VPS VIBRATIONS IN PHYSICAL SYSTEMS

DOI: 10.21008/j.0860-6897.2024.1.01
Vibrations in Physical Systems, 2024, 35(1), 2024101

Acoustic model for the classification of Polish vowels

Karolina PONDEL-SYCZ

Warsaw University of Technology, Nowowiejska 15/19, 00-665 Warszawa

Corresponding author: Karolina PONDEL-SYCZ, email: karolina.pondel.dokt@pw.edu.pl

Abstract The study explored the performance of vowel recognition using an acoustic model built on Audio Fingerprint techniques [1]. The research compares the performance of Support Vector Machines (SVMs), Hidden Markov Models (HMMs), Artificial Neural Networks (ANNs) and k-Nearest Neighbours (k-NN) classifiers in the recognition of isolated and within-word vowels and investigates the importance of different types of acoustic speech features in this process. Temporal, spectral, cepstral, formant, LPC and perceptual features of speech were examined. Importance of features was tested using a random forest classifier. Vowel classification was tested at three confidence levels for feature importance: 90%, 95% and 99%. Two author databases consisting of a total of 1,200 samples from 20 speakers, recorded under household conditions, were used. The classifiers were evaluated by confusion matrix, accuracy, precision, sensitivity and F1 score. A segmentation of words into speech sounds was carried out using a tool based on BiLSTM recurrent neural networks and the BIC criterion. Three most important features were determined: power spectral density, spectral cut-off, and Power-Normalised Cepstral Coefficients. In the isolated vowel recognition task, the SVM classifier was the most effective with a feature significance confidence level of 95% obtaining accuracy = 81%, precision = 81%, sensitivity = 81%, F1 score = 80%. In the task of recognising a vowel within a word, it was verified if the algorithm detected the presence of vowels in the correct segment and if it recognised the correct vowel within it. The best results were obtained by the k-NN classifier (statistical confidence level of feature importance of 99.9%). However, these results were low, correct recognition of the vowel in the word: A, E, U: 20%, I, O: 7%, Y: 23%. This indicates strong influence of the neighbourhood of other speech sounds in speech on the acoustic model of vowels and their recognition.

Keywords: ASR, MFCC, PNCC, HMM, SVM, ANN, k-NN.

1. Introduction

Among automatic speech recognition (ASR) systems, there are two main architectures: conventional and End-To-End (E2E). The conventional approach is based on acoustic (AM) and language (LM) models and a lexicon. In the E2E approach, recognition is carried out by a single, integrated model built on deep neural networks. Parameters of the speech signal in such model are given to the input and on the output is a speech transcription. In conventional architecture, AM and LM can also be deep models, but existing solutions for Polish are mainly based on Hidden Markov Models (AM) and statistical methods (LM). In each of these architectures, parametrisation of the speech signal is required. In E2E systems, the most common approach is to represent the speech signal as a mel-spectrograms. In conventional systems, wide range of speech features of the speech signal are used. This paper describes an acoustic model of speech containing vowels in Polish, composed of various features, used for classification with the decoders widely used in conventional models: Hidden Markov Models (HMM) [2], Support Vector Machine (SVM) [3], artificial neural networks (ANN) [4] and k-nearest neighbours (k-NN) [5].

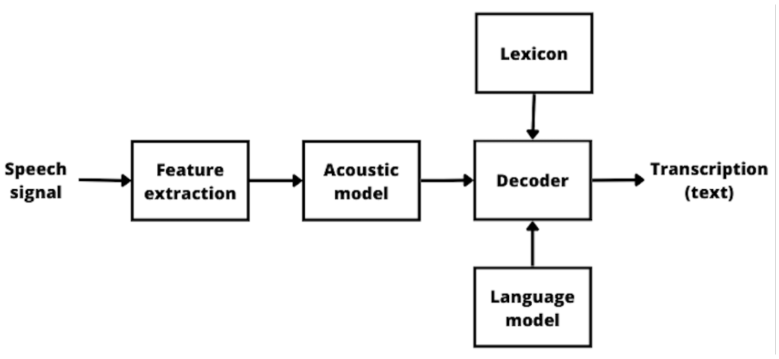

Figure 1. A block diagram of the workflow of a conventional automatic speech recognition system based on an acoustic, language and pronunciation model.

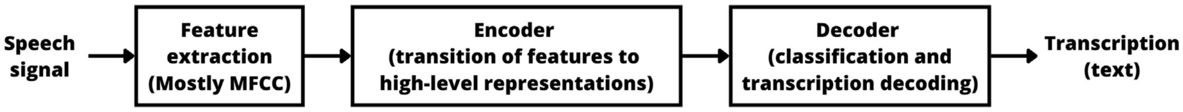
Speech signal
Feature extraction (Mostly MFCC)
Encoder (transition of features to high-level representations)
Decoder (classification and transcription decoding)
Transcription (text)

Figure 2. A block diagram of the workflow of a modern End-To-End automatic speech recognition system in Encoder-Decoder architecture.

2. Speech signal features

An acoustic model of speech was proposed using recognition concepts based on Audio Fingerprint. Research was performed on single vowels, which did not require the use of a language model for classification. The number of classes was assumed to be equal to the number of recognised vowels. Each vowel was assigned a feature vector. This allowed classification based on the acoustic model alone, using classifiers typical of conventional systems. I used speech features such as: temporal (zero crossing density), spectral (power spectral density, roll-off, spread, flatness and special cases of spectral moments: centroid, skewness, kurtosis, flatness), cepstral (linear frequency cepstral coefficients LFCC, power-normalised cepstral coefficients PNCC), perceptual (mel-cepstral coefficients MFCC), formant (first three formant frequencies) and linear predictive coding (linear predictive coefficients LPC, perceptual predictive coefficients PLP).

2.1. Zero-crossing density

The parameter results from the measurement of the points at which a change in the sign of the speech signal. The zero-crossing density is represented by a normalised vector defined for N time intervals [7].

2.2. Power spectral density

In the process of generating and receiving a speech signal, humans operate on the signal in the frequency domain. During speech articulation, is shaped the amplitude-frequency envelope. In receiving, before the speech signal is processed by the neurons, occurs signal frequency components extraction. With the power spectral density function, we can calculate the total power present in each spectral contribution of the signal [6].

2.3. Spectral roll-off

Spectral roll-off characterises the slope of a signal's spectrum. Is defined as the frequency under which a certain percentage of the signal's spectral energy accumulates [8, 9].

2.4. Spectral spread

The spectral spread (spectral width) can be defined in several ways:

- a non-negative real-valued spectral density of a given signal represented as a function of frequency, defined for all frequencies;
- RMS spectral width, which is determined as the standard deviation of the signal density as a function of frequency (Brms);
- equivalent rectangular spectral width, which is determined as the fit of the area under a signal function in the frequency domain to the area of a rectangle of height equal to the maximum value of that function and width 2B (Ber);
- 3 dB slope spectral width, which indicates the "value off" at which the signal spectrum has a value equal to half its peak value (B3dB);
- x Fractional Power Bandwidth (BxF) defined as:

$$\int_{-B_xF}^{B_xF} S(f)df = x \tag{1}$$

x is a particular number in the range $0 < x \leq 1$

Brms, Ber, B3db and BxF are measures of the 'half-width' of a signal as a function of frequency, the nominal width of which is 2B, the signal having most of its area in the region between the positive and negative bandwidths [10].

2.5. Spectral moments

The first three spectral moments, zeroth, first and second, respectively, are measures of energy, frequency, and bandwidth [11]. The research used:

- spectral centroid: the centre mass of the spectrum, a normalised first spectral moment [12];
- spectral skewness: the normalised central spectral moment of the third order;
- spectral flattening (kurtosis): the normalised central spectral moment of the fourth order divided by the square of the variance [13].

2.6. Spectral flatness measure

Spectral flatness determines how similar a sound is to noise, as opposed to tone. A high spectral flatness (closer to 1) indicates that the spectrum is similar to white noise [14].

2.7. Linear frequency cepstral coefficients

LFCC is a linear cepstral representation of sound. The cepstrum is the inverse of the Fourier transform of the signal spectrum, expressed on a logarithmic scale. The linear coefficients are determined using a linear filter bank, which has good resolution in the higher frequency regions [15].

2.8. Power-normalized cepstral coefficients

PNCC uses power law nonlinearities, which replaces the traditional logarithmic nonlinearity used in mel-cepstral coefficients (MFCC). It uses a noise reduction algorithm based on asymmetric filtering, which suppresses background excitations, and a module that implements temporal masking [16].

2.9. Mel-frequency cepstral coefficients

The human ear non-linearly recognises frequencies over the sound spectrum, therefore filter bank analysis is more relevant than linear predictive coding (LPC) analysis. Mel scale is a measure of the perceived frequency (pitch) of a tone. It is based on a frequency division derived from the frequency resolution of the human ear. MFCC coefficients are extracted from the cepstrum of the mel-scale signal by using a transformation of the signal spectrum through mel-scale filters, extraction of the logarithms of the energies of the respective bands, and a discrete cosine transform [17].

2.10. Linear prediction coefficients

LPC imitate the human vocal tract. The technique involves approximating the formants, removing its effects from the speech signal and estimating the concentration and frequency of the remaining bandwidth. Each signal sample is treated as a direct incorporation of previous samples. To separate the remaining bands from the formants and determine their coefficients, a formant measurement is used. The formant positions in the speech signal are predicted by linear prediction coefficients in a sliding window and finding maxima in the spectrum of successive linear prediction filters [18].

2.11. Perceptual linear prediction

PLP combines critical bands, intensity-to-volume compression and equal-volume preemphasis in the extraction of relevant information from speech. It is a combination of spectral analysis and linear prediction. Allows elimination of speaker-dependent features. It gives a representation consistent with a smoothed short-wavelength spectrum, which is equalised and compressed to resemble the response of human hearing. It uses linear predictions to smooth the spectrum. Relevant auditory features are mapped and the speech spectrum is approximated by an autoregressive multipole model. It allows to obtain minimum resolutions at high frequencies (auditory filter bank approach), but its outputs are orthogonal - similar to cepstral analysis [19].

2.12. Formants

The highlighting of certain harmonics (overtones) of the fundamental tone results in acoustically prominent peaks in the spectrum. These maxima are formants, which are characterised by two values: frequency and amplitude. The relative spacing of formants is unique to a particular vowel [20].

3. Experiment

In the study, the developed acoustic model was tested with four classifiers: HMM, SVM, ANN and k-NN. Their performance was evaluated using various quality measures: confusion matrix, accuracy, precision, sensitivity and F1 score. This chapter describes the experimental steps from the creation of the database, signal processing, acoustic model development to the results of the classification.

3.1. Dataset

A dataset consisting of recordings of 6 polish vowels and words containing these vowels was developed for the experiment. The dataset was recorded at a domestic environment using a Zoom H4pro digital recorder and a Shure SM58 cardioid dynamic microphone. The recordings were performed in a general mono PCM 16 bit format with a set sampling rate of 44.1 kHz. The utterances of 20 people (10 females and 10 males) were recorded; 5 repetitions of each isolated vowel and word were taken from each speaker. The dataset contains 100 repetitions of each of the isolated vowels and each of the 6 words. There are 1200 recordings in the dataset. Using Audacity [21], the vowels and words were isolated from the recording and de-noised. The dataset was divided into training and testing in a ratio of 7:3.

3.2. Pre-processing

Two pre-processing methods were used: normalisation and time alignment. I also segmented the vowels from the recordings and performed the necessary linear transformations.

3.2.1. Normalisation

Amplitude normalisation was applied by limiting the dynamic range of the recording to (-1,1). This allows the range of values to be scaled without changing the proportion of data features - the louder parts of the sound recordings do not dominate the quieter ones within a given recording and the entire dataset. For isolated vowels, no segmentation was carried out and normalisation was performed on the raw data. For vowels within words, segmentation into phonemes was done first, followed by normalisation.

3.2.2. Time alignment

Forced time alignment is required by classifiers of conventional ASR systems. In the study a time alignment was used for the parameter vectors for each of the sound samples to have the same length. A linear time alignment based on the median length (number) of samples was used. It calculated the median number of samples for all isolated vowels tested, and then resampled each recording to a number of samples equal to the median using the Fourier method along the time axis.

3.2.3. Segmentation

For the recognition of more complex utterances (e.g. sentences), in conventional ASR systems based on the acoustic and language model, the recordings are divided into segments of fixed length or containing distinct acoustic events (e.g. phonemes). Each segment is then split into frames, where feature extraction takes place. In this study, for samples containing isolated vowels, the segmentation step was omitted and only the division of the signal into frames was performed (the vowel represents a single acoustic event). For words, segmentation was applied to acoustic events using a toolkit [22] based on Bayesian Information Criterion (BIC) [23] and Bi-directional Long Short-Term Memory recurrent neural networks (BiLSTM) [24].

3.2.4. Transformations

Linear transformations are necessary for the extraction of all but temporal features. I used standard time-to-frequency domain transformations: the short-time Fourier transform (STFT).

3.3. Acoustic model

The acoustic model contains all previously described features. It was created by combining the individual feature vectors into a single vector and standardisation. The statistical significance of the characteristics was then assessed.

3.3.1. Combination of feature vectors

The extracted feature vectors have different lengths, so combining them into one feature matrix would result in the shorter features having to be filled with zeros. To avoid this, the vectors of all the features of a given sample were combined into one long vector (18308 elements) - each sample corresponds to a feature vector of the same length. The final feature vector (before the feature selection step based on their importance - see Section 3.3.3) consisted of 146 acoustic parameters. It contained the parameters: zero-crossing density, spectral (power spectral density, roll-off, flatness, centroid, contrast, bandwidth, skewness, kurtosis), first 4 formant frequencies, coefficient (13 PNCC, 40 LFCC, 40 MFCC, 20 LPC and 20 PLP).

3.3.2. Standardisation

Standardization is required to prevent the variability of individual characteristics from affecting the classification results. In the case that any of the characteristics had a wider range, the classifier could find that this characteristic is more relevant than the others, even if this would not be consistent with practice. Applied standardization involves transforming each of the results to obtain a normalized measure, with a mean (expected value) of 0 and a variance of 1 [25].

3.3.3. Statistical features importance

The features importance was tested using a random forest (forest of multiple decision trees) classifier. A class RandomForestClassifier from the ensemble module of the Scikit-learn library was used. The features importance is defined by the random tree attribute feature_importances_, defined as the mean and standard deviation of the accumulation of impurity decrease within each random tree. Three levels of confidence were used: 90%, 95% and 99%. For a confidence level of 99.9%, the 3 statistically most significant features were determined: spectral power density, spectral cut-off and PNCC coefficients.

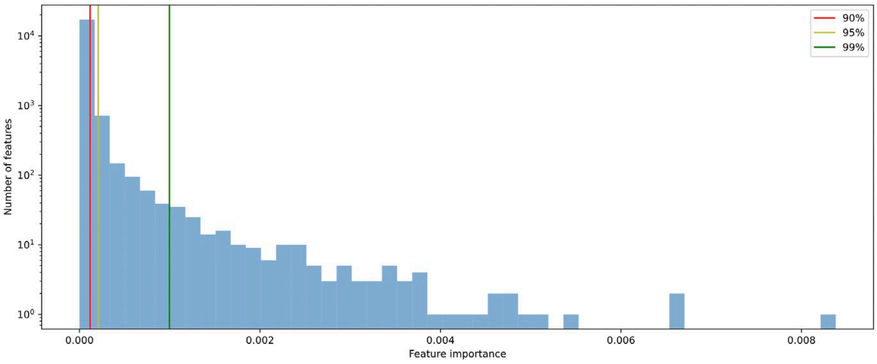

Figure 3. Statistical acoustic features importance depending on the confidence level applied (red 90%, yellow 95%, green 99%).

The vertical axis indicates the number of features and the horizontal shows their importance. The efficiency of the models was tested for the above confidence levels of statistical importance of acoustics features. To train and test models with an index of 90, 1831-element vector, 95 916-element vector and 99 a 184-element vector was used.

3.4. Classification

The acoustic model of vowels was used to classify speech using HMM, SVM, k-NN and ANN. Their effectiveness was evaluated using quality measures: confusion matrix, accuracy, precision, sensitivity, and F-1 score. The results of the classification are shown in Table 1.

3.4.1. Classifiers' characteristics

Classifiers [26] on the basis of parametric methods (based on known or estimated a'priori information - HMM, SVM, ANN) and non-parametric methods (without requiring initial assumptions and estimation of a'priori information, known as minimum distance algorithms - k-NN) were used [27]. The following sections describe the characteristics of the classifiers used in the study.

3.4.1.1. Hidden Markov model

In HMM recognised classes are described through states, transition probabilities between them and probabilities of observations (parameter vectors). In ASR, individual states can be assigned phonemes and word recognition is based on the transition between them [2]. The study used an implementation of the HMM with Gaussian emissions [28]. A separate HMM model was created for each class. The models were learned in a configuration of 1 class against the others. Different numbers of states (2,3,4,6) and two types of covariance [31] were examined - "spherical" (each state uses a single variance value that applies to all functions) and "diag" (each state uses a diagonal covariance matrix). A Viterbi decoder [29] and a forward-backward algorithm with logarithms were used [30]. The best classifiers (results in attached tables) used "diag" covariance, and the number of states differs depending on the number of features considered in the classification: HMM_{90} and HMM_{99} - 4 states, HMM_{95} - 3 states.

3.4.1.2. Support vector machines

SVM selects a small number of critical boundary samples from each class and builds a linear discriminant function that separates them as widely as possible. There is a kernel that maps the data into a high-dimensional feature space, in which a nonlinear task is transformed into a linear disjoint one. The algorithm seeks optimal hyperplanes that maximize the margin between data classes and minimize errors [3]. The study tested an SVM algorithm with linear and radial basis function (RBF) [32] kernel with a gamma coefficient of 1/number of features. Balanced class weighting was used - automatically adjusting the weights inversely proportional to the frequency of classes in the input data. The best classifiers (results in attached tables) depending on the number of features considered in the classification, differed in the selected kernel: SVM_{90} and SVM_{95} - RBF, SVM_{99} - linear.

3.4.1.3. Artificial neural network

The ANN generally consists of three types of layers: input, hidden and output. Each layer can have several nodes (neurons) performing basic operations, and the overall output is a weighted sum of these operations. Each neuron must be trained so that a given set of inputs generates the desired output of the network. Training can be done by providing the network with learning patterns and allowing it to adjust the weighting function according to predefined learning rules. The study used the Multi-layer Perceptron (MLP) architecture [33] with 40 hidden layers. Two activation functions have been tested [34]: ReLu and Tanh. Two types of learning rate schedule were tested for updating the weights: constant and adaptive [35] (keeps the learning rate constant if the learning loss decreases). The best classifiers (results in attached tables) used a constant-type learning rate of 0.001, and the activation function differs, depending on the number of features considered in the classification: ANN_{90} and ANN_{95} - ReLu, ANN_{99} - Tanh.

3.4.1.4. k-nearest neighbours

The k-NN algorithm determines the number of k nearest neighbours of the sample and verifies which class most of them belong to. This is performed by calculating the distance between the unknown sample vector and all learning vectors using a distance or proximity function [5]. The k-NN classifier was tested for various numbers of k nearest neighbours: 3, 4, 5, 6, 7, 8, 9 and 10. Two types of distance were tested: Manhattan and Euclidean [36]. In prediction, were tested two types of weighting functions: uniform, for which all points in each neighbourhood are weighted equally, and distance weighting points by the inverse of their distance. The best classifiers (results in attached tables) used Manhattan distance, distance weighting function and the number of neighbours k varied, depending on the number of features considered in the classification: KNN_{90} and KNN_{95} - 9, KNN_{99} - 6.

3.4.2. Classification results

Tables 1-5 and 11-12 show the results for isolated vowels (best classifier), and Tables 6-10 for vowels within words. Table 1 shows the accuracy, precision, sensitivity and F-1 score for the best of the classifiers tested. Tables 2-5 show the diagonal confusion matrices for these classifiers. Index is the assumed confidence level of feature importance. Tables II-VI show the values from the diagonal confusion matrix for isolated vowel recognition.

Table 1. Classification of vowels.

Quality [%]	SVM_{95}	HMM_{99}	ANN_{90}	$\mathrm{k\text{-}NN}_{95}$
Accuracy	81	58	74	74
Precision	80	63	73	74
Sensitivity	81	58	74	74
F-1 score	80	58	73	74

Table 2. SVM confusion matrix diagonal.

Vowel	SVM_{90} [%]	SVM_{95} [%]	SVM_{99} [%]
A	53	53	57
E	67	73	73
I	93	97	97
O	80	80	67
U	100	100	97
Y	70	80	77

Table 3. HMM confusion matrix diagonal.

Vowel	HMM_{90} [%]	HMM_{95} [%]	HMM_{99} [%]
A	0	30	43
E	90	60	60
I	43	77	93
O	43	43	43
U	67	70	67
Y	27	50	77

Table 4. ANN confusion matrix diagonal.

Vowel	ANN$_{90}$ [%]	ANN$_{95}$ [%]	ANN$_{99}$ [%]
A	63	47	47
E	53	53	70
I	80	87	97
O	77	77	60
U	100	100	90
Y	70	53	57

Table 5. k-NN confusion matrix diagonal.

Vowel	k-NN$_{90}$ [%]	k-NN$_{95}$ [%]	k-NN$_{99}$ [%]
A	57	57	50
E	73	67	63
I	90	93	97
O	70	80	63
U	100	97	93
Y	37	53	60

By using the parts of the dataset with words containing the vowels under study, it was examined how the influence of neighbouring phonemes affects recognition. In this case, the aim was to classify vowels within a word in such a way that the results allow the following questions to be answered:

1. How many vowels were correctly recognised (the correct vowel in the correct segment).
2. For how many words is the probability of occurrence of the searched vowel the highest (even if the wrong segment is selected).
3. For how many words was the desired vowel indicated in the wrong segment?
4. For how many words were the desired vowel not recognised in any of the segments?
5. To what extent was the desired vowel in the correct segment confused with other vowels?

The results are influenced not only by the performance of the classifiers, but also by the extent of correct segmentation, which varied due to the speakers and the phonetic content of the word. In this part of the study, only the k-NN classifier was used.

A classification of isolated vowels was also carried out at for k-NN with a feature importance level of 99.9 (to compare recognition performance of articulated vowels in a word with isolated vowels).

Table 6. Correct vowel in correct segment.

Vowel	Word	Correct recognition [%]
A	BAT	20
E	JEŻ	20
I	NIT	7
O	ROK	7
U	CUD	20
Y	DYM	23

Table 7. The correct vowel most probably (regardless of the segment indicated).

Vowel	Word	Highest probability [%]
A	BAT	20
E	JEŻ	13
I	NIT	23
O	ROK	17
U	CUD	47
Y	DYM	57

Table 8. Indication of the occurrence of the desired vowel in the word (regardless of if the correct segment was indicated and if this vowel obtained the highest probability of occurrence).

Vowel	Word	Occurs in any segment [%]
A	BAT	40
E	JEŻ	67
I	NIT	57
O	ROK	17
U	CUD	47
Y	DYM	77

Table 9. A searched vowel is not found in any of the segments.

Vowel	Word	Doesn't occur in any segment [%]
A	BAT	60
E	JEŻ	33
I	NIT	43
O	ROK	83
U	CUD	53
Y	DYM	23

Table 10. Confusion rate in the correct segment for the searched vowel.

	Vowel					
Word	Correct	Recognized [%]				
BAT	A	E	I	O	U	Y
		20	0	3	37	20
JEŻ	E	A	I	O	U	Y
		7	7	0	3	83
NIT	I	A	E	O	U	Y
		0	10	0	23	67
ROK	O	A	E	I	U	Y
		20	37	0	20	20
CUD	U	A	E	I	O	Y
		3	27	7	0	40
DYM	Y	A	E	I	O	U
		10	40	3	0	23

Table 11. k-NN confusion matrix diagonal with a confidence level of feature importance = 99.9%.

Vowel	k-NN$_{99.9}$ [%]
A	60
E	60
I	87
O	50
U	73
Y	57

Table 12. k-NN quality evaluation (confidence level of feature importance = 99.9 %).

Quality [%]	k-NN$_{99.9}$
Accuracy	64
Precision	66
Sensitivity	65
F-1 score	65

4. Conclusions

The developed acoustic model can be used to recognise isolated vowels, as indicated by the classification results using the SVM$_{95}$ classifier. In the case of in-word vowel detection, there was a significant classification degradation. Based on a comparison of the recognition of isolated vowels and their detection inside the word using the k-NN$_{99.9}$ classifier, it was found that the articulation of vowels in the neighbourhood of other phonemes has a substantial (negative) effect on the recognition performed. This is since the neighbourhood of successive phonemes in an utterance influences the phonetics of a given phoneme (e.g., a voiceless sound becomes more sonorous under the influence of a voiced neighbour). Acoustic models of ASR systems use, e.g., triphones (a set of three consecutive phonemes). Firstly, recording isolated phonemes into the database is difficult (especially consonants), and secondly, this considers the mutual influence of neighbouring phonemes. The quality of detection of the correct vowel in the correct segment also depends on the quality of segmentation of the sample. The quality of segmentation was not the same for every sample. It is affected by the phonetic content of the word in question (the influence of successive phonemes on each other) and the way the speaker pronounces it.

Acknowledgment

This endeavour would not have been possible without Stefan Brachmański, BEng, PhD (Wrocław University of Science and Technology), his inspiration to conduct research in the field of speech acoustics and support at the outset of my scientific journey. I would like to express my deepest gratitude to the head of the Facility of Electroacoustics at the Warsaw University of Technology, Jan Żera BEng, PhD, DSc, ProfTit, for the opportunity to continue my research with his team. I would also like to express my sincere thanks to my supervisor and mentor Piotr Bilski, BEng, PhD, DSc.

Additional information

The author(s) declare: no competing financial interests and that all material taken from other sources (including their own published works) is clearly cited and that appropriate permits are obtained.

References

1. P. Cano, E. Batlle, T. Kalker, J. Haitsma; A review of audio fingerprinting; 2002 IEEE Workshop on Multimedia Signal Processing, 169-173, 2003; DOI: 10.1109/MMSP.2002.1203274
2. M.J.F. Gales, S. Young, The application of hidden Markov models in speech recognition; Foundations and Trends in Signal Processing, 2007, 1(3), 195–304; DOI: 10.1561/2000000004
3. I. Steinwart, A. Christmann; Support vector machines; Wiley Interdisciplinary Reviews: Computational Statistics, 2008
4. J.M. Tebelskis; Speech recognition using neural networks; Carnegie Mellon University, 1995
5. L. Golipour. D. O'Shaughnessy; Context-independent phoneme recognition using a k-nearest neighbour classification approach; In: 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. IEEE, 2009, 1341–1344; DOI: 10.1109/ICASSP.2009.4959840
6. J. Saini, R. Mehra; Power spectral density analysis of speech signal using window techniques; International Journal of Computer Applications, 2015,131(14), 33–36
7. L.R.Rabiner, M.R.Sambur; An algorithm for determining the endpoints of isolated utterances; Bell System Technical Journal, 1975, 54(2), 297–315; DOI: 10.1002/j.1538-7305.1975.tb02840.x
8. M. Kos, Z. Kačič, D. Vlaj; Speech bandwidth classification using general acoustic features, modified spectral roll-off and artificial neural network; In: Mathematical models and methods in modern science Conf., 14th, Mathematical models and methods in modern science, 2012, 212–217
9. M. Kos, Z. Kačič, D. Vlaj; Acoustic classification and segmentation using modified spectral roll-off and variance-based features; Digital Signal Processing, 2013, 23(2), 659–674
10. R.A. Scholtz; How do you define bandwidth?; International Telemetering Conf. Proc, 1972, 8
11. P. Tsiakoulis, A. Potamianos, D. Dimitriadis; Spectral moment features augmented by low order cepstral coefficients for robust asr; IEEE Signal Processing Letters, 2010, 17(6), 551–554; DOI: 10.1109/LSP.2010.2046349
12. B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, O. Nieto; librosa: Audio and music signal analysis in Python; In: Proc. of the 14th Python in science conf., 2015, 8, 18–25
13. P. Virtanen et. al.; SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python; Nature Methods, 2020, 17, 261–272; DOI: 10.1038/s41592-019-0686-2
14. A. Gray, J. Markel; A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis; IEEE Transactions on Acoustics, Speech, and Signal Processing, 1974, 22(3), 207–217; DOI: 10.1109/TASSP.1974.1162572

15. J.J. Noda, C.M. Travieso-González, D. Sánchez-Rodríguez, J.B. Alonso-Hernández; Acoustic classification of singing insects based on mfcc/lfcc fusion; Applied Sciences, 2019, 9(19), 4097; DOI: 10.3390/app9194097
16. C. Kim and R.M. Stern; Power-normalized cepstral coefficients (pncc) for robust speech recognition; IEEE/ACM Transactions on audio, speech, and language processing, 2016, 24(7), 1315–1329; DOI: 10.1109/TASLP.2016.2545928
17. C.K. On, P.M. Pandiyan, S. Yaacob, and A. Saudi; Mel-frequency cepstral coefficient analysis in speech recognition; in 2006 Int. Conf. on Computing & Informatics. IEEE, 2009, 1–5; DOI: 10.1109/ICOCI.2006.5276486
18. U. Shrawankar, V.M. Thakare; Techniques for feature extraction in speech recognition system: A comparative study; arXiv (Cornell University), 2013; DOI: 10.48550/arXiv.1305.1145
19. H. Hermansky; Perceptual linear predictive (plp) analysis of speech; the Journal of the Acoustical Society of America, 1990, 87(4), 1738–1752; DOI: 10.1121/1.399423
20. N. Kraus, T. Nicol; Brainstem origins for cortical 'what'and 'where'pathways in the auditory system; Trends in neurosciences, 2005, 28(4), 176–181; DOI: 10.1016/j.tins.2005.02.003
21. Audacity® software is copyright © 1999-2021 audacity team. the name audacity® is a registered trademark." accessed on: Jun. 2023, Available: https://audacityteam.org/
22. A toolkit to implement segmentation on speech based on bic and nerual network, such as bilstm; https://github.com/wblgers/py_speech_seg (accessed on 2023.06.20)
23. S. Chen, P. Gopalakrishnan et al.; Speaker, environment and channel change detection and clustering via the bayesian information criterion; In: Proc. DARPA broadcast news transcription and understanding workshop, Landsdowne Conference Resort, Landsdowne, 1998, 8, 127–132.
24. R. Yin, H. Bredin, C. Barras; Speaker change detection in broadcast tv using bidirectional long short-term memory networks; In: Interspeech 2017. ISCA, 2017; DOI: 10.21437/Interspeech.2017-65
25. F. Pedregosa et al.; Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011, 12, 2825–2830
26. M.S Sonwane, C.A Dhawale; Evaluation and analysis of few parametric and nonparametric classification methods; In 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT), Ghaziabad, India, 2016, 14–21; DOI: 10.1109/CICT.2016.13
27. C.Zhang, C. Liu, X. Zhang, G. Almpanidis; An up-to-date comparison of state-of-the-art classification algorithms; Expert Systems with Applications, 2017, 128-150; DOI 10.1016/j.eswa.2017.04.003
28. J. Bilmes; Gaussian models in automatic speech recognition; In: D. Havelock, S. Kuwano, M. Vorländer, Eds. Handbook of Signal Processing in Acoustics. Springer, New York, 2008 521-555; DOI :10.1007/978-0-387-30441-0_29
29. Y.R. Kumar, A.V. Babu, K.N. Kumar, J.S.R. Alex; Modified Viterbi decoder for HMM based speech recognition system; In 2014 Int. Conf. on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, 470–474
30. VM. Ilic; Entropy semiring forward-backward algorithm for HMM entropy computation; arXiv (Cornell University), 2021; DOI: 10.48550/arXiv.1108.0347
31. H.Lu, Y.J. Wu, K. Tokuda, L.R. Dai, R.H. Wang; Full covariance state duration modeling for HMM-based speech synthesis; In: 2009 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, IEEE, 2009, 4033-4036; DOI: 10.1109/ICASSP.2009.4960513
32. A. Patle, D.S. Chouhan; SVM kernel functions for classification; In: 2013 Int. Conf. on advances in technology and engineering (ICATE); IEEE, 2013, 1–9
33. A.Ahad, A. Fayyaz, T. Mehmood; Speech recognition using multilayer perceptron; In: IEEE Students Conf., ISCON'02. Proc., IEEE, 2022, 1, 103–109; DOI: 10.1109/ISCON.2002.1215948
34. S. Sharma, S, S. Sharma, A. Athaiya; Activation functions in neural networks; Int. Journal of Engineering Applied Sciences and Technology, 2020, 4(12), 310–316
35. X. Wu, X. R. Ward, L. Bottou; Wngrad: Learn the learning rate in gradient descent; arXiv (Cornell University), 2018; DOI: 10.48550/arXiv.1803.02865
36. M. Mohibullah, M.Z Hossain, M. Hasan; Comparison of euclidean distance function and manhattan distance function using k-mediods; Int. Journal of Computer Science and Information Security (IJCSIS), 2015, 13(10), 61–71