Audio Engineering Society

Convention Paper 10070

Presented at the 145th Convention
2018 October 17–20, New York, NY, USA

Machine Learning Applied to Aspirated and Non-aspirated Allophone Classification – an Approach Based on Audio "Fingerprinting"

Magdalena Piotrowska[1], Grazina Korvel[2], Adam Kurowski[1], Bożena Kostek[3] and Andrzej Czyżewski[1]

[1] *Multimedia System Department, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk 80-233, Poland, {mplewa, adakurow, andcz}@sound.eti.pg.gda.pl*

[2] *Institute of Data Science and Digital Technologies, Vilnius University, Akademijos str. 4, LT-04812, Vilnius, Lithuania, grazina.korvel@mii.vu.lt*

[3] *Audio Acoustics Laboratory, Telecommunications and Informatics, Gdansk University of Technology, Gdansk 80-233, Poland, bokostek@audioacoustics.org*

Correspondence should be addressed to Bożena Kostek (bokostek@audioacoustics.org)

ABSTRACT

The purpose of this study is to involve both Convolutional Neural Networks and a typical learning algorithm in the allophone classification process. A list of words including aspirated and non-aspirated allophones pronounced by native and non-native English speakers is recorded and then edited and analyzed. Allophones extracted from English speakers' recordings are presented in the form of two-dimensional spectrogram images and used as input to train the Convolutional Neural Networks. Various settings of the spectral representation are analyzed to determine adequate option for the allophone classification. Then, testing is performed on the basis of non-native speakers' utterances. The same approach is repeated employing learning algorithm but based on feature vectors. The achieved classification results are promising as high accuracy is observed.

1 Introduction

Speech recognition is still the most important research related to human-machine communication. It is predicted that approximately 50% of search queries will be provided with voice in the future, the same concerns communication with a mobile phone, laptop, machine, robot, etc. However, before this aim is achieved, more research is needed that includes all aspects of speech production and perception. Such analyses may also be supported by machine learning as a tool for human assisted evaluation, e.g. applications checking the subject's pronunciation. These aspects refer to processes such as speech recognition based on phonemes or, even more necessarily, on allophone models being a foundation of words and sentences properly synthesized. The presented study is dedicated to recognition of aspirated and non-aspirated allophones. An allophone may be defined as a semantically non-contrastive positional or contextual variant of a particular phoneme. It should be noted that the phonetic phenomenon of aspiration is selected as particularly difficult to Polish learners of English. Therefore, a set of English words, selected with regard to aspiration phenomena particularly difficult for Polish learners, was recorded. Recordings encompass speech of 9 native English speakers and 9 non-native English speakers.

The proposed approach involves both convolutional neural networks (CNN) classification based on spectrograms, processed for that purpose and exported as images to the CNN input and a learning algorithm utilizing feature vectors. For the CNN-based classification a two-dimensional representation of speech feature space is employed and makes the allophonic samples more populated than in the case of one-dimensional feature vectors. It should be remembered that machine learning needs pre-processing and parameterization as the first steps of speech recognition processes. Therefore, this study addresses these issues by demonstrating methods of extracting a two-dimensional representation of speech dedicated to allophone evaluation, even though CNNs may act as a feature extractor themselves. We propose an approach that may be referred as "fingerprinting", used with success within the Music Information Retrieval (MIR) field, mainly for music or music genre recognition. Contrarily, for the learning algorithm audio parameters based on the MPEG 7 standard as well as features found in the MIR area are utilized.

The results included in this work can support English language education within tools that would automatically evaluate quality pronunciation with a focus on particular phenomena and related allophones. Recent studies on automatic quality of allophones/phonemes evaluation utilize speech recognition technology, (i.e. creating speech databases, feature extracting, machine learning) [1-5] and to less extent concentrate on the phenomenon mechanism.

In our study we decided to take a hybrid approach to feature extraction and machine learning as a tool for human-assisted evaluation. Using various audio features to determine whether allophone is pronounced correctly or not may potentially lead to identification what causes the mistake. Objectivization of the phonology phenomena evaluation results by correlating it with feature vectors may also enable to automatically recognize proper/improper target pronunciation.

2 Allophonic material

In the presented study, authors focus on the phenomenon of aspiration. Aspiration of voiceless stop consonants is represented using the parameter of VOT (Voice Onset Time), which is defined as the time interval between the burst that marks the stop release and the onset of periodicity that reflects laryngeal vibration [6]. Polish and English differ in their implementation of VOT in cuing the contrast between /b, d, g/ and /p, t, k/. Polish uses pre-voicing or negative VOT values for voiced /b, d, g/ and short-lag VOT values for voiceless /p, t, k/ [7]. On the other hand, English contrasts short-lag VOTs for voiced /b, d, g/ and long-lag VOTs for voiceless /p, t, k/ [6,8]. Long VOTs in English voiceless stops are temporal representation of what is traditionally known as aspiration. Polish learners transfer pronunciation habits from their native language and do not produce sufficiently long VOT in English /p, t, k/ [9]. 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. Aspiration occurs immediately before a stressed vowel; therefore information regarding energy distribution within the allophone is crucial. This information can potentially be extracted from visual representation such as spectrograms utilized in the presented study.

The experiment performed consisted of two stages. In **part I** a set of 30 words including aspirated and non-aspirated variants of plosive phonemes /p, t, k/ was used. Nine native speakers (6 male and 3 female) were asked to pronounce the words. In addition, parameterization of the recorded speech signals is performed using selected features.

Additionally in **part II** six words including aspirated /p, t, k/ was recorded by nine non-native speakers. Collected data are evaluated by the phonology expert with regard to the aspiration phenomena. Allophones, edited from speech of non-native English speakers are also evaluated by the phonology expert to check whether results of automatic evaluation based on the machine learning meet subjective expertise. The second set was collected to investigate the predictive properties of the implemented analyses.

3 Experiment I

Different acoustic features have been proposed to separate speech units. Based on our experiments carried out in the context of allophones, we can state that using standard speech parameters along with descriptors from the music area, the phoneme recognition accuracy is better, regardless of singular and specific features of voice exhibited by a speaker [10,11]. That is why the MPEG 7 standard-based features, as well as features derived from the MIR domain, are used in this research [12,13]. As already mentioned, aspiration occurs immediately before a stressed vowel. Therefore information regarding energy distribution within the allophone is crucial. Our initial investigations of phonological processes show that in order to determine aspiration of voiceless stop consonants, the parameters include energy measures of temporal distribution should be used [14]. The description of the features chosen for this research is included in the summary below. A list of these features is given in Table 1.

1	Number of samples
2	RMS Energy
3	Temporal Centroid (TC)
4	Number of samples exceeding RMS
5	Number of samples exceeding $2\times RMS$
6	Number of samples exceeding $3\times RMS$

Table 1. A list of features used for the automatic evaluation of aspiration.

The parameters no. 1-3 refer to the time domain and are widely used in the speech analysis. The first of them (number of samples) indicates the number of the samples included in the allophone. Root Mean Square (RMS) Energy gives a mean energy in the analyzed signal frame. Temporal Centroid is the time average over the signal energy envelope. The parameters no. 4-6 are also time domain representation but they are not extensively used because they were invented quite recently. These dedicated parameters proposed by Kostek and her coworkers [13] are the number of samples exceeding levels RMS, 2xRMS, 3xRMS.

Before the feature calculation starts, the speech signal is divided into short-time segments, the length of which is an integer power of 2. We use this approach, which is typical for an audio analysis and signal analytics, in order to get more accurate information.

Techniques of allophone classification

Denote by L – the number of allophones in training set. Let

$$x_l = \{x_{l1}, x_{l2}, ..., x_{ld}\} \quad (1)$$

be the d-dimensional feature vector of the l-th allophone *(l=1...L),* and

$$c = \{c_1, c_2, ..., c_K\} \quad (2)$$

the set of class labels to which these allophones belong.

Consider a feature vector of the test allophones set which does not have a class label:

$$x = \{y_1, y_2, ..., y_d\} \quad (3)$$

Our goal is to build a classifier to assign its unknown class label. In the experiment, two classical machine learning algorithms to compare classification rates are used. First of them is the Naive Bayes classification method based on Bayes theory [16]. The probability that a phoneme with feature vector y belongs to class c_k *(k=1...K)* is:

$$P(c_k \mid y) = \frac{P(y \mid c_k)P(c_k)}{P(y)} \quad (4)$$

As we see from Eq. (4), the class condition density *P(y|c_k)* needs to be estimated. In this paper, Gaussian kernel density estimation is used. According to this theory, we obtained that:

$$P(c_k \mid y) = \prod\nolimits_{j=1}^{d} \left(\frac{1}{Nh} \sum\nolimits_{i=1}^{N} \frac{1}{\sqrt{2\Pi}} e^{-1/2(\frac{y_j - x_{ij}}{h})} \right) \quad (5)$$

where h is the bandwidth for control of the smoothness of the density curve [17], N – the number of phonemes in class c_k. The test data are assigned to the class with the maximum class probability.

The second classification algorithm employed is *k*-Nearest Neighbors (*k*NN) [18], based on Euclidean distances among the elements of the test and training datasets. The Euclidean distance between the l-th allophone of training set and the unlabeled allophone of the test set is defined as:

$$P(y, x_l) = \sqrt{\sum\nolimits_{i=1}^{d} (x_{li} - y_i)^2} \quad (6)$$

The optimum number of nearest neighbors is established by performing a series of preliminary tests.

The feature vector listed in Tab. 1 was calculated for each analyzed sample. The average parameter values normalized to the range [0,1] are included in Fig. 2 for aspirated and non-aspirated allophones.

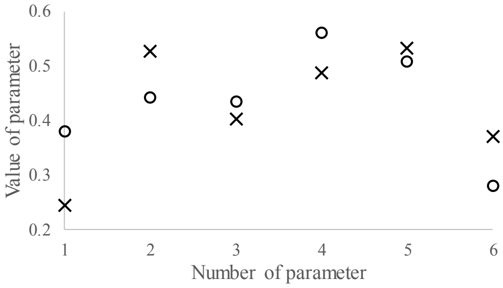

Figure 1. The normalized values of parameters for aspirated and non-aspirated allophones averaged for all speakers (aspirated allophones are marked with a circle, non-aspirated ones are denoted with a cross)

The feature automatic classification was performed. Naive Bayes and *k*NN algorithms to compare classification rates were used. In order to verify the statistical significance of our calculations the cross-validation technique is used. The data are partitioned into 9 equally sized segments, each consisting of one speaker's utterances. In the process, 9 iterations are performed. Within each iteration, the training of classifiers on the allophones of 8 speakers is performed, and estimation of system accuracy on the basis of the speech of the held-out speaker is obtained. The average results of allophones classification with regard to the aspiration for all plosive allophones and each allophone group separately are presented in Table 2.

Allophone	Naive Bayes		*k*NN	
	Mean	STD	Mean	STD
/p/	87.78	13.944	83.33	14.142
/t/	**90.00**	13.229	**91.11**	11.667
/k/	86.67	14.142	88.89	13.642
All	87.78	11.902	90. 37	9.196

Table 2. The results of accuracy [in %] of classification of aspirated and non-aspirated allophones.

A possibility of classification of detection of aspirated vowels with the use of convolutional neural networks (CNN) was the next step of the research. An architecture proposed in the previous work of authors was employed for this task [19]. An architecture of the network was optimized by adding layers of networks until the speed of convergence measured in terms of speed of the loss decreasing rate, where the loss is a measure of an error between desired answers and ones obtained from the CNN. Additionally, dropouts were used to increase generalization abilities of the network [20]. Adam optimizer algorithm was employed as a learning rate adjustment procedure [21]. The input data were 270 samples containing allophones with and without aspiration (135 examples of each). Examples were pre-processed by calculation of spectrograms. Each of examples was padded to the length of samples. Spectrogram was calculated with the use of Hamming window and with the overlap factor of 0.75. The length of a single frame assigned to a time step of a spectrogram was 128 samples. The input for the network consisted of two channels. The one calculated as an absolute value of a spectrogram (amplitude spectrogram denoted as s_{abs}) calculated as

$$s_{abs} = abs(s) \quad (7)$$

and a value of phase (phase spectrogram denoted as s_{phase}) calculated according to the following formula:

$$s_{phase} = \tan^{-1}(s) \, . \quad (8)$$

In both formulas s denotes complex-valued spectrogram. Then, both spectrograms were normalized by division of their standard deviation and subtraction of their mean value. An example of a pair of spectrograms is depicted in the Fig. 3.

The network was trained for 20 epochs and 200 examples were used as a training set and 70 were used as a validation set. The process of training was repeated 100 times to derive mean measures of performance. Learning rate was set to *1e-3*, the batch size was 20. Examples were assigned to training and validation set in a random way. To assess the uncertainty of outcomes, a confidence interval for mean accuracy of classifier was calculated according to the following formula

$$f_{conf} = \mathrm{t}_{\alpha/2}(n-1)\frac{\sigma}{\sqrt{n}} \, , \quad (9)$$

where f_{conf} is a confidence coeficient, α is a significance level which in case of our study was assumed to be 0.05, σ is standard deviation and n is a number of samples used for calculation of the mean.

A mean confidence interval can be then calculated from the following equation

$$x_m \in \bar{x} \pm f_{conf'} \qquad (10)$$

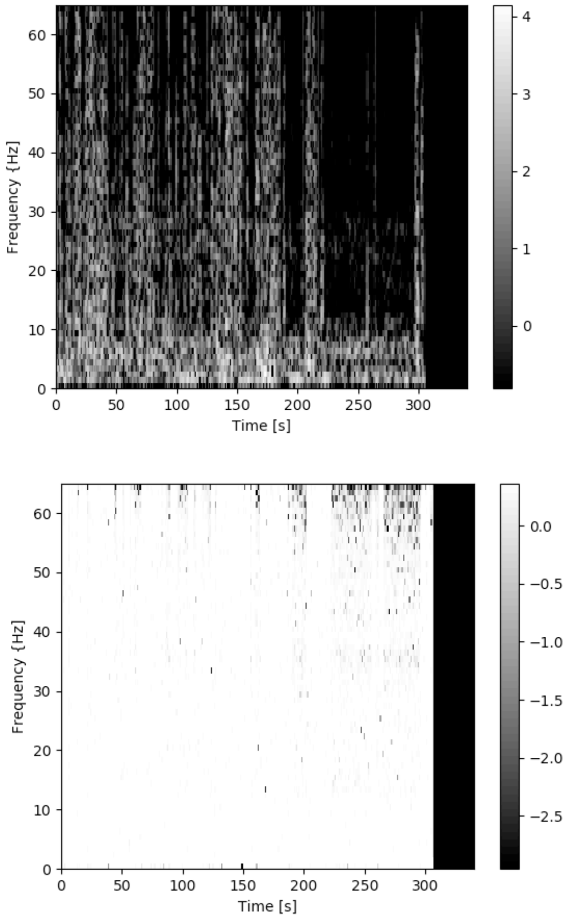

Figure 3. A pair of spectrograms calculated for one of the examples from the database of allophones.

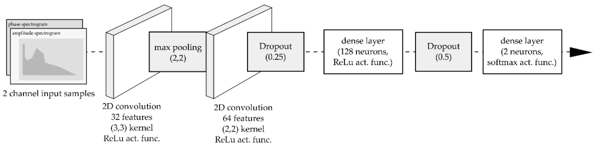

Figure 4. The structure of the CNN used in our study.

where x_m is a true mean of variable x and $\bar{x}$ is an estimate of a mean calculated as an arithmetic mean of samples – in our case, accuracies from 100 allophone classification experiments.

The mean accuracy obtained in experiment repeated 100 times is 0.844. The confidence interval calculated for this case is (0.818; 0.869), the standard deviation is equal to 0.128. The worst accuracy was 0.414 and the highest value of accuracy was 0.986.

4 Experiment II

To further investigate the predictive properties of an automatic classifiers, classification of another set of 114 examples of allophones was performed. All of them were examples of allophones with aspiration. Occurrence of aspiration was also subjectively assessed by a group of phonology experts.

The result of allophone aspiration recognition based on feature extraction and machine learning approach is given in Table 3.

	Naive Bayes	*k*NN
Accuracy	65	**69**

Table 3. The results of automatic evaluation of aspirated allophones pronunciation, in [%].

We also tried to employ CNN model used for classification to perform analogous task and discern examples in which aspiration was really introduced by a speaker from ones which do not have such.

The CNN classification was performed 10 times with CNNs connected to accuracies greater than 92% and 10 networks of such type were used. The only modification with respect to the previous design of network was lowering the learning rate to the value of 1e-4 and increasing number of learning epochs from 20 to 50. Then, a percentage of classification errors was assessed. Results were analyzed with the use of formulas 9 and 10. Also, Pearson correlation factor was calculated for a series of answers obtained from CNNs and given by the experts. The calculation was performed accordingly to the formula 11 [22].

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} \qquad (11)$$

where *r(x,y)* is a Pearson correlation factor calculated for a pair of vectors *x* and *y*. Standard deviations of these vectors are denoted as σ_x and σ_y respectively.

The CNN network error rate was on average 41.6% when compared to answers provided by experts. The confidence interval for the error rate is correspondingly 39.8% and 43.4%, and for correlation between CNN and experts' answers it is 0.076 and 0.170. The mean accuracy of the base dataset is 91.4%.

5 Conclusions

Automatic classification of aspirated and non-aspirated /p, t, k/ allophones returned good results. High accuracy (91%, 90% and 84%) was achieved for all tested methods (respectively for kNN, Naive Bayes and CNN).

The analysis presented in this paper shows the potential for automated evaluation of pronunciation focused on phonological aspiration challenge for non-native speakers.

We achieved accuracy of 69% with *k*NN in the second part of the experiment, where we aimed for automatic evaluation of the correctness of aspiration for non-native speakers. The CNN was able to label correctly approximately 60% of examples in the second stage of experiment, however more research performed on larger database is needed in future to fully explore possibilities of such classification. The architecture of network presented in our paper may however serve as a starting point for such research.

All these remarks can lead to the conclusion that the approach proposed can be utilized in the automatic recognition of allophones and in future can be potentially used for automatic pronunciation evaluation.

Acknowledgments

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

References

[1] Z. Ge, S. R. Sharma and M. J. Smith, "Adaptive frequency cepstral coefficients for word mispronunciation detection", in Image and Signal Processing (CISP), 2011, IEEE 4th International Congress on, vol. 5, pp. 2388-2391 (2011).

[2] A. Czyżewski, M. Piotrowska, and B. Kostek, "Analysis of allophones based on audio signal recordings and parameterization", The Journal of the Acoustical Society of America 141 (5), pp. 3521-3521 (2017).

[3] P. Dalka, P. Bratoszewski and A. Czyżewski, "Visual Lip Contour Detection for the Purpose of Speech Recognition", In *Signals and Electronic Systems (ICSES),* IEEE *2014 International Conference on*, pp. 1-4 (2014).

[4] S. Wei, G. Hu, Y. Hu and R. Wang, "A new method for mispronunciation detection using support vector machine based on pronunciation space models", Speech Communication, vol. 51, no. 10, pp. 896-905 (2009).

[5] G. Korvel and B. Kostek, "Voiceless Stop Consonant Modelling and Synthesis Framework Based on MISO Dynamic System", Archives of Acoustics, 3, 42, pp. 375 - 383, 10.1515/aoa-2017-0039 (2017).

[6] L. Lisker and A. S. Abramson, "A cross language study of voicing in initial stops: Acoustic measurements. *Word* 20, pp. 384-422 (1964).

[7] Mikoś, J. L. P. A. Keating and B. J. Moslin, "The perception of voice onset time in Polish", *Journal of the Acoustical Society of America* (S1), 63, S19 (1978).

[8] P. A. Keating, W. Linker and M. Huffman, "Patterns of allophone distribution for voiced and voiceless stops", *Journal of Phonetics* 11, pp. 277-290 (1983).

[9] E. Waniek-Klimczak, *Temporal parameters in second language speech: An applied linguistics phonetics approach*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego, (2005).

[10] G. Korvel, B. Kostek, "Examining Feature Vector for Phoneme Recognition", Proceeding of IEEE International Symposium on Signal Processing and Information

Technology, ISSPIT 2017 – Bilbao, Spain, pp. 394-398 (2017).

[11] G. Korvel, A. Kurowski, B. Kostek, A. Czyzewski, "Speech Analytics Based on Machine Learning", Machine Learning Paradigms, Springer, Cham, 129-157 (2019).

[12] J. Pohjalainen, O. Räsänen, S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits", Computer Speech & Language 29 (1), 145-171 (2015).

[13] B.Kostek, A. Kupryjanow, P. Zwan, W. Jiang, Z. Raś, M. Wojnarski, J. Swietlicka, "Report of the ISMIS 2011 contest: music information retrieval", Foundations of Intelligent Systems, 715-724 (2011).

[14] K. Hyoung-Gook, N. Moreau, T. Sikora, "MPEG-7 Audio and Beyond: Audio Content Indexing and Retrieval", Wiley & Sons (2005).

[15] M. Piotrowska, G. Korvel, B. Kostek, A. Rojczyk, A. Czyżewski, "Objectivization of phonological evaluation of speech elements by means of audio parametrization", 11th International Conference on Human System Interaction, July 04-06, Gdańsk, Poland (2018).

[16] J. K. Ghosh, M. Delampady, T. Samanta, "An introduction to Bayesian analysis: theory and methods", 1st edn, Springer Science Business Media, LLC (2006).

[17] S. T. Chiu, "Bandwidth Selection for Kernel Density Estimation S. T., The Annals of Statistics, Vol. 19, No. 4 (Dec., 1991), pp. 1883-1905 (1991).

[18] S. Manocha, M. A. Girolami, "An empirical analysis of the probabilistic K-nearest neighbour classifier", Pattern Recognition Letters, 28, 1818–1824 (2007).

[19] G. Korvel, A. Kurowski, B. Kostek, A/ Czyzewski, "Speech Analytics Based on Machine Learning", Machine Learning Paradigms, Springer, Cham, 129-157 (2019).

[20] N. Buduma, "Fundamentals of Deep Learning. Designing Next-Generation Machine Intelligence Algorithms", O'Reilly Media. (2017).

[21] P. D. Kingma. J. L. Ba, "ADAM: A Method For Stochastic Optimization", International Conference on Learning Representations, ICLR 2015 https://arxiv.org/pdf/1412.6980.pdf (accessed Jan 2018) (2015).

[22] Mason, R., Gunst, R., Hess, J., "Statistical Design and Analysis of Experiments: With Applications to Engineering and Science", Wiley (2003).