Vowel-Based Classification with ResNet

In the second experiment, the classification was based on isolated vowels /u/ and /y/. This experiment was thus assessing classification based on segmental accent. It was implemented using the same preprocessing techniques and using the same pretrained 18-layer ResNet model as in the first experiment. Here too, there were 5 separate models generated using cross-validation and 80% train, 10% validation and 10% test sets.

Table 5 shows that all five models reached an average of 78% accuracy on test (unseen) data at the end of the 15th epoch. In vowel-based classification, the alternative metrics of F1 and ROC-AUC were, again, comparable to accuracy (F1 = 81%; ROC-AUC = 76%). For the visual representation of the ROC curve, see Fig. 7b.

Unsurprisingly, the accuracy in accent classification based on isolated vowels decreased compared to word-based classification. Importantly, the accuracy remained fairly high (the difference between the two datasets was only around 10%) although each sample of isolated vowels in experiment 2 contained much less information than the full words in experiment 1. Specifically, each word in experiment 1 contained either the target vowel /u/ or /y/, but also other sounds that were potentially pronounced differently by native and non-native speakers. Thus, words contained much more information, and the model could learn more distinct features. The fact that isolated vowels yielded comparable levels of accuracy in classification to full words points to the fact that the target vowels /u/ and /y/ accounted for disproportionately much of the accent.

Overall, the accuracy of this model can be considered comparable to other studies addressing similar problems on non-native speech data, as their accuracy typically hovers around 80%, as can be seen from Table 6, comparing existing approaches with the current one.

4.3 Experiment 3: Comparison of ResNet Performance with the Baseline SVM

Since SVM takes as input only data in 1D format, the 2D spectrograms were converted into 1D arrays. Those arrays could be fed to the SVM model together with class labels. SVM classification was implemented using Python library *scikit-learn* [58]. The same datasets were used as for the ResNet implementation. However, they were split into 80% training and 20% sets. The validation set was omitted for simplicity reasons because we did not intend to implement optimal parameter search and only used this model as a baseline [59].

Table 7 shows that the SVM model reached 73% accuracy for classification based on single vowels and 79% accuracy for classification based on whole words. These results are much lower than those obtained using ResNet. Note that ResNet reached similar accuracy levels on isolated vowel classification as SVM did on full words.

5 Discussion

This study aimed at investigating to what extent difficult sounds can be representative of a foreign accent and whether this accent can be detected when the algorithm is trained on isolated difficult vowels (segmental accent) versus whole words containing these vowels (global accent). For this purpose, we formulated an accent identification problem, where the models were trained on two types of data from a novel professionally annotated dataset: full words (dataset I) versus isolated difficult vowels (dataset II). The results showed that although word-based classification yielded better accuracy (Accuracy = 90%; F1 = 89%); vowel-based classification still showed fairly high performance (accuracy = 78%; F1 = 81%). Crucially, this ~ 10% difference in the results is much smaller than the difference in the acoustic information contained in words vs. isolated difficult vowels. In particular, full words contained many more sounds and thus acoustic

Table 4 Cross validation results for word classification (dataset I) (in %)

Fold	Accuracy	Recall	Precision	F1 score	ROC-AUC
1	92.36	93.01	93.27	92.73	92.55
2	87.5	73.13	94.44	81.93	86.57
3	93.06	93.96	94.3	93.93	92.99
4	92.71	88.62	97.58	92.16	93.37
5	86.81	98.57	81.8	88.88	83.03
Average	90.49	89.46	92.28	89.93	89.7

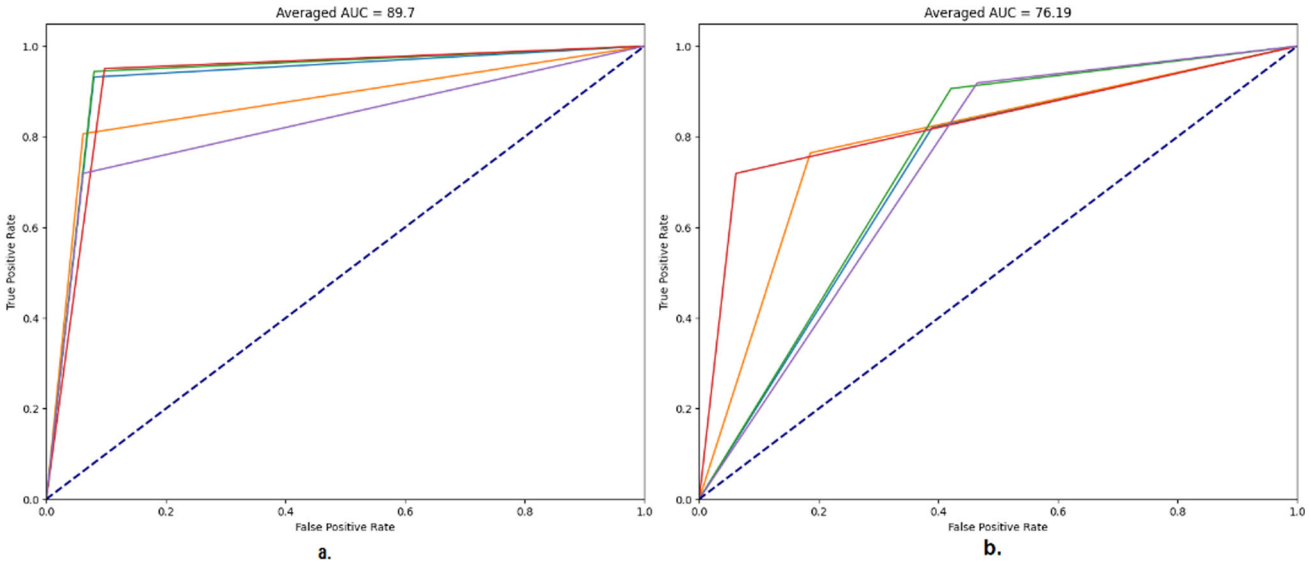

Fig. 7 ROC curve for word (**a**) and vowel (**b**) classification

Table 5 Cross-validation results for vowel classification (in %)

Fold	Accuracy	Recall	Precision	F1 score	ROC-AUC
1	75	88.76	73	79.14	72.38
2	81.25	94.85	77.76	84.66	79
3	78.13	81.51	78.54	79.56	73.97
4	81.6	80.02	87.46	82.52	82.78
5	75.69	90.95	73.86	80.03	72.8
Average	78.33	87.22	78.13	81.18	76.19

features than the isolated vowels. However, the isolated vowels alone could account for much of the accent as the model trained on dataset II yielded around 80% accuracy (and F1) in the identification of the accent. This is in line with several papers in psycholinguistics, which found significant correlations between global and segmental foreign accent [19, 60], using subjective impressionistic judgments.

Based on these results, we suggest that ASR systems and accent identification systems in particular could strongly benefit from datasets focusing specifically on difficult non-native phonemes. In particular, the problem of dataset scarcity that impedes the development of ASR systems could at least partly be mitigated by creating datasets which would be representative of the most difficult sounds for speakers of a particular foreign language (as these sounds are likely to be the most mispronounced). That is, instead of training the model on all sounds of the language, it should be possible to target a few crucial ones. In this way, much time and resources could be saved, as the datasets could be much smaller and require less effortful annotation. A similar approach has been successfully explored by [61] who focused on critical phonemes in regional dialect classification in Arabic.

Table 6 Comparison with existing approaches

Paper	Dataset size	Classification technique	Input features	Baseline	Accuracy	F1
[12]	56,452 samples	DNN + RNN	MFCCs	SVM	50.2	–
[11]	20 h of recordings	Transformer- based hybrid CTC/attention architecture	80-D FBanks and 3-D pitches	Transformer-12L ResNet E2 + ASR-init	72.4	–
[16]	16,422 words	Hybrid CNN-LSTM	Mel-spectrograms	CNN, LSTM, Bi-LSTM, GRU	85.8	85.8
[34]	20 h of recordings	ASR MTL	40-D FBank spectrum	Transformer	82.2	–
[33]	39,000 samples	CNN	Mel-spectrograms	ANN-RNN fusion	78.5	77
[10]	20 h of recordings	Persistent accent memory method	WavLM SSLRs	Model trained on Fbank) features	81.4	–
Current	2738 words	ResNet	Mel-spectrograms	SVM	90.5	89.9
	2738 vowels				78.3	81.2

Table 7 SVM-based classification metrics, compared to ResNet classification (in %)

		Accuracy	Recall	Precision	F1 score	ROC-AUC
ResNet	Word	90.49	89.46	92.28	89.93	89.7
	Vowel	78.33	87.22	78.13	81.18	76.19
SVM	Word	79.38	83.99	80.06	81.98	78.77
	Vowel	73.36	76.92	77.88	77.40	72.54

Note that in our experiments the classification based on a single difficult contrast /u/-/y/ yielded around 80% accuracy. We expect that adding one or several difficult sounds to the dataset would allow to reach even higher accuracy. Future studies should examine how many sounds on average are necessary to account for a satisfactory level of the accent, and what proportion of it should still be covered by recordings of full words and sentences. Importantly, the proposed approach operates with the restriction that different languages and different accents will require different datasets of difficult sounds. Thus, it is possible that two difficult sounds will be sufficient to identify an English accent in French, but four will be required for another pair of languages. This suggests that an optimal number of difficult sounds should be identified in each case separately. However, the identification of these difficult sounds baring crucial information on the accent can be done not only by psycholinguists, phoneticians, but also by second language teachers. In fact, for many large languages, such as English or French, these sounds have already been to a large extent identified and much studied [62–66]. For other, rather under-resourced languages, such research would benefit both the linguistic and teaching communities, as well as speech technologies.

In our analysis, we employed fivefold cross-validation to evaluate the generalizability of our accent identification models across data subsets. Despite these precautions, we observed some variations in model performance across folds, which may stem from the inherent complexity of foreign-accent variations in the speech data and potential underrepresentation of certain accent characteristics within training subsets. Specifically, the target vowels /u/ and /y/ are well-studied examples of difficult sounds for English speakers, as they lack of corresponding sounds in their native phonological inventory [23, 24, 67]. This leads to systematic pronunciation errors that can vary significantly even among speakers with similar language proficiency levels [68]. Note, that variability in pronunciation among non-native speakers is typically more pronounced than in regional accents, largely because foreign accent involves not only different phonetic realizations influenced by the speaker's native language phonology but also a wider range of prosodic and articulatory deviations due to less exposure and familiarity with the target language [69]. Additionally, the cognitive and perceptual challenges faced by non-native speakers in mastering a second language's phonetics often lead to more inconsistent and varied pronunciation patterns compared to regional variations, which are generally more subtle and consistent within a given language community.

Enhancing the model's performance could involve refining feature engineering techniques that focus on the spectral qualities specific to /u/ and /y/. Features like formant frequencies, which are critical in distinguishing these vowels, could be more heavily emphasized during the model training phase. Additionally, although the Adam optimizer was used for adaptive learning, exploring additional adaptive techniques like AdaBound or Lookahead could potentially enhance training stability and convergence rates. Moreover, while batch normalization is already implemented, integrating further regularization methods such as spatial or variational dropout might improve the model's generalization capabilities by providing more robust control over overfitting. Additionally, incorporating L2 regularization could further mitigate overfitting by penalizing excessive weight magnitudes, thereby maintaining model simplicity.

Finally, we found that higher accuracy can be reached using a pretrained ResNet compared to the baseline SVM method. Importantly, when the baseline method was used, the difference between word-based and vowel-based classification remained approximately the same (around 10%), bringing supporting evidence that the segmental accent bares crucial acoustic information that can be used to classify accented vs. non-accented speech.

6 Conclusion

To sum up, the results of the present study provide promising prospects for the future development of novel accented speech datasets. Such datasets could target difficult sounds of a foreign language, which bare most of the segmental accent. This would help to at least partly overcome the data scarcity issue inherent to ASR for accented speech. Further research could include classifying more than two accent groups, focusing on different sounds or including different languages.

Acknowledgements We would like to kindly thank Dr. Sharon Peperkamp for allowing us to use the data collected during a prior collaboration.

Declarations

Conflict of interest The authors report no conflict of interests.

References

1. Piske, T.; MacKay, I.R.A.; Flege, J.E.: Factors affecting degree of foreign accent in an L2: a review. J. Phon. **29**, 191–215 (2001). https://doi.org/10.1006/jpho.2001.0134
2. Foucart, A.; Santamaría-García, H.; Hartsuiker, R.J.: Short exposure to a foreign accent impacts subsequent cognitive processes. Neuropsychologia **129**, 1–9 (2019). https://doi.org/10.1016/j.neuropsychologia.2019.02.021
3. Romero-Rivas, C.; Martin, C.D.; Costa, A.: Processing changes when listening to foreign-accented speech. Front. Hum. Neurosci. **9**, 1–15 (2015). https://doi.org/10.3389/fnhum.2015.00167
4. Lev-Ari, S.; Keysar, B.: Why don't we believe non-native speakers? The influence of accent on credibility. J. Exp. Soc. Psychol. **46**, 1093–1096 (2010). https://doi.org/10.1016/j.jesp.2010.05.025
5. Foucart, A.; Brouwer, S.: Is there a foreign accent effect on moral judgment? Brain Sci. **11**, 1–11 (2021). https://doi.org/10.3390/brainsci11121631
6. Moussalli, S.; Cardoso, W.: Intelligent personal assistants: can they understand and be understood by accented L2 learners? Comput. Assist. Lang. Learn. **33**, 865–890 (2020). https://doi.org/10.1080/09588221.2019.1595664
7. Shi, X.; Yu, F.; Lu, Y.; Liang, Y.; Feng, Q.; Wang, D.; Qian, Y.; Xie, L.: The accented English speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 6918–6922 (IEEE) (2021). https://doi.org/10.1109/ICASSP39728.2021.9413386
8. Viglino, T.; Motlicek, P.; Cernak, M.: End-to-end accented speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2019).https://doi.org/10.21437/Interspeech.2019-2122.
9. Callan, D.; Callan, A.; Jones, J.A.: Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. Front. Neurosci. **8**, 1–15 (2014). https://doi.org/10.3389/fnins.2014.00275
10. Li, R.; Xie, Z.; Xu, H.; Peng, Y.; Liu, H.; Huang, H.; Chng, E. S.: Self-supervised Learning Representation based Accent Recognition with Persistent Accent Memory. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, August 2023, pp. 1968–1972. (2023) https://doi.org/10.21437/Interspeech.2023-1702.
11. Gao, Q.; Wu, H.; Sun, Y.; Duan, Y.: An end-to-end speech accent recognition method based on hybrid CTC/attention transformer ASR. In: Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing, June 2021, pp. 7253–7257 (2021) https://doi.org/10.1109/ICASSP39728.2021.9414082.
12. Jiao, Y.; Tu, M.; Berisha, V.; Liss, J.: Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 08–12-Sept, 2388–2392 (2016) https://doi.org/10.21437/Interspeech.2016-1148.
13. Sancinetti, M.; Vidal, J.; Bonomi, C.; Ferrer, L. A: Transfer learning approach for pronunciation scoring. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), pp. 6812–6816 (2022). https://doi.org/10.1109/ICASSP43922.2022.9747727 https://doi.org/10.1109/ICASSP43922.2022.9747727.
14. Melnik-Leroy, G.A.; Bernataviˇienė, J.; Korvel, G.; Navickas, G.; Tamulevičius, G.; Treigys, P.: An overview of lithuanian intonation: a linguistic and modelling perspective. Informatica (2022). https://doi.org/10.15388/22-INFOR502
15. Fukuda, T.; Fernandez, R.; Rosenberg, A.; Thomas, S.; Ramabhadran, B.; Sorin, A.; Kurata, G.: Data augmentation improves recognition of foreign accented speech. In: In: Proceedings of the Annual Conference of the International Speech Communication Association INTERSPEECH, September 2018, pp. 2409–2413 (2018) https://doi.org/10.21437/Interspeech.2018-1211.
16. Wubet, Y.A.; Balram, D.; Lian, K.Y.: Intra-native accent shared features for improving neural network-based accent classification

and accent similarity evaluation. IEEE Access **11**, 32176–32186 (2023). https://doi.org/10.1109/ACCESS.2023.3259901
17. Tong, S.; Garner, P.N.; Bourlard, H.: Cross-lingual adaptation of a CTC-based multilingual acoustic model. Speech Commun. **104**, 39–46 (2018). https://doi.org/10.1016/j.specom.2018.09.001
18. Riney, T.J.; Flege, J.E.: Changes over time in global foreign accent and liquid identifiability and accuracy. Stud. Second. Lang. Acquis. **20**, 213–243 (1998). https://doi.org/10.1017/s0272263198002058
19. Riney, T.J.; Takada, M.; Ota, M.: Segmentals and global foreign accent: The Japanese flap in EFL. TESOL Q. **34**, 711 (2000). https://doi.org/10.2307/3587782
20. Szalay, T.; Shahin, M.; Ahmed, B.; Ballard, K.: Knowledge of accent differences can be used to predict speech recognition. In: Annual Conference of the International Speech Communication Association, vol. 64, pp. 1372–1376 (2022) https://doi.org/10.21437/interspeech.2022-10162.
21. Reynolds, M.G.; Schlöffel, S.; Peressotti, F.: Asymmetric switch costs in numeral naming and number word reading: implications for models of bilingual language production. Front. Psychol. **6**, 1–15 (2016). https://doi.org/10.3389/fpsyg.2015.02011
22. Levy, E.S.: Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. J. Acoust. Soc. Am. **125**, 1138–1152 (2009). https://doi.org/10.1121/1.3050256
23. Levy, E.S.; Law, F.F.: Production of French vowels by American-English learners of French: language experience, consonantal context, and the perception-production relationship. J. Acoust. Soc. Am. **128**, 1290–1305 (2010). https://doi.org/10.1121/1.3466879
24. Melnik-Leroy, G.A.; Turnbull, R.; Peperkamp, S.: On the relationship between perception and production of L2 sounds: evidence from Anglophones' processing of the French /u/–/y/ contrast. Second. Lang. Res. **38**, 581–605 (2022). https://doi.org/10.1177/0267658320988061
25. He, K.; Zhang, X.; Ren, S.; Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), vol. 45, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90.
26. Géron, A.: Hands-on machine learning with Scikit-learn and TensorFlow. O'Reilly Media (2017)
27. Das, H. S.; Roy, P.: A deep dive into deep learning techniques for solving spoken language identification problems. In: Intelligent Speech Signal Processing, Elsevier Inc., (2019). https://doi.org/10.1016/B978-0-12-818130-0.00005-2.
28. Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B.: A study of cross-linguistic speech emotion recognition based on 2d feature spaces. Electron. **9**, 1–13 (2020). https://doi.org/10.3390/electronics9101725
29. Bartz, C.; Herold, T.; Yang, H.; Meinel, C.: Language identification using deep convolutional recurrent neural networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10639 LNCS, pp. 880–889. (2017). https://doi.org/10.1007/978-3-319-70136-3_93.
30. Mukherjee, H.; Ghosh, S.; Sen, S.; SkMd, O.; Santosh, K.C.; Phadikar, S.; Roy, K.: Deep learning for spoken language identification: Can we visualize speech signal patterns? Neural Comput. Appl. **31**, 8483–8501 (2019). https://doi.org/10.1007/s00521-019-04468-3
31. Kakuba, S.; Poulose, A.; Han, D.S.: Deep learning approaches for bimodal speech emotion recognition: advancements, challenges, and a multi-learning model. IEEE Access **11**, 113769–113789 (2023). https://doi.org/10.1109/ACCESS.2023.3325037
32. Cetin, O.: Accent recognition using a spectrogram image feature-based convolutional neural network. Arab. J. Sci. Eng. **48**, 1973–1990 (2023). https://doi.org/10.1007/s13369-022-07086-9
33. Singh, U.; Gupta, A.; Bisharad, D.; Arif, W.: Foreign accent classification using deep neural nets. J. Intell. Fuzzy Syst. **38**, 6347–6352 (2020). https://doi.org/10.3233/JIFS-179715
34. Zhang, Z.; Wang, Y.; Yang, J.: Accent recognition with hybrid phonetic features. Sensors (2021). https://doi.org/10.3390/s21186258
35. Chu, W.; Liu, Y.; Zhou, J.: Recognize mispronunciations to improve non-native acoustic modeling through a phone decoder built from one edit distance finite state automaton. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, October 2020, pp. 3062–3066 (2020) https://doi.org/10.21437/Interspeech.2020-3109.
36. Piotrowska, M.; Czyżewski, A.; Ciszewski, T.; Korvel, G.; Kurowski, A.; Kostek, B.: Evaluation of aspiration problems in L2 English pronunciation employing machine learning. J. Acoust. Soc. Am. **150**, 120–132 (2021). https://doi.org/10.1121/10.0005480
37. Sebastián-gallés, N.; Baus, C.: On the relationship between perception and production in L2 categories. In: Cutler, A (ed.) Twenty-first Century Psycholinguistics: Four cornerstones, pp. 279–292, Erlbaum, New York (2005)
38. Dufour, S.; Nguyen, N. L: Influence de la langue maternelle sur les capacités de l'auditeur dans la perception de la parole. Travaux interdisciplinaires du Laboratoire Parole et langage d'Aix-en-Provence, pp. 38–49. (2008)
39. Trubetzkoy, N.S.: Principles of Phonology. University of California Press, Berkeley (1969)
40. Houston, D. M.: Speech perception in Infants. In: The Handbook of Speech Perception, pp. 416–448. Blackwell Publishing Ltd, Oxford (2008). https://doi.org/10.1002/9780470757024.ch17
41. Kuhl, P.K.; Stevens, E.; Hayashi, A.; Deguchi, T.; Kiritani, S.; Iverson, P.: Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. Dev. Sci. (2006). https://doi.org/10.1111/j.1467-7687.2006.00468.x
42. Darcy, I.; Daidone, D.; Chisato, K.: Asymmetric lexical access and fuzzy lexical representations in second language learners. Ment. Lex. **8**, 372–420 (2013). https://doi.org/10.1075/ml.8.3.06dar
43. Díaz, B.; Mitterer, H.; Broersma, M.; Sebastián-Gallés, N.: Individual differences in late bilinguals' L2 phonological processes: from acoustic-phonetic analysis to lexical access. Learn. Individ. Differ. **22**, 680–689 (2012). https://doi.org/10.1016/j.lindif.2012.05.005
44. Melnik, G.A.; Peperkamp, S.: Perceptual deletion and asymmetric lexical access in second language learners. J. Acoust. Soc. Am. **145**, EL13–EL18 (2019). https://doi.org/10.1121/1.5085648
45. Melnik, G.A.; Peperkamp, S.: High-variability phonetic training enhances second language lexical processing: evidence from online training of French learners of English. Biling. Lang. Cogn. **24**, 497–506 (2021). https://doi.org/10.1017/S1366728920000644
46. Davidson, L.; Shaw, J.A.: Sources of illusion in consonant cluster perception. J. Phon. **40**, 234–248 (2012). https://doi.org/10.1016/j.wocn.2011.11.005
47. Masuda, H.; Arai, T.: Perception and production of consonant clusters in Japanese-English bilingual and Japanese monolingual speakers. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 1988–1991 (2008)
48. Schmitz, J.; Díaz, B.; Fernández Rubio, K.; Sebastian-Galles, N.: Exploring the relationship between speech perception and production across phonological processes, language familiarity, and sensory modalities. Lang. Cogn. Neurosci. **33**, 527–546 (2018). https://doi.org/10.1080/23273798.2017.1390142
49. Zimmerer, F.; Trouvain, J.: Productions of /h/ in German: French versus German speakers. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, January 2015, pp. 1922–1926. (2015)

50. Flege, J.E.: Production and perception of a novel, second-language phonetic contrast. J. Acoust. Soc. Am. **93**, 1589–1608 (1993). https://doi.org/10.1121/1.406818
51. Valente, A.; Pinet, S.; Alario, F.X.; Laganaro, M.: 'When' does picture naming take longer than word reading? Front. Psychol. **7**, 1–11 (2016). https://doi.org/10.3389/fpsyg.2016.00031
52. Sejdić, E.; Djurović, I.; Jiang, J.: Time–frequency feature representation using energy concentration: an overview of recent advances. Digit. Signal Process. **19**, 153–183 (2009). https://doi.org/10.1016/j.dsp.2007.12.004
53. Volkmann, J.; Stevens, S.S.; Newman, E.B.: A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am. **8**, 208–208 (1937). https://doi.org/10.1121/1.1901999
54. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.; McVicar, M.; Battenberg, E.; Nieto, O.: librosa: Audio and Music Signal Analysis in Python. In: Proceedings 14th Python in Science Conferences, pp. 18–24. (2015). https://doi.org/10.25080/majora-7b98e3ed-003
55. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury Google, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Xamla, A. K.; Yang, E.; Devito, Z.; Raison Nabla, M. et al.: NeurIPS-2019-pytorch-an imperative style high performance deep learning library Paper. NeurIPS (2019)
56. Goodfellow, I.; Bengio, Y.; Courville, A.: Deep Learning, MIT Press, (2016)
57. Bishop C. M.: Pattern Recognition and Machine Learning, Springer, (2006)
58. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
59. Korvel, G.; Treigys, P.; Kostek, B.: Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network. J. Acoust. Soc. Am. **149**, 508–523 (2021). https://doi.org/10.1121/10.0003339
60. Saito, K.; Plonsky, L.: Effects of second language pronunciation teaching revisited: a proposed measurement framework and meta-analysis. Lang. Learn. **69**, 652–708 (2019). https://doi.org/10.1111/lang.12345
61. Alsharhan, E.; Ramsay, A.: Robust automatic accent identification based on the acoustic evidence. Int. J. Speech Technol. **26**, 665–680 (2023). https://doi.org/10.1007/s10772-023-10031-2
62. Sturm, J.L.: Current approaches to pronunciation instruction: a longitudinal case study in French. Foreign Lang. Ann. **52**, 32–44 (2019). https://doi.org/10.1111/flan.12376
63. Inceoglu, S.: Effects of perceptual training on second language vowel perception and production. Appl. Psycholinguist. **37**, 1175–1199 (2016). https://doi.org/10.1017/S0142716415000533
64. Liakin, D.; Cardoso, W.; Liakina, N.: Learning L2 pronunciation with a mobile speech recognizer: French/y/. CALICO J. **32**, 1–25 (2015). https://doi.org/10.1558/cj.v32i1.25962
65. Newbill, P.B.; Jones, B.D.: Students' motivations for studying French: examining undergraduates' language orientations, expectancies, and values to promote advocacy. NECTFL Rev. **69**, 69–91 (2012)
66. Simon, E.; Chambless, D.; Kickhöfel Alves, U.: Understanding the role of orthography in the acquisition of a non-native vowel contrast. Lang. Sci. **32**, 380–394 (2010). https://doi.org/10.1016/j.langsci.2009.07.001
67. Levy, E.S.; Strange, W.: Perception of French vowels by American English adults with and without French language experience. J. Phon. **36**, 141–157 (2008). https://doi.org/10.1016/j.wocn.2007.03.001
68. Baker, W.; Trofimovich, P.: Perceptual paths to accurate production of L2 vowels: the role of individual differences. IRAL Rev. Appl. Linguist. Lang. Teach. **44**, 231–250 (2006). https://doi.org/10.1515/IRAL.2006.010
69. Major, R. C.: Foreign Accent, Routledge, (2001). https://doi.org/10.4324/9781410604293