Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220500

Support-Vector Machine-Based Classifier of Cross-Correlated Phoneme Segments for Speech Sound Disorder Screening

Emilian-Erman MAHMUT[a,1], Stelian NICOLA[a] and Vasile STOICU-TIVADAR[a]
[a] *Politehnica University Timisoara, Romania*
Department of Automation and Applied Informatics

Abstract. This paper presents a Support-Vector Machine (SVM) based method of classification of cross-correlated phoneme segments as part of the development of an automated Speech Sound Disorder (SSD) Screening tool. The pre-processing stage of the algorithm uses cross-correlation to segment the target phoneme and extracts data from the new homogeneously trimmed audio samples. Such data is then fed into the SVM-based classification script which currently achieves an accuracy of 97.5% on a dataset of 132 rows. Given the global context of an increasing trend in the incidence of Speech Sound Disorders (SSDs) amongst early-school aged children (5-6 years old), the constraints imposed by the new Corona virus pandemic, and the (consequent) shortage of professionally trained specialists, an automated screening tool would be of much assistance to Speech-Language Pathologists (SLPs).

Keywords. Speech Sound Disorders, Support-Vector Machine, cross-correlation

1. Introduction

Phonemes are the smallest speech sound units capable of changing meaning in a language. They play a central role in the structure of speech and have the potential to transcend all linguistic tiers of human languages, from phonetics all the way through to semantics and beyond (pragmatics/phraseology). Phonemes are a major acquisition in the early stages of both generating and decoding human speech. As substantiated in papers [1] and [2], it is highly likely that such stages rely on statistical models.

Phonological assimilation is a phenomenon whereby in any given word, the pronunciation of a phoneme is affected by the neighboring phonemes. It produces its effects both progressively, when the phoneme *n* is affected by the phoneme *n+1*, and regressively, i.e., the phoneme *n+1* is affected by the phoneme *n*. Mathematical modelling of speech sounds can be particularly challenging given that regressive phonological assimilation seems to operate backwards in time. Paper [2] hints at infants (and adults) being able to compute “both forward and backward conditional probabilities to segment continuous artificial streams” and the neuroscience perspective [3] confirms

[1] Emilian-Erman Mahmut, Politehnica University Timisoara, P-ta Victoriei no. 2, Timisoara, Romania; E-mail: emilian.mahmut@aut.upt.ro

such considerations and acknowledges the shortcomings of the conventional fixed frame size and rate segmentation technique commonly used in automatic speech recognition (ASR).

Paper [4] provides a presentation of the general framework of automated speech segmentation and a thorough review of various segmentation algorithms and feature extraction techniques.

According to the criteria initially adopted for its development, the SSD screening tool should: be language independent (mathematical model), provide real-time or near-real-time feedback (low computational cost), use open-source software (low financial cost), be web-based, and use mobile technology, grant easy and open access to SLPs/researchers and subjects' parents. Similar research projects aimed at providing an automated phoneme classification [5,6,7], reviewed in paper [8] reported classification accuracy rates roughly comprised between 78 and 86%.

This paper presents a method for the automated SSD screening in early-school aged subjects, which generates better results than the known literature.

2. Method

The diagram below provides a synthesis of the 3-stage automated processing of the .wav files: the pre-processing stage consisting of the cross-correlation (X-corr) bases segmentation followed by the feature extraction stage, and the SVM-based classification stage (Figure 1). Python scripts were written to perform the processing stages described above.

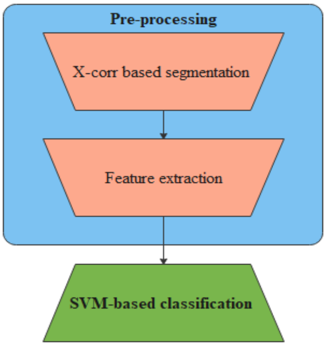

Figure 1. Processing diagram

2.1. Pre-processing

2.1.1. X-corr based segmentation

A total number of 132 .wav files were processed: subjects' pronunciations of words containing the target phoneme /r/ in medial position. The subjects were 5–6-year-old pupils whose pronunciations were recorded during the traditional screening process performed by the SLP of the CNB school in Timisoara (Romania) in 2019.

While segmenting (and extracting features from) a word-initial or a word-final phoneme provides a natural boundary – because of the silence before and after the word

– a rigorous segmentation of a medial target phoneme entails a more complex approach. The words fed into the algorithm were *mere*, *pere*, and *mură* (Romanian words for *apples*, *pears*, and *blackberry*), containing the /r/ phoneme in between vowels. The algorithm was devised to generate homogeneous medial-phoneme segments by cross correlating each of the subjects' utterances with the SLP (reference) pronunciation and it is presented in detail in paper [9]. Providing optimal alignment of the 2 audio signals, cross-correlation solves the issue of the variable length of a pronunciation of the same word by different people. It already starts on the classification problem since it generates homogeneously trimmed target-phoneme segments, and it ensures a smooth transition into the classification stage. The features extracted from the target-phoneme segments, briefly described in the following subsection, are used as input for the classification stage.

2.1.2. Feature extraction

The pre-processing stage results in a 10-column table containing: *Max_left, Max_right, Left_index, right_index, Maxamp_SLP, Maxamp_SUB, Infl_SLP, Infl _SUB, R-squared, Coefficient of determination* (Table 1a and 1b). An 11th column was added to the table to indicate the SLP Opinion expressed as boolean label. *Max_left/Max_right* is the maximum cross-correlation coefficient found between the 2 audio signals by shifting the sample signal (SLP) to the left and, respectively, to the right of the reference signal (Subject). *Left_index/Right_index* are the indices where the maximum cross-correlation coefficient was found. *Maxamp_SLP* and *Maxamp_SUB* report the maximum amplitude value detected in the SLP and the Subject pronunciation of the cross-correlated phoneme segments. *Infl_SLP/Infl_SUB* indicate the total number of inflections (Gaussian smoothing, 5% standard deviation) in each of the cross-correlated target-phoneme segments.

2.2. SVM-based Classification

The features of interest are extracted into a .csv file and are divided into 2 categories: training data and test data. A choice was made to assign 70% of such data to the training data category while the remaining 30% are designated as test data. The main libraries used in the Python script are *pandas*, *numpy*, *sklearn*, and *matplotlib*.

Table 1a. Feature extraction table (first 5 columns)

Subject No.	Max_left	Max_right	Left_index	Right_index	Maxamp_SLP
1	15.57	**31.11**	-1865	998	4503.00
2	17.91	**18.41**	-418	141	7328.00
3	**20.60**	19.22	-99	93	8128.00
4	**52.30**	37.35	-1493	172	7982.00
…					
131	48.35	**74.14**	-503	1755	5469.00
132	**57.74**	42.73	-4836.00	5077.00	6839.00

Table 1b. Feature extraction table (last 6 columns)

Maxamp_SUB	Infl_SLP	Infl_SUB	R^2	Coefficient of determination	SLP Opinion
3337.00	219.00	160.00	0.99	0.07	0
3682.00	168.00	168.00	0.98	0.96	0
10089.00	189.00	185.00	0.99	0.99	1
7765.00	180.00	188.00	0.99	0.99	1

…					
8044.00	174.00	200.00	0.99	0.90	0
12518.00	172.00	182.00	0.99	0.98	1

The Python SVM-based classification algorithm processes the 2 data categories and reports the following metrics: accuracy, precision, recall, and the F1 score.

3. Results

In 68 target-phoneme segments, the segmentation algorithm detected the maximum cross-correlation coefficient (*Max_left*) by shifting the sample signal (Subject's pronunciation) to the left of the reference sample (SLP's pronunciation) while the remaining 64 phoneme segments were generated by shifting the sample signal to the right of the reference sample. The phonetical implication is that 68 subjects needed more time to transition from the preceding vowel and achieve the full vibration of the liquid consonant /r/ in correlation with the reference signal. The other 64 subjects needed less time, as compared to the reference signal, to transition to the target phoneme. Table 2 below illustrates the classification metrics for the 132 .wav files analyzed in this paper.

Table 2. SVM metrics

Accuracy	Precision	Recall	F1 score
0.975	1.00	0.928	0.962

To better illustrate the results, a Confusion Matrix was generated (Figure 2).

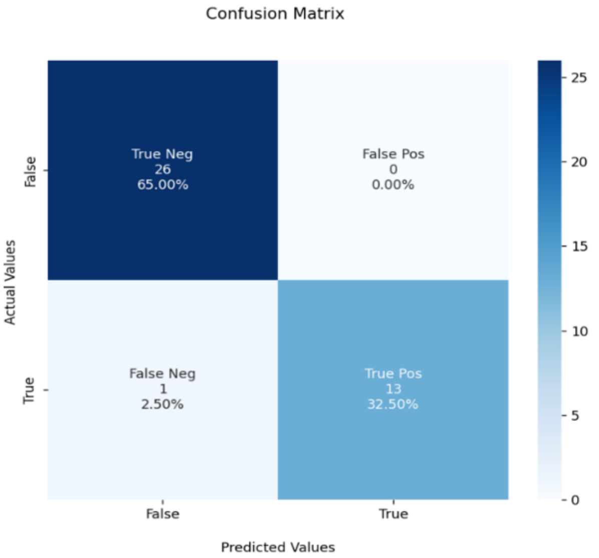

Figure 2. Confusion Matrix

As synthesized in the confusion matrix, 65% of the test data were classified as true negatives (TN), 32.5% are true positives (TP), 2.5% are false negatives (FN) and there were no false positives (FP).

4. Discussion and conclusions

Early diagnosis of SSDs (before the age of 7) is crucial since it allows for speech therapy sessions to be administered with optimal results, preventing therefore the onset of stigma that may lead to serious behavioral issues in early school-aged subjects [10]. An unbiased verdict and the anonymity provided by a web-based/mobile-technology solution would help to tackle uninformed parents' adversity to speech therapy/pathology. The current results are promising, the accuracy achieved is better than the rates reported in the literature.

The *keras* library (TensorFlow framework) was used to configure an artificial neural network in order to compare and validate the metrics described above. Such neural network requires a lot more data rows (in the order of thousands) than the ones currently available so as to generate reliable results. Paper [11] contains a detailed discussion on classification algorithms and proposes a voting classifier to aggregate the predictions of a range of individual classifiers. Other phonemes in various vowel/consonant contexts are being segmented and will be fed both into the SVM-based and into the neural network-based classification scripts.

The final objective of the research project is the development of an application meant to assist the SLPs in their laborious and time-consuming activity, to build a database for interdisciplinary research, and to grant parents access to an unbiased SSD screening tool.

References

[1] Gervain J. Neurocognitive Development: Normative development. In: Handbook of clinical neurology. 2020.

[2] Federmeier K, Huang H-W. Adult and second language learning. In: The psychology of learning and motivation; 2020 May. vol. 72.

[3] Lee B, Cho KH, Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. Scientific Reports 6, article number: 37647; 2016 Nov.

[4] Alaa ES, Sherif MA, Salah EH, Mohsen R. A Review: Automatic speech segmentation, International Journal of Computer Science and Mobile Computing, 2017 Apr; 6(4):308-15

[5] Grigore O, Grigore C, Velican V. Intelligent system for impaired speech evaluation. Recent Advances in Circuits, Systems and Signals. Book Series: International Conference on Circuits Systems Signals, 2010; Malta; p. 365-8.

[6] Grigore O, Velican V. Self-organizing maps for identifying impaired speech. In: AECE (Advances in Electrical and Computer Engineering); Romania; 2011 Jan; 11(3): 41-8.

[7] Yin S-C, Rose R, Saz O, Lleida E. A study of pronunciation verification in a speech therapy application. In: IEEE International Conference on Acoustics, Speech and Signal Processing; 2009 Apr 19-24; Taipei, Taiwan; p. 4609-12.

[8] Mahmut EE, Stoicu-Tivadar V. Current challenges in the computer-based assessment of speech sound disorders. In: IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI); 2018 May 17-19; Timisoara, Romania; p. 431-6.

[9] Mahmut EE, Nicola S, Stoicu-Tivadar V. Cross-correlation based automatic segmentation of medial phonemes. In: International Symposium on Electronics and Telecommunications (ISETC); 2020 Nov 5-6; Timisoara, Romania; p. 1-4.

[10] Shahin M, Ahmed B, Smith DV, Duenser A, Epps J. Automatic screening of children with speech sound disorders using paralinguistic features. In: IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP); 2019. p. 1-5.

[11] El-Kenawy E-S M, Ibrahim A, Mirjalili S, Eid MM, Hussein SE, Novel feature selection and voting classifier algorithms for COVID-19 classification in CT images. In: IEEE Access; 2020. vol. 8, p. 179317-35.