JASA ARTICLE

Evaluation of aspiration problems in L2 English pronunciation employing machine learning[a)]

Magdalena Piotrowska,[1] Andrzej Czyżewski,[2,b)] Tomasz Ciszewski,[3] Gražina Korvel,[4,c)] Adam Kurowski,[2,d)] and Bożena Kostek[5,e)]

[1]*Composition Department, Faculty of Composition, Interpretation and Musical Education, Academy of Music, Krakow, Poland*

[2]*Department of Multimedia Systems, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland*

[3]*Institute of English and American Studies, Faculty of Languages, University of Gdańsk, Gdańsk, Poland*

[4]*Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania*

[5]*Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Gdańsk, Poland*

ABSTRACT:

The approach proposed in this study includes methods specifically dedicated to the detection of allophonic variation in English. This study aims to find an efficient method for automatic evaluation of aspiration in the case of Polish second-language (L2) English speakers' pronunciation when whole words are analyzed instead of particular allophones extracted from words. Sample words including aspirated and unaspirated allophones were prepared by experts in English phonetics and phonology. The datasets created include recordings of words pronounced by nine native English speakers of standard southern British accent and 20 Polish L2 English users. Complete unedited words are treated as input data for feature extraction and classification algorithms such as *k*-nearest neighbors, naive Bayes method, long-short term memory, and convolutional neural network (CNN). Various signal representations, including low-level audio features, the so-called mid-term and feature trajectory, and spectrograms, are tested in the context of their usability for the detection of aspiration. The results obtained show high potential for an automated evaluation of pronunciation focused on a particular phonological feature (aspiration) when classifiers analyze whole words. Additionally, CNN returns satisfying results for the automated classification of words containing aspirated and unaspirated allophones produced by Polish L2 speakers. © *2021 Acoustical Society of America*.

https://doi.org/10.1121/10.0005480

(Received 10 February 2021; revised 22 May 2021; accepted 8 June 2021; published online 8 July 2021)

[Editor: James F. Lynch] Pages: 120–132

I. INTRODUCTION

The analysis of allophones, representing very short fragments of speech, which are defined as variants of phonemes, is a field of speech analysis that still poses many challenges (Mitterer *et al.*, 2018; Recasens, 2012). Such a "microscopic" approach to speech analysis opens new perspectives in numerous areas, e.g., pre-lexical processing in spoken-word recognition, allophonic and phonemic identity in speech recognition (Rabha *et al.*, 2019), automatic evaluation of pronunciation (Shahin and Ahmed, 2019), differences between speakers' native regional accents (Aubanel and Nguyen, 2010), or the evaluation of speech articulatory disorders (Jiao *et al.*, 2017) as it allows a more in-depth speech analysis.

Recent studies on the quality of automatic allophone evaluation utilize speech recognition technology, i.e., creating speech databases, feature extraction, machine learning (Almpanidis and Kotropoulos, 2007; Czyżewski *et al.*, 2017a; Czyżewski *et al.*, 2017b; Dalka *et al.*, 2014; Ge *et al.*, 2011; Korvel *et al.*, 2021; Ge *et al.*, 2011; Wei *et al.*, 2009), and to a lesser extent they concentrate on the mechanism of the phenomenon (Korvel and Kostek, 2017; Korvel and Kostek, 2018; Mitterer *et al.*, 2018). Since articulation is a much-debated topic (Dromey and Black, 2017; Illa and Ghosh, 2020; Pandey and Shah, 2009; Shahin and Ahmed, 2019; Yu *et al.*, 2019), the approach proposed in this study includes methods specifically dedicated to allophonic variants. The potential consequence of the lack of aspiration of /p, t, k/ in the appropriate context is that they may be misperceived by native English speakers as voiced /b, d, g/, which can change the meaning of the word and lead to miscommunication.

The aim of this study is to find an efficient approach for the automatic detection of aspiration and its evaluation in the speech of Polish second-language (L2) users of English based on the analysis of whole words instead of extracted allophones only. It should be noted that the phenomenon of aspiration was selected as a focal point of this study since it

[a)]This paper is part of a special issue on Machine Learning in Acoustics.

[b)]ORCID: 0000-0001-9159-8658.
[c)]ORCID: 0000-0002-1931-6852.
[d)]ORCID: 0000-0002-5132-3016.
[e)]Electronic mail: bokostek@audioacoustics.org, ORCID: 0000-0001-6288-2908.

 0001-4966/2021/150(1)/120/13/$30.00

https://doi.org/10.1121/10.0005480

is particularly difficult for Polish learners of English, both perceptually and articulatorily. The basis of this study is previous research works performed by the authors (Czyżewski *et al.*, 2017a; Piotrowska *et al.*, 2018a; Piotrowska *et al.*, 2018b; Korvel *et al.*, 2019). These works provided some of the recordings and approaches to allophone parametrization. In the study by Piotrowska *et al.* (2018b), a list of words pronounced by native and non-native English speakers was recorded, then edited and analyzed. Aspirated and unaspirated allophones of voiceless plosive consonants were then extracted from the recordings of English native speakers. Automatic classification of aspirated and unaspirated /p, t, k/ allophones returned promising results. However, the experiments pertained to the classification of allophones separated from words by manual editing. In the present study, the authors focus on the automatic recognition of aspiration without extracting relevant allophones from the signal.

The process of segmentation of allophones is extremely arduous. It also requires the phonology experts' involvement, which is a major difficulty for creating a sufficiently large dataset for an automatic system that could provide pronunciation evaluation feedback for the speaker. Few studies address the issue of automatic allophone segmentation (Almpanidis and Kotropoulos, 2007; Rafałko, 2016). However, the proposed solutions cannot be efficiently implemented. Aspiration is a time-based phenomenon, and inaccuracies in segmentation may cause additional bias. That is why the main goal of our work is to use whole words as the input to the classification algorithms employing the allophonic analysis of speech.

In the present work, *k*-nearest neighbors (kNN) and naive Bayes methods are included for comparison with the previous research (Piotrowska *et al.*, 2018b). Since both these learning algorithms belong to baseline classification methods, in the current paper, the main emphasis falls on using more state-of-the-art methods, such as long-short term memory (LSTM), a type of recurrent neural network that deals with sequential data (Salehinejad *et al.*, 2018; Illa and Ghosh, 2020), and convolutional neural network (CNN) (Shahin and Ahmed, 2019; Tsipas *et al.*, 2020; Vrysis *et al.*, 2020).

Moreover, different representations of audio signals, as well as various algorithm configurations, were employed in the course of the study to find the best combination of signal/algorithm representation. The proposed speech signal features were determined in three ways, i.e., first, low-level signal descriptors were calculated according to their definitions, then the mid-term and feature trajectories were built upon these descriptors. Finally, a two-dimensional (2D) signal representation based on spectrograms was used. More detailed information concerning signal representation is included in Sec. III.

Two datasets employed in the experiments were based on a list of English words compiled by three phonology experts, whose task was to evaluate the correctness of aspiration in the recordings auditorily. Nine native English speakers with a standard southern British accent were first recorded while reading the list of target words. Then, the recordings of 2 Polish L2 English speakers were produced. The L2 recordings were made in different conditions than the ones containing the renditions by English native speakers. The recordings are accessible online (Czyżewski *et al.*, 2017b; Piotrowska *et al.*, 2021).

There were two experiment phases, namely, experiment I and experiment II. The first one concerns classifier training and it included speech recordings of native English speakers. The configuration algorithm settings, which returned the best results on this dataset during experiment I, were then utilized on the dataset containing Polish L2 English speakers (experiment II).

The remainder of the paper is organized as follows: Section II briefly introduces the concept of the phonetic/phonological aspects of aspiration. The description of the material used is contained in Sec. III, while the proposed methodology of automatic assessment of aspiration (feature extraction and classification) is discussed in Sec. IV. Section V evaluates the performance of the proposed methods through the analysis of experimental results pertaining to two different datasets. Concluding remarks are presented in Sec. VI.

II. ASPIRATION

Aspiration has been traditionally defined in phonetic literature as an extra "puff of air" or breath following the release of voiceless plosives /p, t, k/, e.g., Heffner (1950) or Jones (1956). A slightly different approach is represented in definitions referring to the so-called voicing lag, i.e., the duration of voiceless period accompanied by glottal friction between the stop release and the onset of F0 of the following vowel, i.e., periodicity that reflects laryngeal vibration (Lisker and Abramson, 1964). Aspiration in English, unlike in, e.g., Hindi (Jensen, 2004; Islam, 2019), is allophonic. The categorical (phonemic) distinction between /p^h/ as /ϕ/ should, however, be maintained due to the presence of burst energy in /p^h/, which is absent in /ϕ/. Presumably, if a language (e.g., Hindi) has a three-way stop contrast [p] vs [p^h] vs [b] misperception does not occur.

The English aspiration rule can be formulated as follows: voiceless stops are aspirated if: (a) a vowel follows, (b) they are not preceded by /s/, (c) they are in the onset position of a stressed syllable. Thus, the /p/ in *pit* is strongly aspirated, whereas it is unaspirated in **spit*, **play*, and **cap*. An acoustic parameter representing pre-voicing, voicing, and aspiration of stops is voiced onset time (VOT), which can be negative, zero, or positive, respectively. It has been established that the duration of the positive VOT in English ranges from approximately 60 to 100 ms (Cho and Ladefoged, 1999), and its extent depends on the immediate phonetic environment and the place of articulation of the plosive. This means that VOT is shorter after bilabials than velars, which suggests that F1 transitions are important clues for voicing perception (Benki, 2001). Polish and English differ in their VOT implementation in cueing the contrast between /b, d, g/ and /p, t, k/. Polish uses pre-voicing, or

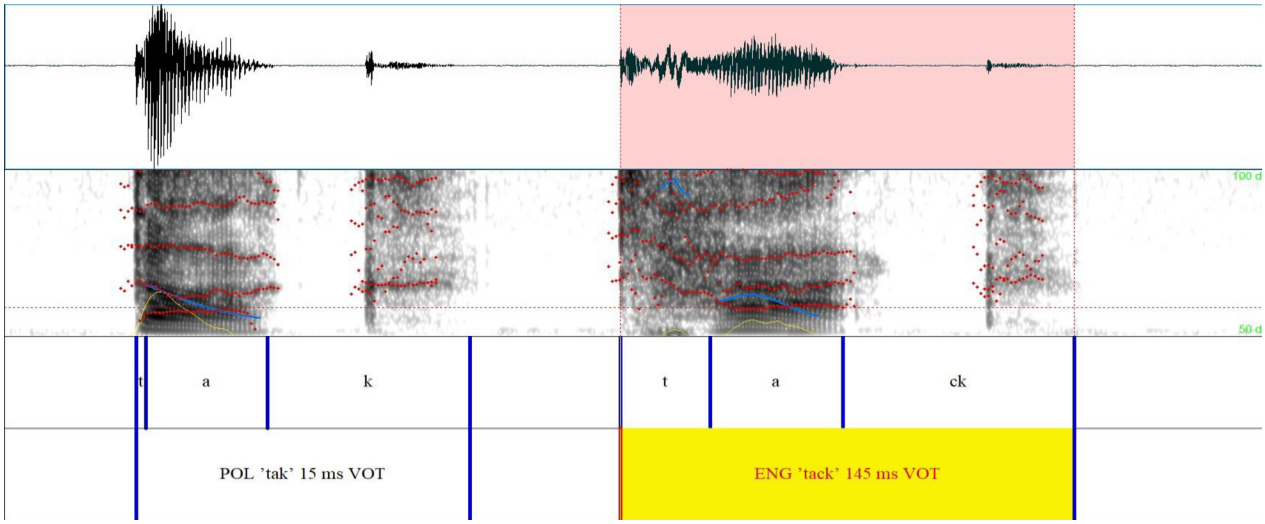

FIG. 1. (Color online) Polish word *tak* (English *yes*) and English word *tack*. Polish /t/ has 15 ms VOT, and English /t/ has 145 ms VOT.

negative VOT values, for voiced /b, d, g/ and short-lag VOT values for voiceless /p, t, k/ (Mikoś *et al.*, 1978; Keating *et al.*, 1981). On the other hand, English contrasts short-lag VOTs for voiced /b, d, g/ and long-lag VOTs for voiceless /p, t, k/ (Lisker and Abramson, 1964; Keating *et al.*, 1983). Polish L2 English learners/speakers transfer pronunciation habits from their native language, and they do not produce sufficiently long VOT in English /p, t, k/ (Rojczyk, 2010; Waniek-Klimczak, 2005). The consequence is that their /p, t, k/ in English have short-lag VOTs and, as a result, are perceived as voiced /b, d, g/ by native speakers of English. Figure 1 shows the Polish word *tak* (English *yes*) and the English word *tack*. Polish /t/ has 15 ms VOT, whereas the English /t/ has 145 ms VOT.

Previous research has shown that Polish speakers can imitate English long VOT in immediate imitation after the model (Rojczyk, 2012). However, when the imitation was distracted by asking participants to read random numbers before imitating the model, the produced VOT was reset to native Polish short-lag values. As already mentioned, aspiration is particularly difficult for Polish learners of English, so that is why the study is motivated by this aspect. It should also be noted that Polish speakers were familiar with the words they were reading, as knowing a word in L2 is important in phonological analysis context (Woore, 2018).

III. MATERIAL

We tested the performance of various feature/algorithms configurations using a set of recordings prepared specifically for the purpose of this research. As already mentioned, two datasets were created. The first one included the recordings of nine native (L1 mother tongue) English speakers and was used for the training stage (experiment I). The second dataset consisted of the recordings of native Polish speakers of English. It was utilized only for testing (experiment II) according to this study goal, whose aim is to arrive at reliable methods of an automatic detection and evaluation of aspirated allophones based on whole words.

A. Dataset I

The recording setup consisted of a shotgun microphone and an audio recorder (Zoom H4). The audio files were recorded with 44 100 S/s/16-bit resolution. The dataset includes the speech of nine native English speakers of Standard Southern British English. The set of target words created by a phonology expert consisted of 30 words (15 aspirated and 15 unaspirated) is listed in Table I.

B. Dataset II

The second dataset consisted of 240 examples collected in various conditions. All recordings were executed with 44 100 S/s/, 16-bit resolution. English speech (six words, repeated twice) of 20 Polish speakers was captured.

TABLE I. List of words used in experiment I (training) containing aspirated and unaspirated variants of allophones.

Unaspirated	Aspirated
speed	peep
spit	pit
spam	pe.g.,
spark	pack
sport	park
steal	team
still	tick
step	tent
stack	tap
start	task
ski	keep
skin	kid
sketch	court
scan	cat
scar	cart

TABLE II. List of words used in experiment II containing aspirated allophones.

Aspirated
peck
park
turn
tent
court
cup
care

The vocabulary set contained only words with aspiration, as listed in Table II. L2 English speakers were familiar with all target words. All renditions were auditorily rated by a phonology expert, whose evaluation notes were treated as a reference in the other experiment. The best networks found with the use of grid search were used for the classification of these examples.

IV. PROPOSED METHODOLOGY

As already mentioned, recent studies on the quality of automatic phoneme evaluation utilize speech recognition technology, i.e., creating speech databases, extracting features, applying machine learning (Almpanidis and Kotropoulos, 2007; Adams *et al.*, 2018; Czyżewski *et al.*, 2017a; Czyżewski *et al.*, 2017b; Korvel *et al.*, 2021; Ge *et al.*, 2011; Wei *et al.*, 2009). The approach proposed in this study includes methods dedicated explicitly to allophonic variants, i.e., aspirated and non-aspirated ones.

A. Feature extraction and signal representation

In this section, we elaborate on the extraction of the speech signal representations and features utilized in the present research. Since we deal here with temporal signal characteristics, it is worth to refer to other approaches to feature extraction of temporal characteristics present in the literature. Vrysis *et al.* (2020) point out that useful structural information is usually hidden within frame-based feature vector sequences, the so-called texture windows. Incorporating and using this information into the feature extraction process is called temporal feature integration (TFI). There exist other similar strategies, such as enhanced temporal or statistical temporal integration (Tsipas *et al.*, 2015; Vrysis *et al.*, 2020). Tsipas *et al.* (2020) extracted a sequence of feature vectors (FVs) using a time window when dealing with semi-supervised audio-driven diarization. Subsequently, they aggregated these FVs to form successive sequences of N aggregated vectors (Tsipas *et al.*, 2020).

In this study, low-level signal descriptors, two variants of time-related parameters built upon these low-level signal features, namely, mid-term statistics and trajectories of acoustic features, and also a 2D (two-dimensional) speech signal representation, i.e., spectrograms, are employed. These approaches bring a different perspective in terms of averaging information included in the speech signal over time.

Our previous investigations of phonological processes show that in order to determine the aspiration of voiceless stop consonants, features, including energy measures of temporal distribution, should be used (Piotrowska *et al.*, 2018a). The description of features selected in the context of the present research is included in the summary below. The corresponding low-level features are given in Table III; they refer to the low-level signal descriptors.

Most of the parameters presented in Table III are extensively used as speech signal descriptors; several are the features derived from the music information retrieval (MIR) domain (Kim *et al.*, 2005; Plewa and Kostek, 2015; Rosner and Kostek, 2018). The choice of all these parameters is justified because using speech features and music descriptors provides a more effective phoneme and allophone recognition than using them alone (Korvel and Kostek, 2018; Korvel *et al.*, 2019; Piotrowska *et al.*, 2018a). The first parameter in Table III refers to the number of samples included in the allophone. Temporal centroid (TC) is the time average over the signal energy envelope (Kim *et al.*, 2005). Zero-crossing rate (ZCR) reflects the number of times the signal crosses the time axis. Root mean square (RMS) energy gives mean energy in the analyzed signal frame. The parameters Nos. 5–28 are dedicated descriptors proposed by Kostek and her collaborators (Kostek *et al.*, 2011). The first group of dedicated parameters consists of the numbers of samples exceeding levels $r1$, $r2$, and $r3$, where $r1$, $r2$, and $r3$ are equal to RMS, 2RMS, 3RMS, respectively. These parameters were calculated in two different ways: for the entire short-time segment and ten sub-segments. Parameters 14–16 are the peak to RMS ratio calculated for the entire short-time segment and ten sub-segments. The last group of dedicated parameters is related to observing the threshold crossing rate (TCR). The calculation procedure consists in computing the number of signal crossings related to zero, $r1$, $r2$, and $r3$ values.

As reported in Sec. II, aspiration of voiceless stops occurs immediately before a vowel in the absolute onset of a stressed syllable. Therefore, obtaining information pertaining to energy distribution within the allophone is crucial. Mid-term statistics cover general changes in features over time and they are commonly used in speech analysis (Giannakopoulos and Pikrakis, 2014; Smailis *et al.*, 2016). However, it should be noted that the analyzed phenomenon of aspiration is associated with very short time segments (about tens of ms). Hence, trajectories of signal features related to a higher time resolution are also utilized. These trajectories are a sequence of numbers containing information on the occurrence or absence of aspiration in the speech signal analyzed. Accordingly, we calculated these two variants of parameters in terms of averaging them in the

TABLE III. List of low-level signal features used for the automatic evaluation of aspiration.

No.	Feature	Formula
1	Number of samples	Indicates the number of samples included in the allophone
2	Temporal centroid (TC)	$$TC = \frac{\sum_{k=1}^{M} k(x_i(k))^2}{\sum_{k=1}^{M} (x_i(k))^2}$$ where $x_i(k)$ are the samples of the ith speech segment, M is the segment length
3	Zero-crossing rate (ZCR)	$$ZCR = \frac{\sum_{k=2}^{M} \lvert s(k) - s(k-1) \rvert}{M-1}$$ where $$s(k) = \begin{cases} 1 & if\ x_i(k) > 0 \\ 0 & if\ x_i(k) \le 0, \end{cases}$$ where $x_i(k)$ are the samples of the ith speech segment, M – the segment length
4	Root mean square (rms) energy	$$RMS = \sqrt{\frac{\sum_{k=1}^{M} (x_i(k))^2}{M}}$$ where $x_i(k)$ are the samples of the ith speech segment, M is the segment length
5 – 7	Number of samples exceeding r_1, r_2, and r_3 threshold	$$p_n = \frac{count(samples_{exceeding} r_n)}{length(x_i(k))}$$ where $n = 1, \ldots, 3$ and $x_i(k)$ represents the analyzed signal segment
8 – 13	The mean (q_n) and variance (v_n) of samples exceeding r_1, r_2, and r_3 threshold averaged for 10 sub-segments	$$q_n = \frac{\sum_{k=1}^{10} p_n^k}{10}$$ $$v_n = \frac{\sum_{k=1}^{10} (p_n^k - q_n)}{9}$$ where p_n^k is a number of samples exceeding the level r_n calculated in the k th sub-segment, $n = 1, \ldots, 3$
14	PEAK TO RMS	$$\mathrm{PRMS} = \frac{max\{\lvert (x_i(1) \rvert,\ \lvert x_i(2) \rvert,\ \ldots,\ \lvert x_i(M) \rvert\}}{RMS}$$ where $x_i(k)$ are the samples of the ith speech segment, M is the segment length
15 – 16	The mean (q) and variance (v) of PEAK TO RMS averaged for 10 sub-segments	$$q = \frac{\sum_{k=1}^{10} \mathrm{PRMS}^k}{10}$$

https://doi.org/10.1121/10.0005480

TABLE III. (*Continued*)

No.	Feature	Formula
		$v = \frac{\sum_{k=1}^{10}(\mathrm{PRMS}^k - q)}{9}$ where PRMS^k is peak to RMS ratio calculated in the k th sub-segment
17 – 20	Number of signal crossings in relation to zero, $r1$, $r2$, and $r3$	$c_1 = \frac{count(samples_{crossing} zero)}{length(x_i(k))}$ $c_n = \frac{count(samples_{crossing} r_{n-1})}{length(x_i(k))}$ where $n = 2, \ldots, 4$ and $x_i(k)$ represents the analyzed signal segment
21 – 28	The mean (q_n) and variance (v_n) of signal crossings in relation to zero, r_1, r_2, and r_3 averaged for 10 sub-segments	$q_n = \frac{\sum_{k=1}^{10} c_n^k}{10}$ $v_n = \frac{\sum_{k=1}^{10}(c_n^k - c_n)}{9}$ where c_n^k is a number of signal crossings in relation to zero, $r1$, $r2$, and $r3$ calculated in the k th sub-segment, $n = 1, \ldots, 4$

time domain. They are built upon the low-level features contained in Table III.

In this research, we are focusing on automatic recognition of aspiration. Therefore, mid-term statistics are applied to the whole word, without extracting allophones from the speech signal. Let the vector $\boldsymbol{x}$ represent samples of the analyzed speech signal:

$$\boldsymbol{x} = \big(x(1), x(2), \ldots, x(N)\big), \tag{1}$$

where N is the sample number. The mid-term statistics extraction process consists of three steps:

Step 1. Signal dividing into segments.
Step 2. Extraction of short-term features.
Step 3. Statistics calculation.

The feature calculation process starts with the division of speech signals into short-time segments. The segment length is 1024 samples, and the overlap between contiguous segments is equal to 50%. The samples of the analyzed segment can be described as follows:

$$\boldsymbol{x}_i = \big(x_i(1), x_i(2), \ldots, x_i(M)\big), \tag{2}$$

where i is the frame number and M is the segment length.

The input signal was divided into short-term segments with the length of, and the low-level features given in Table III were calculated. Based on them the mid-term statistics, the feature trajectories are then calculated. The length of the speech signal after the zero-padding procedure equals 81 022 samples. It should be observed that zero-padding is a technical requirement imposed in an indirect manner by many deep learning frameworks. It is a standard practice used if one has to ensure that all the data fit into a tensor of a strictly defined shape which has to be the same for all examples from the dataset. This step is also required for CNNs because they, by definition, can only process data frames of a fixed size. Zero-padding may potentially influence the internal state of recurrent neural networks if padding takes a significant portion of the word. Some studies suggest such a phenomenon, especially if zero-padding happens at the end of an example (Dwarampudi and Reddy, 2019). This potential problem can be further addressed with the addition of an attention mechanism to the recurrent neural network or a masking mechanism, but such modification further complicates the neural networks, especially in comparison with plain CNNs.

Then the vector of low-level features is extracted from each segment. As a result, we obtained 27 feature sequences. In the last step, the signal is divided into mid-term segments, and for each of them, the feature statistics are calculated. The following pseudo-code presents the procedure of dividing short-term feature vector into mid-term segments:

INPUT:
L_1 – the number of short-term segments
L_2 – the number of mid-term segments
PROCEDURE:
1 $step \leftarrow \lfloor L_1/L_2 \rfloor$
2 $n \leftarrow 1$
3 $segm \leftarrow step \times 2$
4 WHILE $(n \leq L_2)$
5 $P_{start}(n) \leftarrow (n - 1) \times step + 1$
6 $P_{end}(n) \leftarrow (n - 1) \times step + segm$
7 IF $P_{end}(n) > L_1$ THEN
8 $P_{end}(n) = L_1$
9 END IF
10 $n \leftarrow n + 1$
11 END WHILE
OUTPUT:
$P_{start}(1), \ldots, P_{start}(L_2)$ are the start edges of the mid-term segments,
$P_{end}(1), \ldots, P_{end}(L_2)$ are the end edges.
The comment on the algorithm: symbol $\lfloor * \rfloor$ denotes the floor function.

The number of mid-term segments is equal to 6. The following mid-term statistics are employed: the mean value and the standard deviation. As a result, the mid-term statistics are a matrix, which is related to time-series features. The matrix is converted into a single-column vector before these statistics are submitted to the classifier. For this purpose, all rows are transposed and concatenated to produce a single column vector.

In order to obtain trajectories of low-level features, first of all, a zero-padding technique is used. This technique consists in padding the N signal samples by zeros:

$$\boldsymbol{y} = \left(x(1), \ldots, x(N), \underbrace{0, \ldots, 0}_{N_{max}-N}\right), \tag{3}$$

where N_{max} is a sample number of the longest speech signal used in the experiment.

After padding the signal by zeros (each up to 96 000 samples to cover the duration of the longest word in the database), it is divided into short-time segments, and then the acoustic features are extracted with frame 1024 samples (approximately 20 ms) and overlap 256 samples (approximately 5 ms). This time resolution was chosen with regard to the duration of the aspirated period, which may be as long as 150 ms, while the duration of unaspirated stops may be as short as 20 ms or less. In this setup, 27 features contained in Table III were calculated. It is possible for the algorithms to recognize patterns due to the duration of the word instead of recognizing information characteristic for particular phonemes since the zero-padding may emphasize such type of temporal information. However, one has to stress that the duration of a phoneme may be considered as one of the distinctive features useful for phoneme recognition (Hamooni *et al.*, 2016; Kazanina *et al.*, 2018; Nahar *et al.*, 2012). Especially in Hamooni *et al.* (2016) and Nahar *et al.* (2012), one can find that statistically, in terms of mean length, some phonemes can be separated from others, and thus the length of a phoneme becomes a viable parameter that can be employed in a machine learning system for phoneme recognition. This is as good source of information as any other parameter such as spectral centroid or skewness. In combination with all other obtainable metrics, the length of the phoneme may be one of the premises used to infer the type of a phoneme.

Some classification algorithms may need a two-dimensional representation of acoustic signals. An example of such classifiers is a two-dimensional CNN (Brocki and Marasek, 2015; Korvel *et al.*, 2018; Korvel *et al.*, 2019; Saleem *et al.*, 2019; Salehinejad *et al.*, 2018; Vrysis *et al.*, 2020).

Therefore, speech recordings were also transformed into a 2D domain by calculating the spectrograms. Since input fed into a neural network has to be of uniform shape, all audio recordings were padded with zeros to the length of the longest recording. Additional zeros were appended to the end of the audio signal. The sampling rate of recordings is 48 kHz/s.

Various types of framing employed in spectrogram calculation were checked for the purpose of hyperparameter selection. Frames of lengths equal to 128, 256, 512, 1024, and 2048 were tested. Similarly, the overlap factors of 0.05, 0.5, and 0.75 were also checked for their impact on performance models.

The amplitudes obtained from the fast Fourier transform were transformed to the logarithmic domain according to the following formula:

$$\boldsymbol{sp}_{\boldsymbol{dB}} = 20 \log_{10} \max(|\boldsymbol{sp}|, sp_{min}), \tag{4}$$

where $\boldsymbol{sp}_{\boldsymbol{dB}}$ denotes a spectrogram in the logarithmic domain, $\boldsymbol{sp}$ denotes the values of the spectrogram in the linear domain, and sp_{min} determines the lowest possible value of $\boldsymbol{sp}_{\boldsymbol{dB}}$. With the use of the last parameter, it is possible to cut off the low-amplitude noises present in the recording and force machine learning algorithms to focus on the components of signals with greater amplitudes. In the case of our research sp_{min} is equal to

$$sp_{min} = \max(|\boldsymbol{sp}|)/20e3. \tag{5}$$

The last step was the scaling of $\boldsymbol{sp}_{\boldsymbol{dB}}$ values to range from 0 to 1 performed according to the following formula:

$$\overline{\boldsymbol{sp}_{\boldsymbol{dB}}} = \left[\boldsymbol{sp}_{\boldsymbol{dB}} - \min(\boldsymbol{sp}_{\boldsymbol{dB}})\right]/\max(\boldsymbol{sp}_{\boldsymbol{dB}}). \tag{6}$$

The spectrogram with normalized values is denoted as $\overline{\boldsymbol{sp}_{\boldsymbol{dB}}}$.

B. Classification methods

The machine learning methods used for the evaluation of extracted features are described in this section. In the experiment, we use two baseline algorithms, namely the naive Bayes classification method and kNN, to assign the

https://doi.org/10.1121/10.0005480

class of the unknown test object to the known object belonging to the training class. These two methods were implemented due to very promising results, which were obtained at the previous stage of the research (Piotrowska *et al.*, 2018b). In the present study, we also employed state-of-the-art methods using feature-based and convolutional neural networks, including LSTM and CNN (Brocki and Marasek, 2015; Illa and Ghosh, 2020; Korvel *et al.*, 2018; Korvel *et al.*, 2021; Saleem *et al.*, 2019; Tsipas *et al.*, 2020; Vryzas *et al.*, 2018; Vrysis *et al.*, 2020). It is evident that CNNs may be successfully exploited to learn and model high-dimensional hierarchical signal/2D data representations as well as 1D feature vectors (Deng *et al.*, 2020; Vrysis *et al.*, 2020). CNN-based systems are also capable of modeling temporal dependencies, which makes it possible to use them in such tasks as synthesis of emotional speech (Choi *et al.*, 2019) or voice cloning (Partila *et al.*, 2020). Especially the latter CNN use example is important and should be thoroughly examined as in the future it may pose a potential threat to biometric access control systems based on voice recognition. There are examples of systems performing the aforementioned sample tasks based only or mainly on CNNs. Both of them rely heavily on capturing and processing the temporal evolution of generated or processed audio signals. Moreover, LSTM capability to learn acoustic-articulatory mappings through a single acoustic-to-articulatory inversion model, rather than building a separate speaker-specific model (Illa and Ghosh, 2020), is advantageous when dealing with real-life recordings of multi speakers. Tsipas *et al.* (2020) proved that applying LSTM networks to several sequence-to-sequence and sequence-to-vector classification scenarios can provide a mechanism to incorporate temporal characteristics into the diarization process successfully. Thus, based on the proven advantages of these algorithms, we assumed that CNN and LSTM might significantly enhance an automated evaluation of pronunciation focused on a particular phonological feature (aspiration) when classifiers analyze whole words.

1. Baseline classifiers

The naive Bayes classification method is based on the Bayes theory (Ghosh *et al.*, 2007). The test object is assigned to the class with the maximum class probability. The probability that a speech with the parameter vector z belongs to a class c_k ($k = 1 \ldots K$, where K is the number of classes) can be stated as follows:

$$P(c_k|z) = \frac{P(z|c_k)P(c_k)}{P(z)}. \tag{7}$$

Gaussian kernel density is used to estimate the class condition probability

$$P(y|c_k) = \prod_{j=1}^{d}\left(\frac{1}{Lh}\sum_{i=1}^{L}\frac{1}{\sqrt{2\pi}}e^{-1/2\left((y_j - x_{ij})/h\right)}\right), \tag{8}$$

where h is the bandwidth for the control of the smoothness of the density curve (Chiu, 1991) and L is the number of objects in the class c_k. According to the k-Nearest Neighbors (kNN) classification algorithm, the test set objects are classified by calculating the nearest training object distance. In this paper, the Euclidean metric is used. The optimum number of nearest neighbors is established by performing a series of preliminary tests.

2. Neural networks

The classification using neural networks was performed with two types of neural networks, i.e., LSTM (Illa and Ghosh, 2020; Tsipas *et al.*, 2020) and CNN (Buduma and Locascio, 2017; Vrysis *et al.*, 2020). The recurrent LSTM neural network was used in the classification employing trajectory-based parameters, as such the network architecture is suitable for the analysis of temporal data. CNN was employed for the analysis of spectrograms (2D speech representation). The topology of both networks is shown in Fig. 2. Some hyperparameters related to those topologies were changed in the course of a grid search to find the optimal values for each of them.

Each network consists of four layers of neurons. Additionally, CNN also contains two pooling layers after each convolutional layer. The Kernel size of convolutional layers was set to (3,3). The number of neurons and hyperparameters of the learning process was found by grid search. Therefore, the number of neurons in Fig. 2 is denoted as nl_1, nl_2, and nl_3 for three consecutive layers placed before the output layer consisting of two neurons. The output from the network is encoded in one hot manner. One output corresponds to one of two classes to which the input examples are assigned: aspirated or unaspirated. To prevent overfitting, the influence of the L1 and L2 regularization techniques was also investigated to find out if it can be used to enhance the performance of the model. The strength of

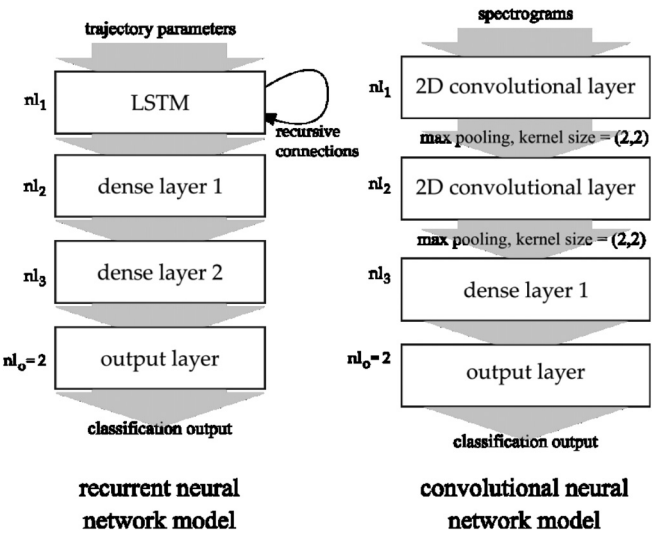

FIG. 2. Topologies of two types of neural networks employed for the classification of allophones. The number of neurons in the first three layers is denoted as nl_1, nl_2, and nl_3.

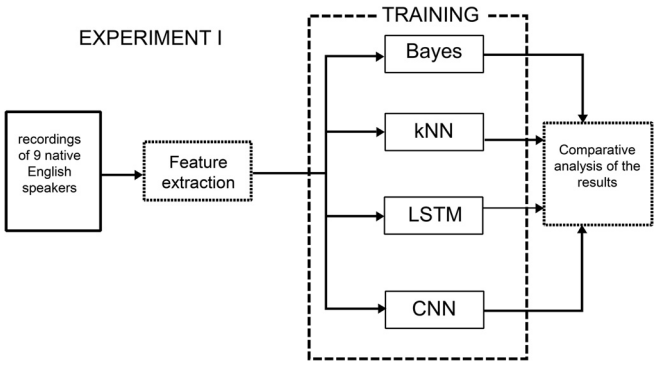

FIG. 3. Schema of experiment I, in which the recordings of nine native English speakers were used for training.

regularization is determined separately by the regularization coefficient for each type of regularization. More details associated with investigated types of activation functions and employed types of regularization may be found in the literature (Buduma and Locascio, 2017; Geron, 2017). The grid search was used to find the following hyperparameters:

- the numbers of neurons in each layer: nl_1, nl_2, and nl_3,
- the activation function of neurons in neural networks: rectified linear unit (ReLU) or exponential linear unit (ELU),
- the type of regularization: no regularization, L1 or L2 regularization with regularization coefficient set to 0.01, or with both L1 and L2 regularization with the coefficient set also to 0.01 for both types of regularization,
- the length of frame used for calculation of spectrogram (only for CNN),
- the overlapping factor used for calculation of spectrogram (only for CNN).

The tested architecture of the network was trained for 100 epochs. ADAM optimizer was employed as the learning rate optimizer algorithm (Kingma and Ba, 2014). Each test produced values of the accuracy of the network associated with training, validation, and test sets. The accuracy of the validation set was used to find the best performing architecture of the neural network. The calculations were performed with the use of TensorFlow and Keras machine learning

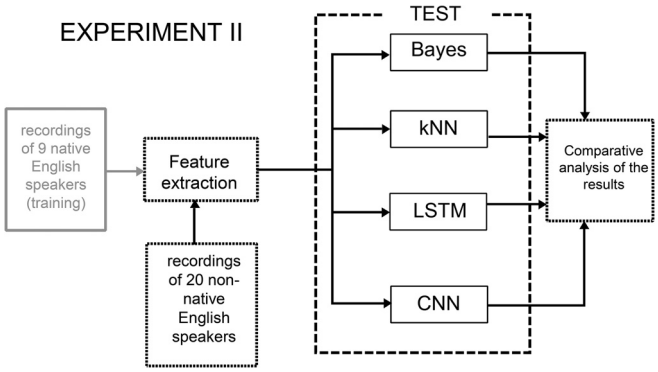

FIG. 4. Block-diagram of experiment II, where the recordings of 20 non-native speakers were used for testing.

TABLE IV. Accuracy (Acc.) of classification of aspirated and non-aspirated allophones for the mid-term parameters.

	Naive Bayes Acc. [%]		kNN Acc. [%]	
Allophone	Mean	STD	Mean	STD
/p/	93.33	8.660	95.56	7.265
/t/	88.89	10.541	97.78	4.410
/k/	86.67	12.247	93.33	8.660
All	89.26	5.720	93.33	5.271

libraries in PYTHON programming language (Abadi, 2019; Chollet, 2019).

V. EXPERIMENTS

As already mentioned, the presented study consists of two experiments (experiments I and II), whose schema are shown in Figs. 3 and 4. In experiment I, the recordings of nine native English speakers were used in order to find the proper setting configuration of the tested classifiers. The dataset I was used for training, validation, and test in proportions 7:2:1. A random split of data into training and testing sets was used. In experiment II, dataset II, which consists of the recordings of 20 Polish speakers of English was utilized, and only the testing part with no training was executed. Evaluation of aspiration is conducted at the level of isolated words. The details of the entire process are described in the following.

A. Experiment I—Training and classifier performance

In the first experiment (see Fig. 3), the feature extraction procedure proposed in Sec. IV A was performed. The obtained mid-term statistics and trajectories of low-level features are normalized to the range [0,1], and then divided into two segments: the first one employed to train the model and the other one to test this model. In order to verify the statistical significance of the results of our calculation, the cross-validation technique was used (Refaeilzadeh *et al.*, 2014).

1. Naive Bayes and kNN performance

The naive Bayes and kNN algorithms are employed as they were used in our earlier studies on allophone classification (Piotrowska *et al.*, 2018b). The performance of these learning algorithms is shown in Tables IV and V, where

TABLE V. Accuracy (Acc.) of classification of aspirated and non-aspirated allophones based on trajectories of low-level features.

	Naive Bayes Acc. [%]		kNN Acc. [%]	
Allophone	Mean	STD	Mean	STD
/p/	47.78	14.814	53.33	15.811
/t/	45.56	25.056	64.44	17.401
/k/	46.67	25.981	70.00	17.321
All	45.56	13.744	66.30	11.954

https://doi.org/10.1121/10.0005480

TABLE VI. Examples of the best sets of the topology parameters found through the grid-search with the use of a CNN. The best performing network is shown in the first position.

	Max accuracy [%]								Learning rate	Regularization		
ID	Training	Validation	Test	t_{trn}	nl_1	nl_2	nl_3	Act. func.		L1	L2	l_{frame}
1	**1.00**	**0.94**	**0.90**	**84.63**	**128**	**64**	**32**	ReLU	**0.001**	**0.01**	**0**	**512**
2	1.00	0.92	0.90	46.47	32	16	8	ReLU	0.001	0.01	0	512
3	1.00	0.92	0.87	85.63	128	64	32	ReLU	0.0001	0.01	0	512
4	1.00	0.92	0.90	83.05	128	64	32	ReLU	0.001	0	0	512
5	1.00	0.92	0.87	85.66	128	64	32	ReLU	0.0001	0.01	0.01	512
6	1.00	0.92	0.90	56.79	64	32	16	ReLU	0.001	0	0.01	512
7	1.00	0.92	0.82	49.85	64	32	16	ReLU	0.001	0	0.01	256
8	1.00	0.92	0.85	76.23	128	64	32	ReLU	0.001	0	0	256
9	1.00	0.92	0.90	85.00	128	64	32	ReLU	0.0001	0	0.01	512
10	0.99	0.92	0.79	38.74	32	16	8	ReLU	0.001	0	0.01	256
11	0.97	0.92	0.90	83.60	128	64	32	ReLU	0.0001	0	0	512
12	0.93	0.92	0.85	47.71	32	16	8	ReLU	0.0001	0.01	0.01	512

average results of allophone classification with regard to the aspiration of a particular allophone and allophone group are presented separately.

It is interesting to observe that much higher accuracies were obtained for both naive Bayes and kNN when the mid-term parameters were employed. Contrarily, the outcome in terms of accuracies for the trajectories built upon low-level features is relatively low. This means that these time-related features may not be suitable for the naive Bayes and kNN algorithms.

2. CNN and LSTM performance

It should be recalled that speech spectrograms were fed to the CNN input, while LSTM employed trajectory-based parameters.

Overall, the input dataset consisted of 384 examples. The input dataset was split into three subsets: training, validation, and test set with proportions of 7:2:1, which gave the training set of 268 examples, the validation set of 77 examples, and the test set consisting of 39 examples. During the grid search, 256 combinations of parameters were tested. The purpose of the validation set was to assess the performance of the models, plus the fine-tuning of the hyperparameters, which were used for the final evaluation of the trained model.

The results obtained for the selected best combinations of network parameters are shown in Tables VI and VII. Maximal accuracies for the validation, training, and test sets are provided, t_{trn} denotes the time of training (in seconds), nl_1, nl_2, nl_3 refer to the number of neurons in layers shown in Fig. 2;

TABLE VII. Examples of the best sets of the topology parameters found through the grid-search using an LSTM. The best performing network is shown in the first position.

	Max accuracy [%]								Learning rate	Regularization	
ID	Training	Validation	Test	t_{trn}	nl_1	nl_2	nl_3	Act. func.	Val.	Trn.	Test
1	**1.00**	**0.92**	**0.87**	**71.65**	**64**	**32**	**16**	**ELU**	**0.001**	**0**	**0.01**
2	0.98	0.90	0.82	80.94	64	32	16	ReLU	0.001	0	0.01
3	0.97	0.90	0.85	81.37	128	64	32	ReLU	0.001	0	0.01
4	0.96	0.90	0.85	79.08	16	8	4	ReLU	0.001	0	0
5	1.00	0.88	0.87	81.18	64	32	16	ReLU	0.001	0	0
6	0.99	0.88	0.82	76.89	16	8	4	ELU	0.001	0	0.01
7	0.98	0.88	0.87	75.01	16	8	4	ELU	0.001	0	0
8	0.97	0.88	0.72	89.19	128	64	32	ReLU	0.0001	0	0.01
9	0.96	0.88	0.85	82.83	16	8	4	ReLU	0.001	0	0.01
10	1.00	0.87	0.90	73.38	64	32	16	ELU	0.001	0	0
11	1.00	0.87	0.82	80.21	128	64	32	ReLU	0.001	0	0
12	1.00	0.87	0.87	71.38	32	16	8	ELU	0.001	0	0.01
13	0.99	0.87	0.85	70.58	128	64	32	ELU	0.001	0	0.01
14	0.99	0.87	0.87	29.53	64	32	16	ReLU	0.001	0.01	0.01
15	0.99	0.87	0.77	86.00	128	64	32	ReLU	0.0001	0	0
16	0.99	0.87	0.85	82.24	32	16	8	ReLU	0.001	0	0.01
17	0.98	0.87	0.85	82.80	32	16	8	ReLU	0.001	0	0
18	0.98	0.87	0.92	14.64	16	8	4	ELU	0.001	0	0

TABLE VIII. Classification accuracy [%] obtained in experiment II.

Method/Acc. [%]	Naive Bayes	kNN	CNN	LSTM
Mean	61.67	29.17	74.17	65
STD	48.72	45.55	43.86	47.8

the remaining columns contain the type of the utilized activation function (act. func.), the learning rate, and the values specified for L1 and L2 regularization parameters. The length of the spectrogram frame is denoted as l_{frame} and the overlap factor was 0. For the CNN-based neural networks, the activation function was a rectified linear unit (ReLU) for all setups, while for the LSTM-based classifier two options were tested: ReLU and exponential linear unit (ELU).

B. Experiment II

In experiment II the best networks found with the use of grid search in experiment I were used to classify aspirated allophones in the case of Polish speakers (English L2 pronunciation). Thus, only testing with no training was executed. The block diagram of experiment II (test) is shown in Fig. 4.

Dataset II consisted of 240 examples collected in different conditions (different room, microphone, etc.) from Polish speakers. All words were supposed to contain aspiration. The auditory evaluation performed by a phonology expert was used for reference in this task. Since in experiment I, the highest scores were obtained for mid-term parameters, thus only these descriptors were used in tests employing kNN and naive Bayes.

The best accuracy with the expert's evaluation obtained for CNN was 74% and 65% for LSTM (all results are presented in Table VIII). This seems a reasonable level of generalization between datasets, which differ in time, size, and acquisition conditions. The winning signal representation/classifier combination is CNN with spectrograms fed at its input. It also seems that LSTM should be exploited in further analysis as the results are promising.

The one-way analysis of variance (ANOVA) test is used to determine whether the differences between the classification accuracies obtained by different algorithms are statistically significant. The test significance level equals 0.05. The ANOVA test results are given in Table IX, where the significant differences are highlighted in bold font.

From the results given in Table IX, we may see that all differences between the classification accuracies—except those between naive Bayes and LSTM—are statistically significant.

VI. CONCLUSIONS

The analysis presented in this paper shows the potential for an automated evaluation of pronunciation focused on a particular phonological feature (aspiration) for non-native speakers based on whole words. The results obtained in experiment I return satisfying results for automated classification of words containing aspirated and unaspirated allophones. The audio features selected for the detection of aspiration in whole words seem appropriate because both approaches, mid-term features and trajectories, can be used for particular methods. The results of experiment II are mostly compatible with the phonology experts' ratings. The best results were achieved for the CNN setup, while the kNN method was not appropriate for this generalization level. Since aspiration is a phenomenon specifically difficult for Polish speakers (Mikoś *et al.*, 1978; Keating *et al.*, 1981), the results may be treated as promising.

Contrarily, as the dataset of recordings contains only Polish L2 speakers' speech, it is not possible to conclude how well the proposed method will perform for groups with different mother tongues. However, this aspect will be pursued in a future study, and recordings of aspirated allophones of L2 English speakers will be made and then tested.

Even though the datasets are relatively small, the list of words was carefully compiled by phonology experts. Thus, the approach proposed in the present study may be treated as a kind of benchmark, especially focusing on CNN and LSTM methods that could be utilized in automated support in the pronunciation learning process. However, the application of this task would require building a much larger corpus of recordings made in various acoustic conditions in order to guarantee the robustness of the evaluation systems and the reliability of results.

It should also be noted that the overall performance of both deep learning-based models was satisfactory, and the LSTM-based scores did not differ in a drastic manner from the performance of a CNN-based neural network, either. However, examining the possible influence of zero-padding is a very promising direction for future research, as there are only a few papers related to this topic.

Moreover, in the future approach, we will follow the work of Palaz *et al.* (2019) and search for relevant features automatically employing convolutional neural networks

TABLE IX. The result of the ANOVA test.

Methods/Acc. [%]	Naive Bayes /kNN	Naive Bayes /CNN	Naive Bayes /LSTM	kNN/CNN	kNN/LSTM	CNN/LSTM
F-value	**56.99**	**8.73**	0.57	**121.54**	**70.69**	**4.79**
F-value	**<0.00001**	**0.0033**	0.4497	**< 0.00001**	**< 0.00001**	**0.02908**

(CNNs) and reveal them as HMM states class conditional probabilities at the CNN output.

ACKNOWLEDGMENTS

This research was sponsored by the Polish National Science Centre, Dec. No. 2015/17/B/ST6/01874.

Abadi, M. (**2019**). "Tensorflow," https://www.tensorflow.org/ (Last viewed February 2020).
Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, S. (**2018**). "Evaluating phonemic transcription of low-resource tonal languages for language documentation," *LREC 2018 (Language Resources and Evaluation Conference)*, May 2018, Miyazaki, Japan, pp. 3356–3365, https://halshs.archives-ouvertes.fr/halshs-01709648v4/document (Last viewed February 2021).
Almpanidis, G., and Kotropoulos, C. (**2007**). "Automatic phonemic segmentation using the Bayesian information criterion with generalized gamma priors," in *15th European Signal Processing Conference*, pp. 2055–2059.
Aubanel, V., and Nguyen, N. (**2010**). "Automatic recognition of regional phonological variation in conversational interaction," Speech Commun. **52**(6), 577–586.
Benki, J. R. (**2001**). "Place of articulation and first formant transition pattern both affect perception of voicing in English," J. Phon. **29**(1), 1–22.
Brocki, Ł., and Marasek, K. (**2015**). "Deep belief neural networks and bidirectional long-short term memory hybrid for speech recognition," Arch. Acoust. **40**(2), 191–195.
Buduma, N., and Locascio, N. (**2017**). *Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms* (O'Reilly Media, Sebastopol, CA).
Chiu, S.-T. (**1991**). "Bandwidth selection for kernel density estimation," Ann. Stat. **19**(4), 1883–1905.
Cho, T., and Ladefoged, P. (**1999**). "Variation and universals in VOT: Evidence from 18 languages," J. Phon. **27**(2), 207–229.
Choi, H., Park, S., Park, J., and Hahn, M. (**2019**). "Multi-speaker emotional acoustic modeling for CNN-based speech synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6950–6954.
Chollet, F. (**2019**). keras-team/keras, https://github.com/keras-team/keras (Last viewed February 2021).
Czyżewski, A., Piotrowska, M., and Kostek, B. (**2017a**). "Analysis of allophones based on audio signal recordings and parameterization," J. Acoust. Soc. Am. **141**(5), 3521–3521.
Czyżewski, A., Kostek, B., Bratoszewski, P., Kotus, J., and Szykulski, M. (**2017b**). "An audio-visual corpus for multimodal automatic speech recognition," J. Intell. Inf. Syst. **49**, 167–192.
Dalka, P., Bratoszewski, P., and Czyżewski, A. (**2014**). "Visual lip contour detection for the purpose of speech recognition," *IEEE 2014 International Conference on Signals and Electronic Systems*, pp. 1–4.
Deng, C., Ji, X., Rainey, C., Zhang, J., and Lu, W. (**2020**). "Integrating machine learning with human knowledge," iScience **23**(11), 101656.
Dromey, C., and Black, K. M. (**2017**). "Effects of laryngeal activity on articulation," IEEE/ACM Trans. Audio Speech Lang. Proc. **25**(12), 2272–2280.
Dwarampudi, M., and Reddy, N. V. (**2019**). "Effects of padding on LSTMs and CNNs," arXiv:1903.07288.
Ge, Z., Sharma, S. R., and Smith, M. J. (**2011**). "Adaptive frequency cepstral coefficients for word mispronunciation detection," *4th International Congress on Image and Signal Processing (CISP)*, IEEE, pp. 2388–2391.
Geron, A. (**2017**). *Hands-on Machine Learning with Scikit-Learn and Tensor-Flow: Concepts, Tools, and Techniques to Build Intelligent Systems* (O'Reilly Media, Sebastopol, CA).
Giannakopoulos, T., and Pikrakis, A. (**2014**). *Introduction to Audio Analysis: A MATLAB Approach* (Academic Press, New York).
Ghosh, J. K., Delampady, M., and Samanta, T. (**2007**). *An Introduction to Bayesian Analysis: Theory and Methods* (Springer Science & Business Media, New York).
Hamooni, H., Mueen, A., and Neel, A. (**2016**). "Phoneme sequence recognition via DTW-based classification," Knowl. Inf. Syst. **48**, 253–275.
Heffner, R.-M. S. (**1950**). *General Phonetics* (University of Wisconsin Press, Madison).
Illa, A., and Ghosh, P. K. (**2020**). "Closed-set speaker conditioned acoustic-to-articulatory inversion using bi-directional long short term memory network," J. Acoust. Soc. Am. **147**(2), EL171–EL176.
Islam, J. (**2019**). *Phonetics and Phonology of 'Voiced-Pirated' Stops: Evidence from Production, Perception, Alternation and Learnability* (Georgetown University–Graduate School of Arts & Sciences, Washington, DC).
Jensen, J. T. (**2004**). *Principles of Generative Phonology: An Introduction* (John Benjamins, New York), p. 250.
Jiao, Y., Berisha, V., Liss, J., Hsu, S.-C., Levy, E., and McAuliffe, M. (**2017**). "Articulation entropy. An unsupervised measure of articulatory precision," IEEE Sign. Proc. Lett. **24**, 485–489.
Jones, D. (**1956**). "The hyphen as a phonetic sign," STUF Lang. Typol. Univ. **9**(1-4), 99–107.
Kazanina, N., Bowers, J. S., and Idsardi, W. (**2018**). "Phonemes: Lexical access and beyond," Psychon. Bull. Rev. **25**, 560–585.
Keating, P. A., Mikoś, M. J., and Ganong, W. F. III (**1981**). "A cross-language study of range of voice onset time in the perception of initial stop voicing," J. Acoust. Soc. Am. **70**(5), 1261–1271.
Keating, P., Linker, W., and Huffman, M. (**1983**). "Patterns in allophone distribution for voiced and voiceless stops," J. Phon. **11**(3), 277–290.
Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv:1412.6980.
Kim, H.-G., Moreau, N., and Sikora, T. (**2005**). *MPEG-7 Audio and beyond: Audio Content Indexing and Retrieval* (Wiley, New York).
Korvel, G., and Kostek, B. (**2017**). "Voiceless stop consonant modelling and synthesis framework based on MISO dynamic system," Arch. Acoust. **42**(3), 375–383.
Korvel, G., and Kostek, B. (**2018**). "Examining feature vector for phoneme recognition," *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, Bilbao, Spain, pp. 394–398.
Korvel, G., Treigys, P., and Kostek, B. (**2021**). "Highlighting interlanguage phoneme differences based on similarity matrices and convolutional neural network," J. Acoust. Soc. Am. **149**(1), 508–523.
Korvel, G., Treigys, P., Tamulevičius, G., Bernatavičienė, J., and Kostek, B. (**2018**). "Analysis of 2D feature spaces for deep learning-based speech recognition," J. Audio Eng. Soc. **66**(12), 1072–1081.
Korvel, G., Kurowski, A., Kostek, B., and Czyżewski, A. (**2019**). "Speech analytics based on machine learning," in *Machine Learning Paradigms. Intelligent Systems Reference Library*, edited by G. Tsihrintzis, D. Sotiropoulos, and L. Jain (Springer, Cham), Vol. 149, pp. 129–157.
Kostek, B., Kupryjanow, A., Żwan, P., Jiang, W., Raś, Z. W., Wojnarski, M., and Świetlicka, J. (**2011**). "Report of the ISMIS 2011 contest: Music information retrieval," in *International Symposium on Methodologies for Intelligent Systems* (Springer, Berlin), pp. 715–724.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**(3), 384–422.
Mikoś, M., Keating, P., and Moslin, B. (**1978**). "The perception of voice onset time in Polish," J. Acoust. Soc. Am. **63**(S1), S19–S19.
Mitterer, H., Reinisch, E., and McQueen, J. M. (**2018**). "Allophones, not phonemes in spoken-word recognition," J. Mem. Lang. **98**, 77–92.
Nahar, K. M. O., Elshafei, M., Al-Khatib, W. G., Al-Muhtaseb, H., and Alghamdi, M. M. (**2012**). "Statistical analysis of Arabic phonemes used in Arabic speech recognition," in *Neural Information Processing*, Vol. 7663 of ICONIP 2012. Lecture Notes in Computer Science, edited by T. Huang, Z. Zeng, C. Li, and C. S. Leung (Springer, Berlin).
Pandey, P. C., and Shah, M. S. (**2009**). "Estimation of place of articulation during stop closures of vowel consonant vowel utterances," IEEE Trans. Audio Speech Lang. Proc. **17**(2), 277–286.
Palaz, D., Magimai-Doss, M., and Collobert, R. (**2019**). "End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition," Speech Commun. **108**, 15–32.
Partila, P., Tovarek, J., Ilk, G. H., Rozhon, J., and Voznak, M. (**2020**). "Deep learning serves voice cloning: How vulnerable are automatic speaker verification systems to spoofing trials?," IEEE Commun. Magn. **58**(2), 100–105.
Piotrowska, M., Czyżewski, A., Ciszewski, T., Korvel, G., Kurowski, A., and Kostek, B. (**2021**). "Alofon repository corpus and extras," www.modality-corpus.org (Last viewed 6/29/2021).

Piotrowska, M., Korvel, G., Kostek, B., Rojczyk, A., and Czyżewski, A. (**2018a**). "Objectivization of phonological evaluation of speech elements by means of audio parametrization," *11th International Conference on Human System Interaction (HSI)*, pp. 325–331.

Piotrowska, M., Korvel, G., Kurowski, A., Kostek, B., and Czyżewski, A. (**2018b**). "Machine learning applied to aspirated and non-aspirated allophone classification—An approach based on audio fingerprinting," 145 Audio Engineering Society Convention, New York (October 17–20).

Plewa, M., and Kostek, B. (**2015**). "Music mood visualization using self-organizing maps," Audio Eng. Soc. Conv. Arch. Acoust. **40**(4), 513–525.

Rabha, S., Sarmah, P., and Prasanna, S. M. (**2019**). "Aspiration in fricative and nasal consonants: Properties and detection," J. Acoust. Soc. Am. **146**(1), 614–625.

Rafałko, J. (**2016**). "Algorithm of allophone borders correction in automatic segmentation of acoustic units," *145 Audio Engineering Society Convention. IFIP International Conference on Computer Information Systems and Industrial Management* (Springer, Berlin), pp. 462–469.

Recasens, D. (**2012**). "A cross-language acoustic study of initial and final allophones of/l," Speech Commun. **54**(3), 368–383.

Refaeilzadeh, P., Tang, L., and Liu, H. (**2014**). *Cross-Validation. Encyclopedia of Database Systems* (Springer, Berlin), pp. 532–538.

Rojczyk, A. (**2010**). "Preceding vowel duration as a cue to the consonant voicing contrast: Perception experiments with Polish-English bilinguals," in Issues *in Accents* English: Variability *and Norm* (Cambridge Scholars, Newcastle upon Tyne, UK), pp. 341–360.

Rojczyk, A. (**2012**). "Phonetic and phonological mode in second-language speech: VOT imitation," in *EuroSLA 22–22nd Annual Conference of the European Second Language Association*, Poznań, Poland, pp. 5–8.

Rosner, A., and Kostek, B. (**2018**). "Automatic music genre classification based on musical instrument track separation," J. Intell. Inf. Syst. **50**(2), 363–384.

Saleem, N., Irfan Khattak, N., Ali, M. Y., and Shafi, M. (**2019**). "Deep neural network for supervised single-channel speech enhancement," Arch. Acoust. **44**(1), 3–12.

Salehinejad, H., Sankar, S., Barfett, J., Colak, E., and Valaee, S. (**2018**). "Recent advances in recurrent neural networks," https://arXiv:1801.01078 (Last viewed February 2021).

Shahin, M., and Ahmed, B. (**2019**). "Anomaly detection based pronunciation verification approach using speech attribute features," Speech Commun. **111**, 29–43.

Smailis, C., Sarafianos, N., Giannakopoulos, T., and Perantonis, S. (**2016**). "Fusing active orientation models and mid-term audio features for automatic depression estimation," in *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, 39.

Tsipas, N., Vrysis, L., Dimoulas, C., and Papanikolaou, G. (**2015**). "Methods for Speech/Music Detection and Classification," in *Proceedings of MIREX 2015*.

Tsipas, N., Vrysis, L., Konstantoudakis, K., and Dimoulas, C. (**2020**). "Semi-supervised audio-driven TV-news speaker diarization using deep neural embeddings," J. Acoust. Soc. Am. **148**(6), 3751–3761.

Vrysis, L., Tsipas, N., Thoidis, I., and Dimoulas, C. (**2020**). "1D/2D deep CNNs vs. temporal feature integration for general audio classification," J. Audio Eng. Soc. **68**(1/2), 66–77.

Vryzas, N., Kotsakis, R., Liatsou, A., Dimoulas, C., and Kalliris, G. (**2018**). "Speech emotion recognition for performance interaction," J. Audio Eng. Soc. **66**(6), 457–467.

Waniek-Klimczak, E. (**2005**). *Temporal Parameters in Second Language Speech: An Applied Linguistic Phonetics Approach* (Łódź University Press, Poland).

Wei, S., Hu, G., Hu, Y., and Wang, R.-H. (**2009**). "A new method for mispronunciation detection using support vector machine based on pronunciation space models," Speech Commun. **51**(10), 896–905.

Woore, R. (**2018**). "Learners' pronunciations of familiar and unfamiliar French words: What can they tell us about phonological decoding in an L2?," Language Learn. J. **46**(4), 456–469.

Yu, J., Markov, K., and Matsui, T. (**2019**). "Articulatory and spectrum information fusion based on deep recurrent neural networks," IEEE/ACM Trans. Audio Speech Lang. Proc. **27**(4), 742–752.