INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
ISSN 1841-9836, 13(5), 853-864, October 2018.

Comparative Analysis of Various Transformation Techniques for Voiceless Consonants Modeling

G. Korvel, B. Kostek, O. Kurasova

Grazina Korvel*, Olga Kurasova
Institute of Data Science and Digital Technologies, Vilnius University
Akademijos str. 4, LT-04812, Vilnius, Lithuania
*Corresponding author: grazina.korvel@mii.vu.lt
olga.kurasova@mii.vu.lt

Bozena Kostek
Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics,
Gdansk University of Technology
G. Narutowicza 11/12, 80-233 Gdansk, Poland
bokostek@audioacoustics.org

Abstract: In this paper, a comparison of various transformation techniques, namely Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT) and Discrete Walsh Hadamard Transform (DWHT) are performed in the context of their application to voiceless consonant modeling. Speech features based on these transformation techniques are extracted. These features are mean and derivative values of cepstrum coefficients, derived from each transformation. Feature extraction is performed on the speech signal divided into short-time segments. The kNN and Naive Bayes methods are used for phoneme classification. We consider both classfication accuracies and computational time. Experiments show that DFT and DCT give better classification accuracy than DWHT. The result of DFT was not significantly different from DCT, but it was for DWHT. The same tendency was revealed for DCT. It was checked with the usage of the ANOVA test that the difference between results obtained by DCT and DWHT is significant.

Keywords: Discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Discrete Walsh Hadamard Transform (DWHT), cepstrum coefficients.

1 Introduction

The state-of-the-art methods applied to speech technology are mostly based on the extraction of features and machine learning. A wide range of speech signal features was conceived and used for classification tasks [19], speech recognition [9], emotional speech recognition [17, 28], phoneme modeling [10], and speech analytics tasks [2]. There are also other approaches employed for processing speech signals, where feature extraction process is discarded. For example, the use of fuzzy logic [29, 30] applied to speech technology, specifically to voice activity detection (VAD), speech segmentation, and coding, cannot be disregarded in this aspect. Moreover, very intense activities connected to the usage of resources for large-scale deep learning analysis applied to speech recognition or emotional speech recognition may be observed in the last few years [15,24,32]. However, when we are talking about speech analytics and modeling, speech synthesis or audio-visual speech recognition, the progress in these fields is below expectations. Secondly, this research area requires a different approach, a thorough analysis of individual spoken elements needs to be performed as there is basic knowledge still missing in this context.

The phoneme mathematical models, utilized as tools for describing speech, are of great importance not only for speech synthesis. The need for research on phoneme models of speech is justified by its numerous possible uses. The following can be named: speech recognition, helping with pronunciation and learning foreign languages (a comparison of phoneme utterances and its model enabling to demonstrate differences of pronunciation), studies in linguistics, medical field (e.g. disturbances in speech present in stroke and neurodegenerative diseases, disorder in one or more prosodic functions, deficits in speech production, etc.). In some of the envisioned applications the obtained results can be a part of a larger multimodal Human-Computer Interaction system consisting of three modalities: vocal, facial and gesture based recognition.

The object of this research is the consonant phoneme signals which are more difficult for analysis, modeling and classification tasks as those of vowels and semivowels. The character of the consonant signals is consonant-dependent and varying. Stop consonants can be considered as quasiperiodic signals in noise, while fricative consonants as aperiodic signals. We can also divide those phonemes into two sets: voiced and voiceless sounds [6]. This means that the vocal folds are apart while saying these sounds. In speech processing, sounds can be represented as a source-filter model [22]. The filter represents the vocal tract, which is excited with a source. A source is a pulse sequence for the voiced sounds and noise for the unvoiced sounds. A commonly used technique for separating source and filter in a speech signal is cepstral analysis [14]. The cepstrum is widely used in speech processing [8, 16].

In all the areas mentioned above automatic feedback for systems and applications is also of importance, thus in a given methodology both feature extraction and machine learning should be applied. To create a mathematical model of a phoneme, it is important to find a suitable parametric description of speech. The speech signal is converted to the appropriate space domain and preprocessing is carried out. The two main domains of analysis are time and frequency. The first of them shows the time varying character of the signal, the second mirrors how the energy of the signal is contained within the frequency range. In the frequency domain, parameters are often based on the Fourier spectrum. In this paper, we perform a comparative experiment based on DFT (Discrete Fourier Transform), DCT (Discrete Cosine Transform) and DWHT (Discrete Walsh Hadamard Transform). The results returned by this study should enable us to verify which transformation method along with feature extraction work better when such a methodology is applied to check phoneme modeling precision.

The relationship between the performance of transformation techniques in the context of feature vectors derivation has been investigated by many researchers, in various speech classification tasks. In the paper of Velican et al. [33], a comparison of DWHT and DCT as feature selection tools in the case of identifying rhotacism is performed. The experiment result shows that classification rate in the case of DWHT is better than the rate obtained with DCT. In the paper of Kekre and Kulkarni [7] a comparison of the performance DCT and DWHT for various feature vector sizes with and without overlap based on speech utterances is given for speaker identification. The results show that DCT performs better than DWHT. The comparison of two fundamentally different approaches the Fast Fourier Transform (FFT) and Hilbert-Huang Transform (HHT) is given in paper of Donnelly [4]. The behavior and flexibility of these two transforms are examined for a number of different time domain signal types.

The targeted consonant phonemes are also more susceptible to noise than vowels mainly due to their lower intensity. This means that in many conditions they may easily be masked by signals interfering with speech. That is why it is important to find optimized feature vectors that will perform in both quiet and noise conditions. This paper deals with a domain-dependent analysis and classification of consonant phonemes utilizing the cepstrum analysis. The feature vectors consist of cepstrum coefficients derived from the Fourier, Cosine and Walsh Hadamard transforms.

2 Transformation techniques

Let $x(k)$ is a signal with length K, where K is an integer power of 2 ($k = 1, \ldots, K$). The frequency domain representation shows how the energy of this signal is contained within the frequency range. The techniques of signal transformation from time to the frequency domain are given in this section.

2.1 Discrete Fourier transform

Fourier analysis is based on decomposing signals into sinusoids [26]. DFT is a family member of this analysis used with digital signals. The transform decomposes the signal $x(k)$ into the sequence of complex numbers $y(1), ..., y(K)$ according to the formula:

$$y(k) = \sum_{n=1}^{K} x(n) e^{\frac{(-2\pi j)}{K}(n-1)(k-1)} \tag{1}$$

where the symbol j denotes the imaginary unit.

To convert signal data from the frequency to the time domain the Inverse Discrete Fourier Transform (IDFT) is applied. The IDFT is defined as follows:

$$x(k) = \frac{1}{K} \sum_{n=1}^{K} y(n) e^{\frac{2\pi j}{K}(n-1)(k-1)} \tag{2}$$

The result of IDFT will be used in the construction of the signal cepstrum.

2.2 Discrete Walsh-Hadamard transform

DWHT is a non-sinusoidal technique that represents a signal as a set of orthogonal rectangular waveforms. The transform is given by the formula:

$$y(k) = \sum_{n=1}^{K} x(n) W_K(k, n) \tag{3}$$

The basis function is described as follows:

$$W_K(k, n) = \prod_{l=1}^{L-1} (-1)^{n_l k_{M-1-l}} \tag{4}$$

where $L = \log_2 K$, n_l is the l^{th} bit in the binary representation of n [27].

As we see from Eq. (4), DWHT takes the binary value 1 or -1. The Inverse Discrete Walsh Hadamard Transform (IDWHT) is defined as follows:

$$y(k) = \frac{1}{K} \sum_{n=1}^{K} x(n) W_K(k, n) \tag{5}$$

The only difference between DWHT and IDWHT (see Eq. (3) and Eq. (5)) is a constant divisor.

2.3 Discrete Cosine transform

DCT decomposes a signal into cosine functions. The transformation has several standard variants. These variants and the mathematical properties of DCT are presented in works of Rao and Yip, Oppenheim *et al.* [18, 23]. In this paper, the Discrete Cosine transform of the signal $x(k)$ is computed according to the formula:

$$y(k) = \sqrt{\frac{2}{K}}\beta(k)\sum_{n=1}^{K} x(n)cos(\frac{\pi(2n-1)(k-1)}{2K}) \tag{6}$$

where coefficient $\beta(k)$ is defined as follows:

$$\beta(k) = \begin{cases} \frac{1}{\sqrt{2}}, & if \quad k = 1 \\ 1, & if \quad k \neq 1 \end{cases} \tag{7}$$

The formula of the Inverse Discrete Cosine Transform (IDCT) is given below:

$$x(k) = \sqrt{\frac{2}{K}}\sum_{n=1}^{K} \beta(k)y(n)cos(\frac{\pi(n-1)(2k-1)}{2K}) \tag{8}$$

DCT and other transformation techniques analyzed in this Section are orthogonal transforms. Therefore, they can be computed using the fast algorithms.

3 Feature extraction

All the N samples of the analyzed phoneme are collected into a vector:

$$x = \left[\begin{array}{ccc} \mathrm{x(1)}, & \mathrm{x(2)}, \ldots, & \mathrm{x(N)} \end{array}\right] \tag{9}$$

The phoneme signal is divided into short-time frames, the length of which is M samples. A process of dividing a signal into frames is typical for the speech signal analysis. To each of these frames, a window function $w(n)$ is used. Due to the window procedure, a part of the signal data is lost. Therefore, an overlap of segments is utilized. How much should the segments overlap can be seen in [5].

Denote by L the number of the overlapped samples. Then the number of intervals can be obtained by the following formula:

$$P = \left[\frac{N-M}{M-L}\right] + 1 \tag{10}$$

where $[\alpha]$ stands for an integer part of the real number.

Then the phoneme signal can be written as the following matrix:

$$X = \begin{bmatrix} w(1) \times x(1) & \ldots & w(M) \times x(M) \\ w(1) \times x(M-L+1) & \ldots & w(M) \times x(2M-L) \\ \ldots & \ldots & \ldots \\ w(1) \times x\Big((P-1) \times (M-L+1)\Big) & \ldots & w(M) \times x\Big((P-1) \times (2M-L)\Big) \end{bmatrix} \tag{11}$$

The calculation procedure of the cepstrum coefficients constitutes a part of the algorithm prepared. The consecutive steps of the algorithm are listed below:

Step 1. The selected transform is applied to each row of the matrix X.

Step 2. The absolute values are taken out.

Step 3. The logarithm is calculated.

Step 4. The inverse transform is applied.

Consequently, we obtain the matrix of the cepstrum coefficients:

$$C = \begin{bmatrix} c_{11} & \dots & c_{1M} \\ c_{21} & \dots & c_{2M} \\ \dots & \dots & \dots \\ c_{k1} & \dots & c_{kM} \end{bmatrix} \tag{12}$$

The mean values of the columns of the matrix C are calculated. All the obtained values are collected into a vector c:

$$c = \begin{bmatrix} \mathrm{c(1)}, & \mathrm{c(2)}, \dots, & \mathrm{c(M)} \end{bmatrix} \tag{13}$$

The mean cepstrum values given in Eq. (13) are used as representative features.

In order to determine whether a function is increasing or decreasing, additionally the first-order delta derivatives are calculated. The first-order dynamic coefficients are calculated from the static cepstrum coefficients using the following regression formula:

$$d_m = \frac{\sum_{n=1}^{N} n(c_{ex}(m+n) - c_{ex}(m-n))}{2\sum_{n=1}^{N} n^2} \tag{14}$$

where

$$c_{ex} = \begin{bmatrix} \underbrace{0, \dots, 0,}_{M}, & c(1), & \dots, & c(M), & \underbrace{0, \dots, 0}_{M} \end{bmatrix} \tag{15}$$

$m = 1...M$, N is the regression window size.

4 Classification methods

A vast literature on the application of machine learning to the classification task exist. Researchers developed many approaches to the problem of classification, including methods for inducing rule sets, models in the form of a tree structure, linear discriminants, statistical learning algorithms, and artificial neural networks [24]. In an experiment, we use two classical machine learning algorithms to compare classification rates. First of them is the Naive Bayes classification method, based on Bayes theory [11]. This algorithm is widely used because it often outperforms more sophisticated classification methods. It falls into the statistical learning algorithms and provides the probability of each attribute set belonging to a given class.

The second classification algorithm used in this experiment is k-Nearest Neighbors (kNN), based on Euclidean distances between the elements of the test dataset and elements of the training dataset [13]. The number of nearest neighbors is set by performing preliminary tests.

5 Experimental results

The experiment was performed on Lithuanian speech recordings, created during the project LIEPA (Services controlled by the Lithuanian Speech) [20]. The database consists of 100 hours of words, phrases and sentences recordings, different speakers, both male and female voices and is adapted for scientific research. In the present experiment, we consider the extracted consonant phonemes from this database for our analysis. The phonemes are the following: /t/, /k/, /s/, /ʃ/. The first two (/t/ and /k/) are called stop consonants because the air in the vocal tract is stopped at some period. An example of the phoneme /t/ signal is given in Figure 1.

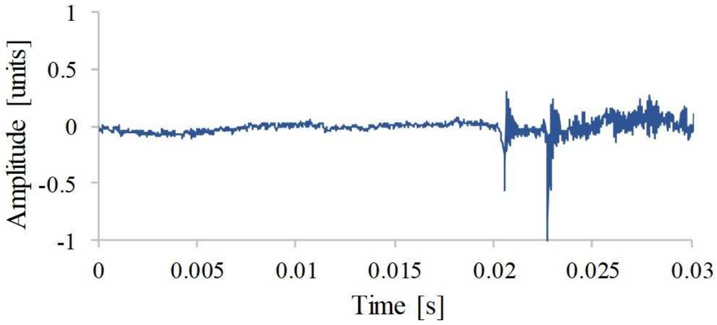

Figure 1: The plot of the consonant phoneme /t/

The next two phonemes (/s/ and /ʃ/) are called fricative consonants, which are produced when the air is squeezed out through a small hole in the mouth. An example of a phoneme /ʃ/ signal is given in Figure 2. The audio data used in the analysis are wav files with the following parameters: sampling frequency: 22 kHz, quantification: 16 bits, the number of channels is 1.

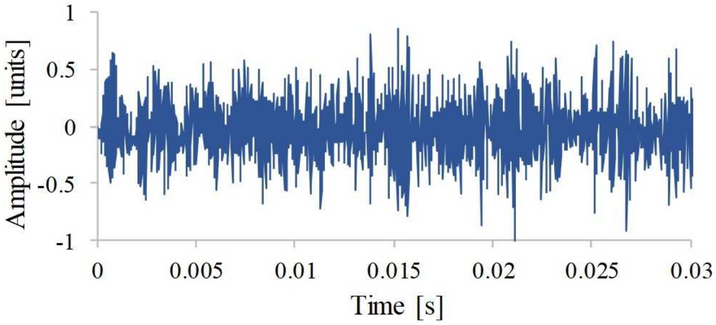

Figure 2: The plot of the consonant phoneme /ʃ/

The feature extraction procedure proposed in Section 3 involves several steps. First of all, the signal pre-processing is carrying out. Then the signal is converted to the appropriate space domain and the extraction of features is performed. In this experiment, signal pre-processing is performed using the following parameters: the input signal is divided into frames of 512 samples, and then for each frame, the Hamming window is chosen. The overlap of 50% is used. Therefore, the number of cepstrum coefficients is equal to the number of coefficients of transformation (i.e. 512). An observation reveals that only part of them is useful for separation of the consonant classes. That is why only the first 12 coefficients were selected as representative features. In Figure 3, DFT cepstrum is shown. It can be seen from Figure 3 that the cepstrum coefficients present differences between consonant classes, this is especially visible in the case of the first four coefficients.

It was checked in further analyses that cepstra of both DWHT and DCT followed the same trend. The plots of these cepstra are shown in Figures 4 (DWHT-based) and 5 (DCT-based).

Additional 12 features are derived from computing the first order derivatives.

In the experiment, 480 utterances (120 for each phoneme) were considered. These phonemes were cut out of the recordings of 15 speakers (9 females and 7 males). We extracted parameters for all these phonemes. The data are divided into two segments: one employed to teach a model and another one utilized to test this model. The test set for models is constructed of 10% randomly selected phonemes.

Due to the fact that the set of samples is not very big, and it is important to estimate the true error rate of a given classifier, an experiment was repeated 50 times for each case and the

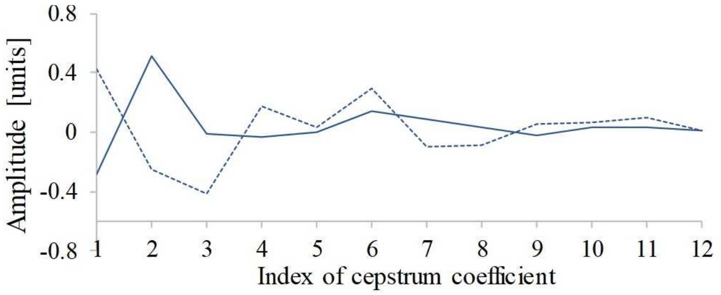

Figure 3: The cepstrum coefficients obtained by DFT (the solid line–consonant phoneme /t/, the dotted line–consonant phoneme /ʃ/)

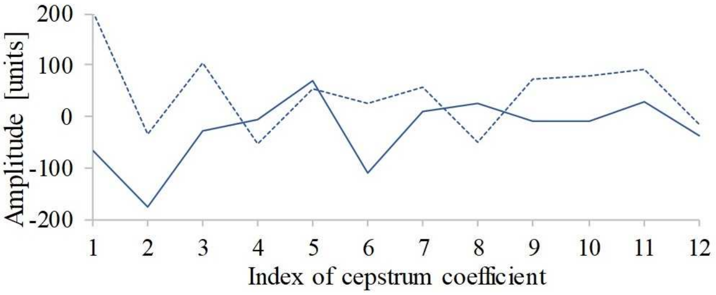

Figure 4: The cepstrum coefficients obtained by DWHT (the solid line–consonant phoneme /t/, the dotted line–consonant phoneme /ʃ/)

arithmetic mean was calculated. A comparison of the performance of two selected classification methods averaged for all speakers, males and females separately is given in Table 1.

In order to determine whether the differences between the means of the three parametrization techniques are statistically significant, the one-way analysis of variance (ANOVA) test is used. The test significance level α equals to 0.05. We state a null hypothesis (H_0) that in each case both samples are from populations with the same means. The decision rule to reject this hypothesis is as follows:

$$reject\ H_0\ if\ F > F_{critical}(1-\alpha) \tag{16}$$

where F is the test statistic calculated as the ratio of the difference between the means over a distribution of their data points, and $F_{critical}$ is the critical value taken from the F distribution table [3]. The results of ANOVA test are given in Table 2.

The obtained F values (see Table 2) are compared with the critical value for F distribution.

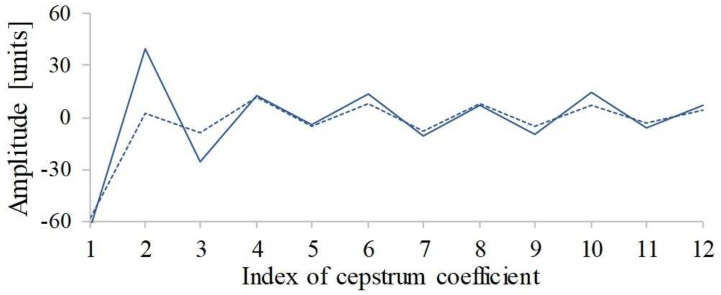

Figure 5: The cepstrum coefficients obtained by DCT (the solid line–consonant phoneme /t/, the dotted line–consonant phoneme /ʃ/)

Table 1: The classification accuracy [%] for 4 consonant classes

		k-Nearest Neighbors			Naive Bayes		
		DWHT	DFT	DCT	DWHT	DFT	DCT
All	Mean	82.67	85.13	**86.04**	87.00	**90.50**	89.42
	Std.Dev.	5.97	4.89	4.44	4.16	4.00	4.41
Male	Mean	86.71	90.71	**90.86**	96.29	98.14	**98.29**
	Std.Dev.	11.55	7.53	8.29	5.25	3.77	3.40
Female	Mean	76.94	79.56	**81.88**	84.38	87.06	**87.81**
	Std.Dev.	6.05	6.35	6.19	4.77	4.93	4.86

Table 2: The result of ANOVA test for kNN and Naive Bayes

		k-Nearest Neighbors			Naive Bayes		
		DWHT/ DFT	DWHT/ DCT	DFT/ DCT	DWHT/ DFT	DWHT/ DCT	DFT/ DCT
All	F-value	**5.07**	**10.28**	0.963	**18.42**	**7.95**	1.66
	p-value	0.026581	0.001815	0.328837	0.000042	0.005826	0.201148
Male	F-value	**4.21**	**4.25**	0.008	**4.13**	**5.11**	0.039
	p-value	0.042865	0.041988	0.928349	0.044831	0.026022	0.842655
Female	F-value	**4.48**	**16.28**	3.40	**7.68**	**12.75**	0.59
	p-value	0.036904	0.000109	0.068158	0.006691	0.000553	0.445528

In the experiments performed, the obtained F is significant at a given level if it is equal to or greater than 4.03 ($F_{critical} = 4.03$). According to these assumptions, the differences between DWHT and DFT as well as differences between DWHT and DCT are statistically significant. Meanwhile, the differences between DFT and DCT are not statistically significant.

The experiments were performed using MATLAB on a Laptop with IntelR CoreTM i5-6200U 2.20 GHz CPU and 8 GB of RAM. The computation time is given in Table 3. From this table we see, that the computational time of kNN is much smaller than Naive Bayes.

We also compare the results obtained on Lithuanian consonants with the classification effectiveness collected from the literature for other languages (see Table 4). Obviously, such a comparison cannot be performed straightforward as the studies recalled here concern different languages and also a variety of features and classification methods as well as they are researched for different purposes (e.g. speech recognition, clean and telephone speech differentiation, speech productionÂ models and mechanisms, pathology disorder, etc.). Thus data contained in Table 4 may serve only to a limited extent when comparing algorithmic performances.

Though different classification methods are employed in the studies recalled, we see that our results are consistent with the results of other researchers, however they are dependent more on the vector feature content than on the type of a classifier.

Table 3: Computational time [s] for the classifiers

	k-Nearest Neighbors			Naive Bayes		
	DWHT	DFT	DCT	DWHT	DFT	DCT
All	0.1464	0.1456	0.1457	2.5672	2.5742	2.5599
Male	0.0270	0.0270	0.0270	1.8016	1.7982	1.8021
Female	0.0908	0.0902	0.0899	2.2227	2.1915	2.1928

Table 4: Consonants classification performance in literature

Reference	**Dataset**	**Parameters**	**Classification technique**	**Overall classification accuracy**
Thasleema and Narayanan, 2018 [31]	Malayalam (India) (unaspirated, aspirated, nasal, approximants, fricatives)	Normalized Wavelet Hybrid Feature (NWHF) vector based on Wavelet Transform	*k*-Nearest Neighbors (kNN), Artificial Neural Network (ANN), Support Vector Machine (SVM)	From 34.2% to 60.2% for kNN, from 45.9% to 63.7% for ANN and 55.4% to 79.9% for SVM (depending on the mother wavelet)
Korvel and Kostek, 2017 [9]	MODALITY database (English stop consonants)	Descriptors coming from music information retrieval	*k*-Nearest Neighbors (kNN)	73%
Lee and Choi, 2012 [12]	TIMIT database (American English)	Mel-frequency cepstral coefficients (MFCCs), first and second derivatives plus acoustic parameters such as band-limited RMS energy, amplitude of the first harmonic and peak normalized cross correlation values (PNCC)	Gaussian mixture models (GMMs)	Depending on the type of consonants, i.e.: stops, fricatives, and, affricates classification accuracies are as follows: 82.2%, 80.6%, and 78.4% respectively
Ali et al., 2001 [1]	American English stop consonants	The acoustic-phonetic characteristics	The authors proposed classification system combining the voicing detection and the place of articulation detection	86%
Pruthi and Espy-Wilson, 2003 [21]	TIMIT database (Nasals and semivowels)	The acoustic parameters which include F1 measure, a pick onset/offset measure, an energy ratio, and a formant density measure	Support Vector Machine (SVM) based classifiers	Accuracies of 88.6%, 94.9% and 85.0% were obtained for prevocalic, postvocalic and intervocalic sonorant consonants, respectively

6 Conclusions

In the paper, we have compared the performance of DFT, DWHT and DCT for voiceless consonant (/t/, /k/, /s/, /ʃ/) classification. In order to evaluate the classification accuracy, two methods, namely kNN and Naive Bayes were used. The analyses were performed for the whole group of speakers, and for male and female speakers separately. The highest classification accuracy for all speakers (86.04%) has been achieved for features based on DCT technique, in the case of kNN method. While for Naive Bayes classifier, the highest accuracy for all speakers was equal to 90.50% for DFT. In the cases of the analysis of male and female recordings separately, the highest accuracies have been achieved for features based on DCT technique for both classifiers. These accuracies are as follows: for kNN classifier the highest accuracy for male group was equal to 90.86%, for female group – 98.29%, while for Naive Bayes classifier the highest accuracy for male group was equal to 81.88%, for female group – 87.81%.

The employment of one-way analysis of variance (ANOVA) test for results of both selected classification methods revealed the same tendency for different groups of speakers and different classifiers: the difference between results, obtained by DFT and DCT is not significant, meanwhile difference between results, obtained by DWHT and the other two transformations (DFT and DCT) is significant.

A comparison of the results obtained on Lithuanian consonants with other results in the literature was also performed. A literature review shows, that our results are consistent with those of other researchers.

It is important to mention that our primary intention was not to obtain high classification accuracy, but the goal was to determine which transformation method returns better results when applying a given feature vector and a regular machine learning algorithm. This is important in the context of the feedback needed on phoneme modeling precision to verify the model consistency with the initial phoneme target. As seen from observations the created feature vector is not complete as the accuracy obtained is not fully satisfying. Therefore, in the future research, we will investigate the possibility of extending the created feature vector with additional signal descriptors applicable to short-segmented speech units. In addition, more effective classification algorithms based on the weighted features are to be considered. Finally, some additional tests should be executed on the same feature vectors but taking into account the phoneme neighbors and also presence of noise.

Funding

This research is funded by the European Social Fund under the No 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities" measure.

Bibliography

[1] Ali, A. M. A.; Van der Spiegel, J.; Mueller, P. (2001); Acoustic-Phonetic Features for the Automatic Classification of Stop Consonants, *IEEE Transactions on Speech and Audio Processing*, 9(8), 833–841, 2001.

[2] Czyzewski, A.; Piotrowska, M.; Kostek B. (2017); Analysis of Allophones Based on Audio Signal Recordings and Parameterization, *The Journal of the Acoustical Society of America*, 141 (5), 3521–3521, 2017.

[3] De Muth, J. E. (2014); *Basic Statistics and Pharmaceutical Statistical Applications*, 3rd edn, CRC Press, 2014.

[4] Donnelly, D. (2006); The Fast Fourier and Hilbert-Huang Transforms: A Comparison, *International Journal of Computers Communications & Control*, 1 (4), 45–52, 2006.

[5] Heinzel, G.; Rudiger; A., Schilling, R, (2002); Spectrum and Spectral Density Estimation by the Discrete Fourier Transform (DFT), Including a Comprehensive List of Window Functions and Some New Flat-top Windows, *Internal Report, Max-Planck-Institut fur Gravitationsphysik, Hannover*, 2002.

[6] Kasparaitis, P. (2005); Diphone Databases for Lithuanian Text-to-speech Synthesis. *Informatica*, 193–202, 2005.

[7] Kekre, H. B., Kulkarni, V. (2011); Speaker Identification using Row Mean of DCT and Walsh Hadamard Transform, *International Journal on Computer Science and Engineering*, 3(3), 1295–1301, 2011

[8] Kim C.; Stern R. M. (2016); Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 24(7), 1315–1329, 2016.

[9] Korvel, G.; Kostek, B. (2017); Examining Feature Vector for Phoneme Recognition, *Proceeding of IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2017*, Bilbao, Spain, 394–398, 2017.

[10] Korvel, G.; Kostek, B. (2017); Voiceless Stop Consonant Modelling and Synthesis Framework Based on MISO Dynamic System, *Archives of Acoustics*, 3, 42, 375–383, 2017.

[11] Kotsiantis, S. B. (2007); Supervised Machine Learning: A Review of Classification Techniques, *Informatica*, 31(3), 249–268, 2007.

[12] Lee, S. M.; Choi J. Y.(2012); Analysis of Acoustic Parameters for Consonant Voicing Classification in Vlean and Telephone Speech, *The Journal of the Acoustical Society of America*, 131, EL197 (2012); doi: 10.1121/1.3678667

[13] Manocha S.; Girolami M. A. (2007); An Empirical Analysis of the Probabilistic K-nearest Neighbour Classifier, *Pattern Recognition Letters*, 28, 1818–1824, 2007.

[14] Milner, B.; Shao X. (2002); Speech Reconstruction from Mel-Frequency Cepstral Coefficients using a Source-Filter Model, *7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2421–2424, 2002.

[15] Mitra V.; Sivaraman G.; Nam H.; Espy-Wilson C.; Saltzman E.; Tiede M. (2017); Hybrid Convolutional Neural Networks for Articulatory and Acoustic Information Based Speech Recognition, *Speech Communication*, 89, 103-112, 2017.

[16] Mitra, V.; Franco, H.; Graciarena, M.; Vergyri D. (2014); Medium-Duration Modulation Cepstral Feature for Robust Speech Recognition., *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1749–1753, 2014.

[17] Noroozi, F.; Kaminska, D.; Sapinski, T.; Anbarjafari, G. (2017); Supervised Vocal-Based Emotion Recognition Using Multiclass Support Vector Machine, Random Forests, and Adaboost, *Journal of the Audio Engineering Society*, 65(7/8), 562–572, 2017.

[18] Oppenheim, A. V.; Schafer, R. W.; Buck, J. R. (1999); *Prentice-Hall Signal Processing Series Discrete-Time Signal Processing*, 2nd edn. Prentice Hall, Inc., New Jersey, 1999.

[19] Pravin, S. C.; Anjana, R.; Pandiyan, T. P.; Ranganath, S. K.; Rangarajan P. (2017); ANN Based Disfluent Speech Classification, *Artificial Intelligent Systems and Machine Learning*, 9(4), 77-80, 2017.

[20] Project LIEPA Homepage, https://www.rastija.lt/liepa/about-project-liepa/7596, accessed on 2018/03/02.

[21] Pruthi T.; Espy-Wilson C. (2003); Automatic Classification of Nasals and Semivowels, *ICPhS 2003-15th International Congress of Phonetic Sciences*, 3061–3064, 2003

[22] Pyz, G.; Simonyte, V.; Slivinskas, V. (2014); Developing Models of Lithuanian Speech Vowels and Semivowels, *Informatica*, 25 (1), 55–72, 2014.

[23] Rao, K. R.; Yip, P. (1990); *Discrete Cosine Transform: Algorithms, Advantages, Applications*, 1st edn, Academic Press, 1990.

[24] Ravanelli M.; Brakel P.; Omologo M.; Bengio Y. (2017); A Network of Deep Neural Networks for Distant Speech Recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4880–4884, 2017.

[25] Sammut C.; Webb G. I. (2011); *Encyclopedia of Machine Learning. Springer Science & Business Media*, Springer New York, 2011.

[26] Smith, S. W. (1999); *The Scientist and Engineer's Guide to Digital Signal Processing*, 2nd edn. California Technical Publishing, San Diego, California, 1999.

[27] Sundararajan, D. (2001); *The Discrete Fourier Transform - Theory, Algorithms and Applications*, World Scientific, 2001.

[28] Tamulevicius, G.; Liogiene, T. (2015); Low-Order Multi-Level Features for Speech Emotion Recognition, *Baltic Journal of Modern Computing*, 4(3), 234–247, 2015.

[29] Teodorescu H.N.L. (2015), A Retrospective Assessment of Fuzzy Logic Applications in Voice Communications and Speech Analytics, *International Journal of Computers Communications & Control*, 10 (6), 105–112, 2015.

[30] Teodorescu H.N.L. (2015); Fuzzy Logic in Speech Technology-Introductory and Overviewing Glimpses, *Fifty Years of Fuzzy Logic and its Applications*, 581–608, 2015.

[31] Thasleema T. M.; Narayanan N. K.: Consonant Classification using Decision Directed Acyclic Graph Support Vector Machine Algorithm, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 6(1), 59–74, 2013.

[32] Tzinis E.; Potamianos A. (2017); Segment-Based Speech Emotion Recognition using Recurrent Neural Networks, *Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 190–195, 2017.

[33] Velican, V.; Strungaru, R.; Grigore, O. (2012); Automatic Recognition of Improperly Pronounced Initial 'r' Consonant in Romanian, *Advances in Electrical and Computer Engineering*, 12 (3), 79-84, 2012.