Applied Acoustics 136 (2018) 1–8

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification

Turgut Özseven[a,*], Muharrem Düğenci[b]

[a] Department of Computer Engineering, University of Gaziosmanpaşa, 60250 Tokat, Turkey
[b] Departmant of Industrial Engineering, University of Karabük, 78050 Karabük, Turkey

ARTICLE INFO

Keywords:
Speech feature extraction
Speech classification
Speech toolbox
Speech processing toolbox

ABSTRACT

Background and objective: The acoustic analysis, an objective evaluation method, is used to determine the descriptive attributes of the voices. Although there are many tools available in the literature for acoustic analysis, these tools are separated by features such as ease of use, visual interface, and acoustic parameter library. In this work, we have developed a new toolbox named SPAC for extracting and simulating attributes from speech files.
Methods: SPAC has a modular structure and user-friendly interface, which will make up for the shortcomings of existing vehicles. In addition, modules can be used independently of each other. With SPAC, about 723 attributes can be extracted from each voice file in 9 categories. A validation test was applied to verify the validity of the toolbox-derived attributes. When the validation test was performed, the attributes obtained with Praat and OpenSMILE were grouped as standard, the attributes obtained with SPAC as test data, and the general differences between the attributes were evaluated with mean square error and mean percentage error. In another method used for verification, the classification performance is tested using the SPAC-derived attributes for classification.
Results: According to the validation test results, SPAC attributes differ between 0.2% and 9.7% compared to other toolboxes. According to the results of the classification test, the SPAC attribute clusters can identify each class and the classification success varies between 1% and 3% according to the attributes obtained from other toolboxes. As a result, the attributes obtained with SPAC accurately describe the voice data.
Conclusions: SPAC's superiority over existing toolboxes is that it has an easy-to-use user-friendly interface, it is modular, allows graphical representation of results, includes classification module and allows to work with SPAC data or data obtained from different toolboxes. In addition, operations performed with other tools can be performed more easily with SPAC.

1. Introduction

Speech processing is the processing of sounds and speech with Digital Signal Processing (DSP) methods. For many years, increasing the quality of voice has been used in pattern recognition studies such as person recognition, speech recognition, emotion recognition and communication. Pattern recognition consists of two major areas: feature extraction and classification [1]. Feature extraction is used for obtaining the data which will be used in classification. Feature extraction is usually divided into two groups, namely temporal and spectral, and can be grouped in four sub-categories [2]: (1) acoustic features, (2) linguistic features, (3) context information, and (4) hybrid features which combine acoustic features with other information sources. Speech feature extraction process is conducted with the help of ready toolboxes or codes developed by researchers. Toolboxes help reduce workload and increase the reliability of extracted attributes. The existing toolboxes have been developed for the specific workspace and many do not have a user-friendly interface. In addition, the extracted attributes differ according to the toolboxes. For example, the Hidden Markov Toolkit (HTK) [3] has the ability to extract a large number of attributes over multiple conversations, but it has the difficulty of using it because it does not have a Graphical User Interface (GUI). In addition, working on more than one files can be necessary in speech processing works. While many toolboxes support this, COLEA [4] does not have the ability to process batch files, even though it has a GUI that provides ease of use. PRAAT [5], which has a GUI and is capable of batch file processing, is used in many studies. Although PRAAT has the ability to extract a large number of attributes and has a graphical interface, it is necessary to prepare scripts for batch file processing. Although OpenSMILE [6] allows for the use of a large number of attribute inferences

* Corresponding author.
E-mail address: turgut.ozseven@gop.edu.tr (T. Özseven).

https://doi.org/10.1016/j.apacoust.2018.02.009
Received 2 January 2018; Accepted 8 February 2018

and batch files, it the use of the GUI with additional software. In addition, the settings for extracting the attributes and the attributes to be extracted are performed via the configuration file.

The purpose of this work is to develop a new toolbox to strengthen existing toolboxes for acoustic analysis and to overcome them. One of the biggest shortcomings seen in existing toolboxes is that they do not include attribute extraction and classification. Another shortcoming is that a significant number of toolboxes are available for professional users, so they do not have a user-friendly interface and do not display the results graphically. To strengthen these weaknesses in the existing toolboxes and eliminate the shortcomings, we developed a toolbox in the MATLAB environment called SPeech ACoustic (SPAC). In addition, the feature extraction with SPAC can be realized more easily than with other tools. SPAC consists of 6 Modules (Configuration, Batch Analysis, Pre-Processing, Feature Extraction, Post-Processing and Classification). Because SPAC has a modular structure, modules can be used independently depending on the need. With SPAC, feature can be extracted in 9 different main categories including F0, formant, voice quality, LPCC, MFCC, Zero-Crossing Rate, Signal Energy, Speech and Wavelet. The Pro-processing module includes down sampling, pre-emphasis, noise reduction and DC-offset removal methods. The post-processing module includes z-score for attribute normalization, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Fisher selection methods for attribute selection. In the classification module, Support Vector Machine (SVM), Multilayer Perceptron (MLP) and k-Nearest Neighbors (k-NN) classifiers can be used. SPAC has superior modular structure, user-friendly interface, graphical representation and classification, which are superior to other toolboxes. A validation test and classification test were conducted to test the validity of the SPAC.

Several toolboxes at different platforms including different algorithms were developed for feature extraction in Speech signal processing research area. The most popular of these toolboxes is Praat [5], followed by OpenSMILE [6], OpenEAR [7], HTK [3] and CSL [8] in that order. Other than these toolboxes there are various tools developed in small sizes. Table 1 provides a general summary of the toolboxes developed with the purpose of speech processing and feature extraction.

The following feature sets given in Table 1 consist of the corresponding features: cepstral feature set-MFCC, F0, voice quality feature set-HNR, jitter and shimmer, waveform feature set-zero-crossing, auditory feature set-PLP, formants feature set-formant frequencies and band-with, spectral feature set-energy and entropy, tonal feature set-CHROMA and CENS. In the studies there are also tools developed for researchers except the ready toolboxes given in Table 1 [12–15].

2. Materials and methods

2.1. SPAC toolbox design

SPAC is a toolbox developed in MATLAB environment with the purpose of speech processing. It has a user friendly structure with its graphical interface and graphical display of the results. Toolbox is designed in a modular structure and consists of 6 modules. These;

- Module 1: Configurations
- Module 2: Pre-Processing,
- Module 3: Feature Extraction,
- Module 4: Post-Processing,
- Module 5: Classification
- Module 6: Batch Analysis

The modular structure of SPAC is given in Fig. 1.

Module 1 contains the configuration of parameters such as frame size, windowing and overlap to be used for speech processing. Module 2 is a pre-processing step that includes methods for improving the signal before feature extraction. Module 3 is a speech feature extraction step and can be used on one or more files. Module 4 contains methods for feature normalization and feature selection on the obtained features. Classification can be performed using Module 5, the attribute set obtained by SPAC, or the attributes in an external file. Module 6 enables speech processing and feature extraction operations on multiple files. Since the SPAC toolbox is modularly designed, each module can be used independently. For example, attributes obtained with another toolbox can only be classified on the SPAC with the classification module, or only the pre-processing module can be used to improve the sound quality. A screen shot of processing of voice file with SPAC is given in Fig. 2.

In Fig. 2, the screen shot of processing of a sample voice file with SPAC is provided. Original speech section includes the voice file which is being processed and analysis section includes feature extraction and visualization from voice file. With SPAC each feature can be examined through GUI as seen in the figure and pre-processing can be applied such.

2.1.1. Module 1: Configurations

This module contains the settings to be used for speech processing and feature extraction. Screen shot of Module 1 is given in Fig. 3.

Since the speech is stable in small time intervals, the speech signal is processed by dividing the frames. The Frame Length parameter determines the frame length to be used for speech processing. It is generally preferred to be between 10 and 20 ms [16]. In order to avoid introducing spurious frequency components when each frame of the speech signal is processed, the next frame covers a certain part of the previous one. This process is called overlapping. Windowing is used to adjust the spectral leak on the signal and the intersection due to overlap. Hamming, Hann, Bartlett, Gaussian and Rectangular methods can be used for windowing on SPAC and the method to be used on the module can be selected. Speech of humans have F0 value in a certain interval. For this reason, the frequency range to be used in speech processing can be determined by Maximum-Minimum F0 in order not to consider frequencies outside the human speech on the signal. MFCC Cepstrum Count specifies the MFCC coefficient to be used when

Table 1
Speech processing toolboxes.

Toolbox	Platform	Access	Interface	Extracted Features
Praat [5]	C++	Free	GUI	Signal energy, FFT spectrum, cepstral, F0, voice quality, LPC, auditory, formants, spectral, H1-H2, H1-A1, H1-A2, H1-A3, speech rate
OpenSMILE [6]	C++	Free	Command Line + GUI	Waveform, signal energy, Loudness, FFT spectrum, ACF, cepstrum, Mel/Bark spectr., cepstral, F0, voice quality,
OpenEAR [7]	C++	Free	Command Line + GUI	LPC, auditory, formants, spectral, tonal
HTK [3]	C	Free	Command Line	Signal energy, cepstral, auditory, Mel/Bark spectr., LPC
Voicebox [9]	Matlab	Free	Matlab function	Waveform, signal energy, F0, LPC, cepstral, Mel/Bark spectr.
COLEA [10]	Matlab	Free	GUI	F0, formants, spectral, signal energy
SPEFT [11]	Matlab	NA	GUI	Waveform, signal energy, voice quality, F0, formants, cepstral, auditory, LPC, Mel/Bark spectr
SPAC	Matlab	Free	GUI	F0, formants and bandwidths, voice quality (SNR, SNR Power Jitter, Shimmer), LPCC, MFCC, ZCR, signal energy, speech rate, pause rate, number of pause, wavelet

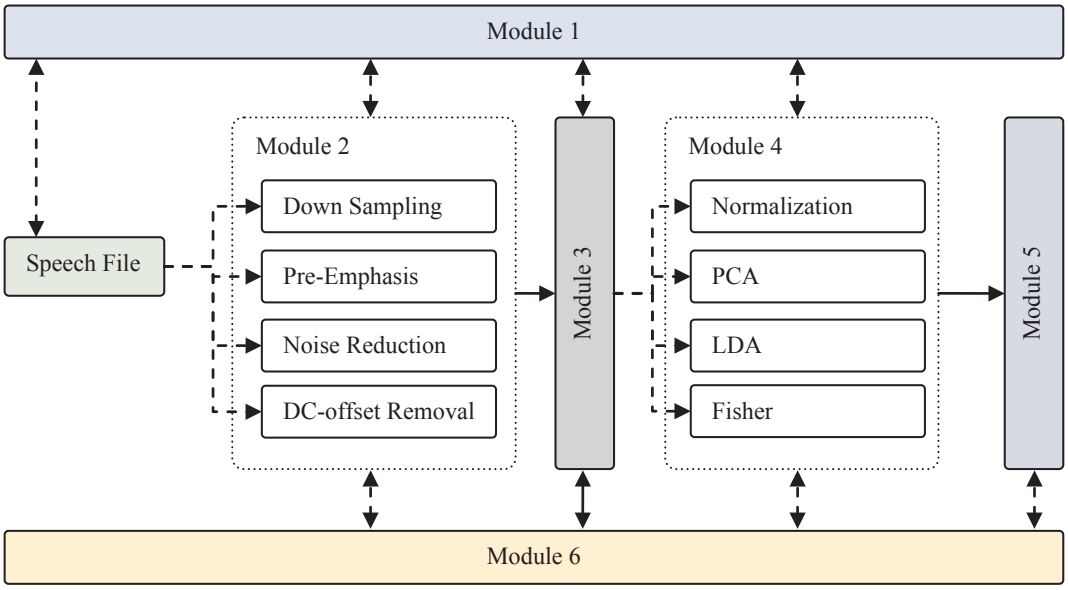

Fig. 1. The modular structure of SPAC.

detecting MFCC attributes over the signal. The Pre-Emphasis Filter specifies the value of the coefficient to be used for the filter. Voicing Threshold is the energy threshold used to identify areas with and without speech. The F0 Detection Method allows the selection of the method to be used to detect F0. Autocorrelation and Cepstrum are two methods for F0 detection. Inverse filtering is used to determine the shape of the voice signal in the larynx. Thus, Glottal Closure Instant (GCI) can be detected by obtaining the shape of the voice in the throat. SPAC consists of two methods, namely DYPSA [9] and SE_ SE_VQ [17], for identifying GCI points. The method to be used with the GCI Algorithm on the module is selected. The GCI values obtained here are used to determine the Jitter features. Mean Normalization determines whether normalization is performed by subtracting the average value of the signal from the signal. Voiced/Unvoiced Control determines which part on the voice signal is to be considered: parts with or without speech. This preference plays an important role in the acoustic features which

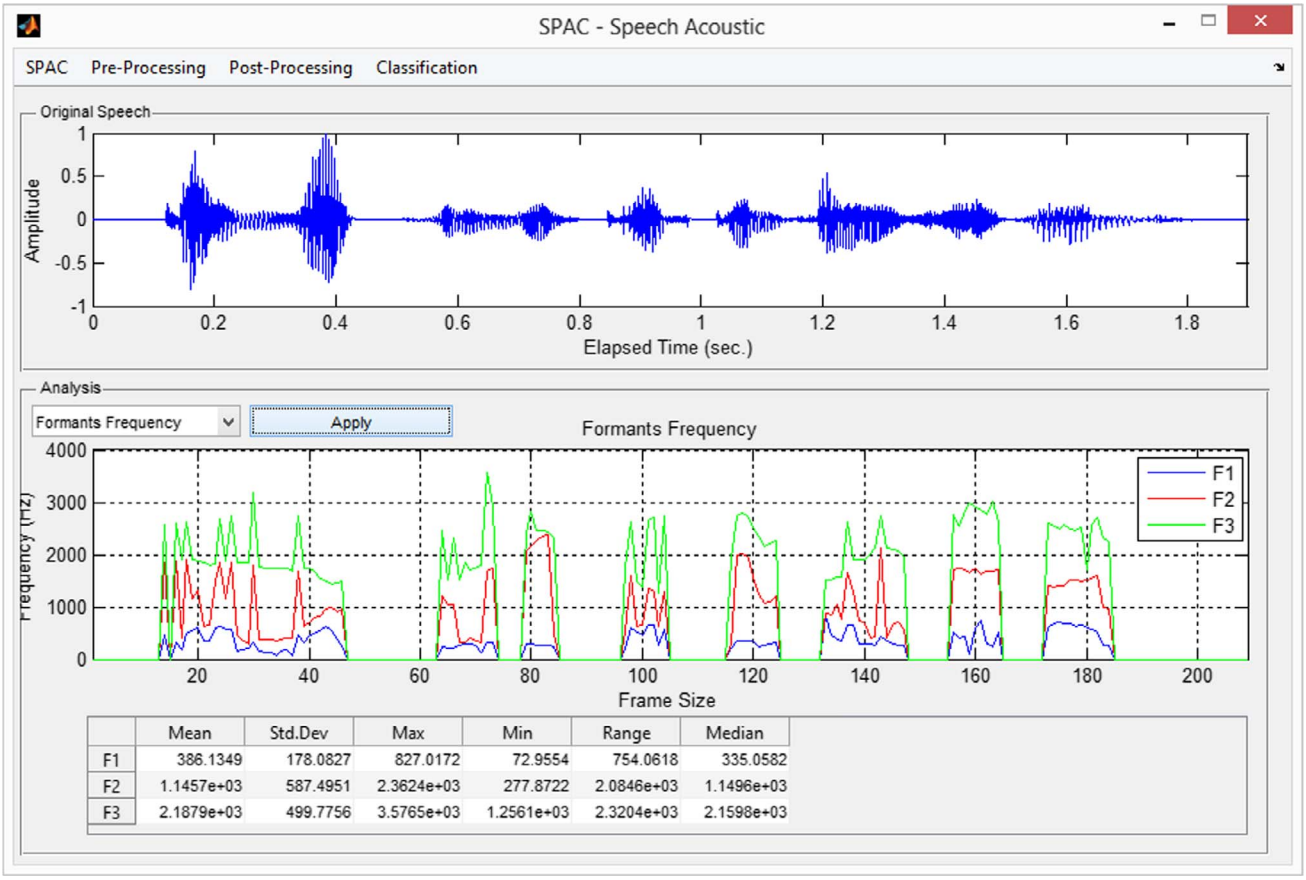

Fig. 2. Example of speech processing with SPAC.

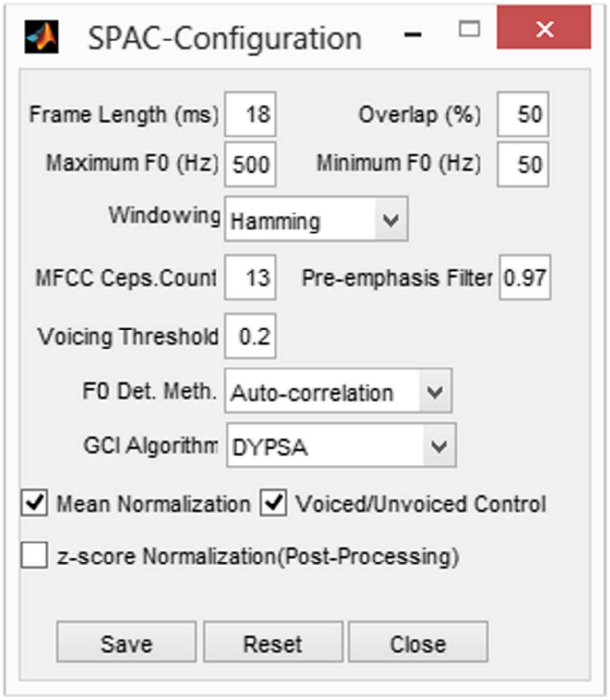

Fig. 3. GUI of the Module 1.

will be extracted. Z-score normalization determines whether normalization is applied to the SPAC-acquired attributes.

2.1.2. Module 2: Pre-processing

Pre-processing is applied to the pre-feature extraction speech signal to improve the performance of the feature extraction algorithms. SPAC toolbox includes down sampling, pre-emphasis, noise reduction and DC offset removal methods and combinations of these methods.

Down sampling is used to reduce the workload by reducing the sampling value of the signal. High sampling values are not a problem, but can increase workload.

Pre-emphasis enhances the Signal to Noise Ratio (SNR) value by amplifying harmonics with low amplitude in the signal [18].

The noise is defined as undesirable signals which cause deterioration in signal in such applications as communication, measuring and signal processing [19]. The noise reduction method should not cause any change in the structure of the main signal to completely remove the noise. For this reason, many methods provide noise suppression at 10–20 dB [20]. SPAC uses wavelet transform for noise reduction.

DC offset is the mean amplitude value of the wave. If mean value is 0 there is no DC offset. This means that the amounts of signal in positive and negative zones are not equal.

2.1.3. Module 3: Feature extraction

Speech feature extraction is the basic requirement for all of the speech processing applications, and the main purpose is to derive descriptive attributes from the signal. With SPAC, about 723 attributes can be obtained with 9 main headings and their statistical variations, and it changes according to the cepstrum count value. Table 2 provides the features that can be obtained with SPAC.

F0 is the frequency at which the vocal cords are opening and closing. On SPAC, autocorrelation and cepstrum methods were used. Which method can be used can be determined via Module 1.

Formant is the resonance in vocal tract. In theory there are infinite number of formats but in practice it only consists of the first 3 or 4 formats. LPC analysis is used for determination of format frequencies.

In a complex voice, exact multiples of fundamental frequency constitute harmonics. SNR is the ratio of the total energy of F0 and harmonics which are its exact multiples to the noise energy. Matlab Signal Processing Toolbox has been used for determination of SNR.

The changes in Jitter indicate the pathologies in vocal folds. As the case in F0 perturbation, short-time amplitude changes in voice signals are measured. The periodic variation between amplitude peaks is called shimmer. This parameter is defined as the glottal flow amplitude changes between cycles with subsequent vibration [21]. GCI algorithm selected from Module 1 is used in calculating Jitter.

Linear Prediction is widely used in speech recognition and synthesis systems, as an efficient representation of a speech signal's spectral envelope [11]. The auto-correlation has been used for calculating Linear Predictive Filter Coefficients (LPC).

The signal is defined by band passing filters and their energies. Thus, estimation of signal spectrum is ensured. Filters used in the structure of Mel-frequency Cepstral Coefficients (MFCCs) consists of a triangular series of filters in equal Mel-scale intervals. Mel is a unit designed to simulate the perceptive features of human ear [22]. The Mel-scale is a logarithmic mapping from physical frequency to perceived frequency. The cepstral coefficients extracted using this frequency scale are called MFCCs [11].

Zero-crossing rate is the number of passing from zero of the amplitude value of speech signal in a certain time interval [23].

The areas remaining below the threshold energy value determined in Module 1 are the silent areas, the others are areas including speech. Speech rate shows the percentage of the signal which includes speech. Pausing or remaining silent reflect mental and emotional aspects [24]. Pause rate gives the rate of passing during speech and number of pause gives the number of pausing.

Using Wavelet transform, 5 level decomposition is done on the signal. Approximation (A) coefficients show the low-frequency components of the signal, detail (D) coefficients show the high-frequency components of the signal. The wavelet coefficients have been calculated for each level as a result of decomposition. Matlab Wavelet Toolbox has been used for determining Wavelet coefficients.

Delta (Δ) and Delta-delta (Δ^2) which are used in several feature sets is the addition of time derivative for relevant feature. It has been realised on SPAC with delta coefficients simple linear slope.

2.1.4. Module 4: Post-processing

Normalization and feature selection operations on the feature set obtained after feature extraction can be performed through Module 4. Normalization is used to remove unit differences between the values contained in the attributes. Feature selection is used to reduce the size of the attribute set. Normalization and feature selection are used to improve classifier performance. SPAC contains z-score for normalization and PCA, LDA and Fisher selection methods for feature selection.

The Eq. (1) provides z-score normalization method for an x feature [25];

$$x_n = \frac{x-\mu}{\sigma} \tag{1}$$

where μ is the mean of x value and σ is the standard deviation.

PCA is a dimensionality reduction method aiming to find projections, which explain most of the remaining variance in the data. When PCA is used for size reduction, the size with lower deviation value is erased and, if necessary, datum is transformed back to its previous size.

LDA is a supervised feature reduction method which searches for the linear transformation that maximizes the ratio of the determinants of the between-class covariance matrix and the within-class covariance matrix [26].

Fisher selection algorithm [27] is a statistical method which is frequently used in obtaining information with individual attributes. The method uses in measuring method the mean of numerical values with different attributes for each class and standard deviation values for each attribute.

Table 2
SPAC feature set.

Feature	Statistical value	Feature size
Fundamental Frequency (F0)	Mean, Std. Dev., Max, Min, Range, Median, Δ, Δ^2	18
Formants (F1, F2, F3 and bandwidths)		108
Voice Quality (SNR, SNR Power Jitter, Shimmer)	SNR and SNR Power (Mean, Δ, Δ^2), Jitter and Shimmer (Value)	18
LPCC	Mean, Std. Dev., Max, Min, Range, Median, Δ, Δ^2	Ceps. countX18
MFCC		Ceps. countX18
Zero-Crossing Rate		18
Signal Energy		18
Speech (Speech Rate, Pause Rate, Number of Pause)	Value	3
Wavelet (D1, D2, D3, D4, D5, A1)	Mean, Std. Dev., Max, Min, Range, Median, L1 Norm, L2 Norm, Max Norm, Mean Abs., Median Abs., Mode	72

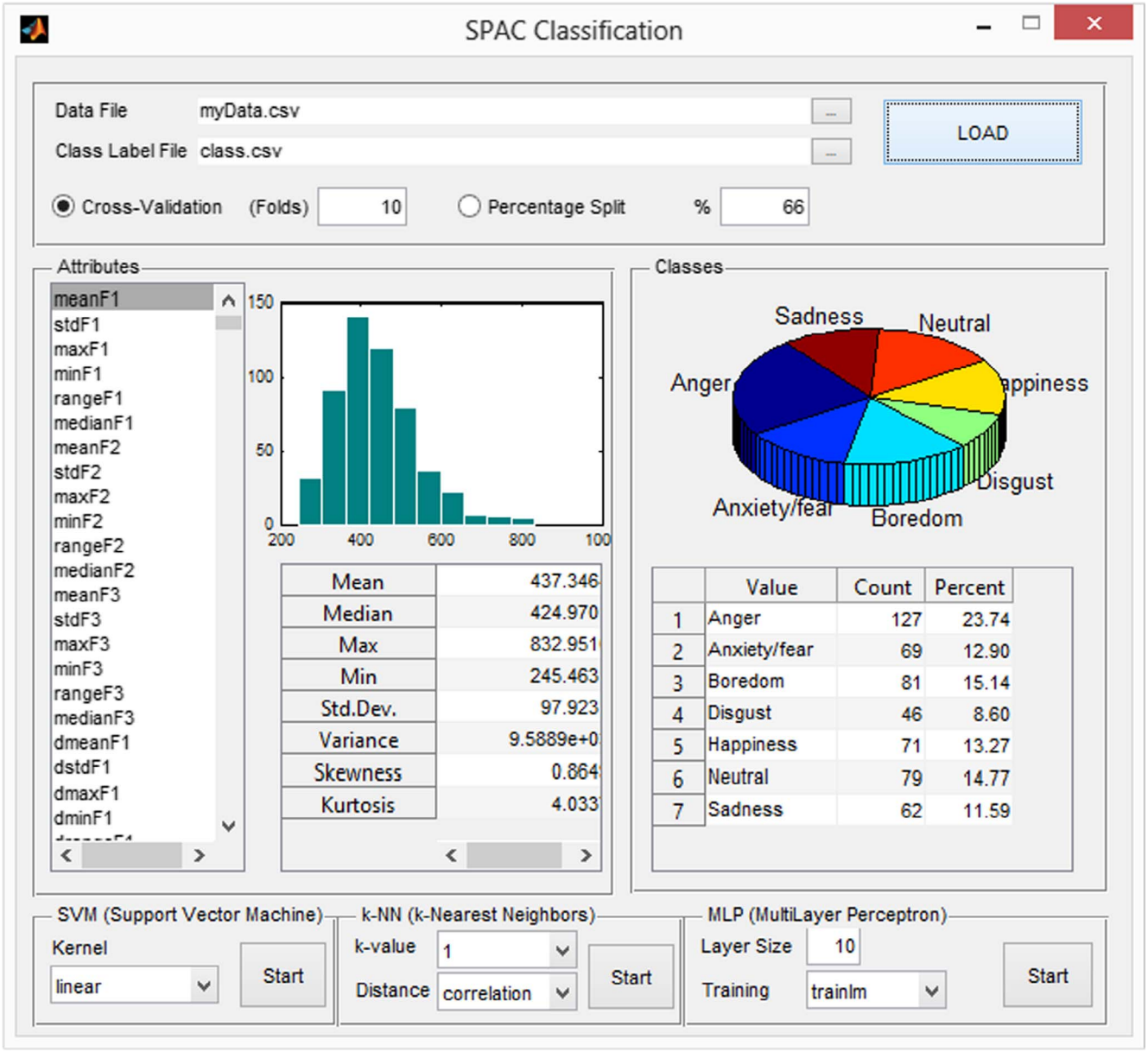

Fig. 4. GUI of the Module 5.

2.1.5. Module 5: Classification

In speech processing applications, classification processes such as emotion, speaker, text are performed after feature extraction. Module 5 consists of SVM, MLP and k-NN classifiers for classification through extracted features. However, as these classifiers are supervised learning, the classes of the used data have to be described. This procedure can be conducted with Class Label File option which is located in the module. GUI belonging to Module 5 is given in Fig. 4.

Through classification module given in Fig. 4, the distribution of statistical data belonging to features and distribution of classes can be

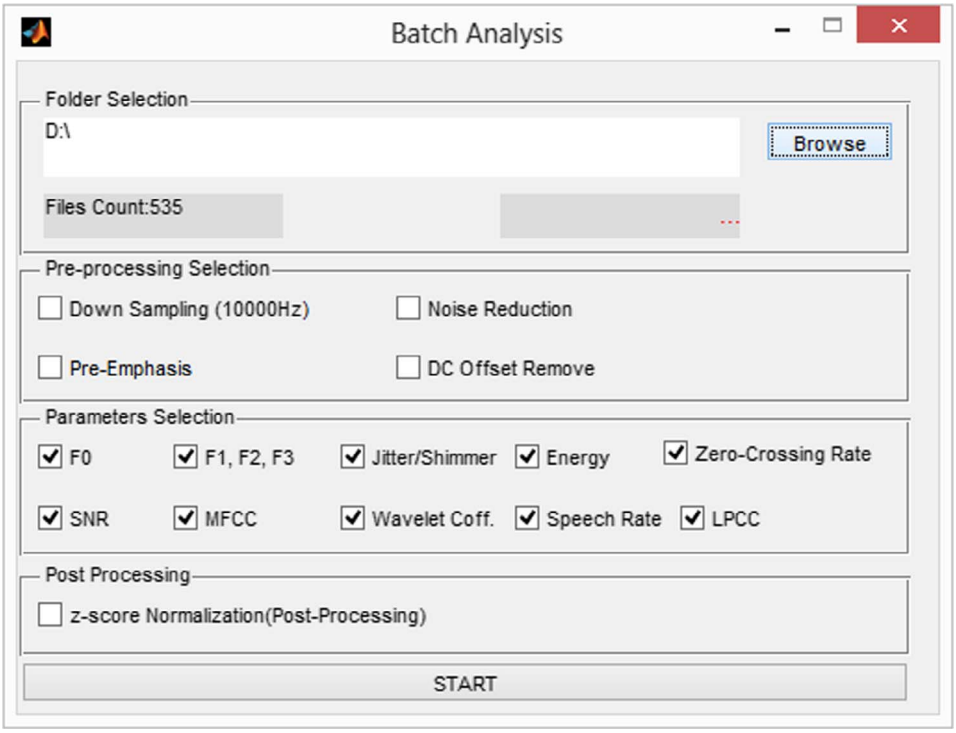

Fig. 5. GUI of the Module 6.

Table 3
MSE and MPE results.

Feature	OpenSMILE (x)-Praat (y)		OpenSMILE (x)-SPAC (y)		Praat (x)-SPAC(y)	
	Avg. MSE	Avg. MPE	Avg. MSE	Avg. MPE	Avg. MSE	Avg. MPE
F0	11.494	4.2%	2.195	1.5%	0.027	0.2%
Formants	–	–	–	–	20.156	5.7%
Voice Quality	–	–	–	–	0.021	2.9%
LPCC	–	–	–	–	0.001	8.7%
MFCC	14.594	70.3%	0.231	30.4%	0.029	9.7%
Zero-Crossing Rate	0.021	5.6%	0.018	5.2%	0.078	11.4%
Signal Energy	0.253	13.1%	0.383	16.2%	0.014	2.7%
Speech	–	–	–	–	0.299	5.3%

Table 4
The classification accuracy of SPAC feature set.

Feature Set	Feature Size	Accuracy (%)		
		SVM	k-NN	MLP
MFCC	234	69.5	58.7	74.1
LPCC	234	67.5	59.4	71.3
Wavelet	72	62.3	49.2	60.3
Others	183	68.6	51.3	72.3
ALL	723	82.1	65.4	85.1

monitored.

2.1.6. Module 6: Batch analysis

Module 6 is used for performing operation on multiple files. It has a user-friendly interface with the GUI in its content and all operations that are performed in other modules can be applied to multiple files by means of this module. GUI belonging to Module 6 is given in Fig. 5.

The folder which includes the files for operation are selected by Folder Selection panel which is found on GUI. Pre-processing Selection panel is used to determine the pre-processing methods to be used, Parameters Selection panel is used to determine the features to be obtained from voice files and Post-processing Panel is used to determine whether or not normalization will be performed. The features obtained after analysis are kept in mat file.

3. Results and discussion

Two methods are used to evaluate the accuracy of features obtained with SPAC. In the first one SPAC feature set and Praat and OpenSMILE feature sets have been compared. In the other method the classification success of features extracted with SPAC have been tested.

3.1. Testing data

Twenty speech file have been chosen from 4 different speech databases (EMO-DB, EMOVA, eNTERFACE05 and SAVEE) in order to test SPAC toolbox. The 5 samples were taken from each database.

Berlin Database of Emotional Speech (EMO-DB) was obtained by expression of different emotions by actors. Voice records are 16 bit mono and have a sampling frequency of 16 kHz [28]. EMOVA is a database built from the voices of up to 6 actors who played 14 sentences simulating emotional states. The recordings were performed with a sampling frequency of 48 kHz, 16 bit stereo, wave format [29]. The eNTERFACE'05 is an audio-visual emotion database. The database contains 42 subjects, coming from 14 different nationalities [30]. Surrey Audio-Visual Expressed Emotion (SAVEE) Database recorded an audio-visual emotional database from four native English male speakers, one of them was postgraduate student and rest were researchers at the University of Surrey [31].

3.2. Testing method

Mean Square Error (MSE) and Mean Percentage Error (MPE) methods were used while comparing the feature sets in validation of SPAC toolbox. In both methods standard data was obtained with Praat and OpenSMILE and, test data was obtained with SPAC. When standard

Table 5
The overall accuracy of SPAC, Praat, OpenSMILE feature sets.

Toolbox	Feature Size	Accuracy (%)		
		SVM	k-NN	MLP
SPAC	723 (F0, F1, F2, F3, jitter, shimmer, ZCR, Energy, MFCC, LPCC, Wavelet)	82.1	65.4	85.1
Praat	123 (F0, F1, F2, F3, jitter, shimmer, HNR, intensity, MFCC, LPCC)	83.4	68.8	85.2
OpenSMILE	1583 (Loudness, MFCC, logMelFreqBand, lspFreq, F0, jitter, shimmer)	85.2	61.4	87.6

data is given as X and test data is given as Y [11];

Mean Square Error (MSE):

$$E_{MSE} = \frac{1}{N}\sum_{i=1}^{N} |x_i - y_i|^2 \tag{2}$$

Mean Percentage Error (MPE):

$$E_{MPE} = \frac{1}{N}\sum_{i=1}^{N} \frac{|x_i - y_i|}{|x_i|} \times 100 \tag{3}$$

where x_i is the i th element in vector X.

In another method used for validation of the Toolbox, classification performances of feature sets obtained with SPAC, Praat and OpenSMILE have been compared. SVM, k-NN and MLP were used as classifiers.

3.3. MSE and MPE results

MSE and MPE were used in order to evaluate the features obtained with OpenSMILE, Praat and SPAC toolboxes and the results are given in Table 3.When examining the results in Table 3, it can be seen that there are variances between SPAC feature vectors and existing speech toolboxes and that the size of variance depends on the relevant feature set. When F0 feature set is examined, 4.2% variance occurs between OpenSMILE and Praat whereas 1.5 variance occurs between OpenSMILE and SPAC and 0.2% variance emerges between Praat and SPAC. SPAC feature sets do not include huge variances compared to the results of other toolboxes, due to which their validity can be accepted. In general sense feature sets obtained with SPAC are closer to Praat.

3.4. Classification results

Another method used to validate SPAC toolbox is usage of features obtained with SPAC in classification. For this purpose, feature extraction was performed from 535 voice files included in EMO-DB database with OpenSMILE, Praat and SPAC toolboxes. The success of feature sets obtained from the three toolboxes was analyzed with SVM, k-NN and MLP classifiers. The linear kernel type for SVM and the euclidean distance measure for k-NN are used. The epoch 500, learning rate of 0.3 and hidden layer neuron number (number of features + number of class)/2 were used for MLP. The 723 features obtained with SPAC are classified with SVM, k-NN and MLP classifiers and results are given in Table 4. The “others” feature set includes F0, formants, voice quality, ZCR, Energy and speech features.

When examining Table 4, it is seen that the classification accuracy is considerably reduced by using only one feature set, but it is increased when more than one feature sets is used. Classification accuracy is most increased by MFCC feature set followed by others, LPCC and wavelet. When the results in all feature sets is examined acceptable success is obtained by SPAC feature set in all classifiers.

In another classification analysis, the classification accuracy of SPAC was compared with the classification accuracy of OpenSMILE and Praat feature sets and overall success results are given in Table 5 and classification results are given in Fig. 6.

According to Table 5 and Fig. 6, the success obtained with SPAC is close to the results obtained with other toolboxes with 1%-3% variance. In addition, classification was performed for each class. This is an indication of the validity of feature sets obtained with SPAC.

4. Conclusions

In this work, we developed a toolbox called SPAC for MATLAB based speech processing and feature extraction. There are various toolboxes prepared for this purpose in the literature, but they have advantages and weaknesses compared to each other. SPAC's superiority over existing toolboxes is that it has an easy-to-use user-friendly interface, it is modular, allows graphical representation of results, includes classification module and allows to work with SPAC data or data obtained from different toolboxes. In addition, operations performed with other tools can be performed more easily with SPAC. The validity of the SPAC was confirmed by comparative analysis with Praat and OpenSMILE. For this purpose, two tests have been performed on the feature sets, including validation testing and classification. In the validation test, the MSE and MPE values for each attribute set are also calculated for the three toolboxes. According to the obtained results, the SPAC attributes vary between 0.2% and 9.7% compared to other toolboxes. The differences in the attributes of the OpenSMILE and Praat

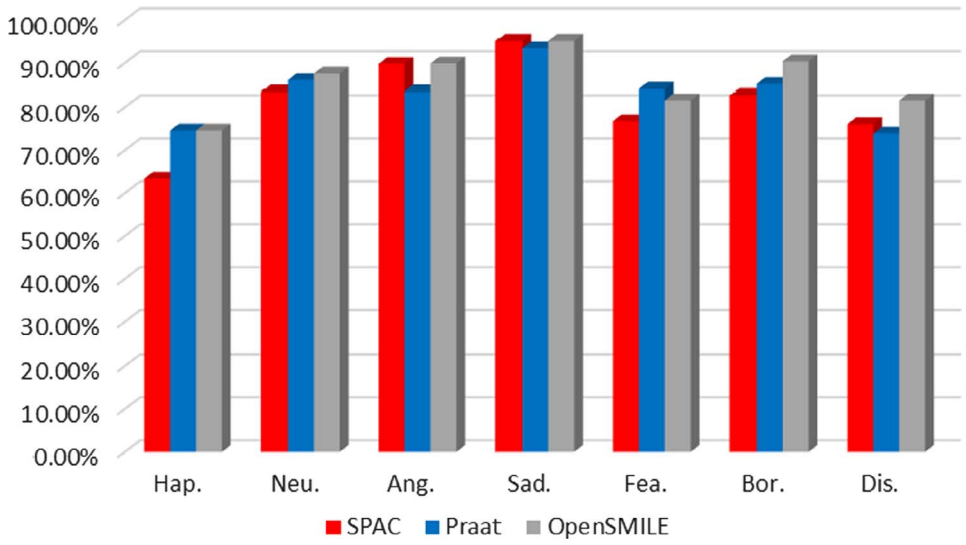

Fig. 6. The class-based accuracy of SPAC, Praat, OpenSMILE feature sets with SVM.

toolboxes accepted in the literature vary between 4.2% and 5.6%. These results confirm the validity of the feature sets obtained by SPAC. The classification results are compared as another method for testing the correctness of the attribute clusters. For this purpose, feature extraction was performed with three toolboxes via EMO-DB and these attributes were classified with SVM, k-NN and MLP. According to the results obtained, the success achieved with SPAC feature set is closer to other toolboxes with 1% -3% difference. In addition, SPAC has implemented classification for each class in the database. This suggests that the attributes obtained with SPAC have validity. In addition, SPAC provides platform independence and simplicity, as compared to the tools developed by other researchers,

Compliance with ethical standards

Conflict of Interest Turgut Özseven and Muharrem Düğenci declares that they have no conflict of interest.

Informed Consent All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Additional informed consent was obtained from all patients for which identifying information is included in this article.

Human and Animal Rights This article does not contain any studies with human or animal subjects performed by the any of the authors.

References

[1] Tien D, Liang Y, Sisodiya A. Speech feature extraction and data visualisation-vowel recognition and phonology analysis of four Asian ESL accents. Inf Technol Ind 2015;3(1):16–22.

[2] Huang Z, Xue W, Mao Q. Speech emotion recognition with unsupervised feature learning. Front Inf Technol Electron Eng 2015;16:358–66.

[3] Bou-Ghazale SE, Hansen JH. A comparative study of traditional and newly proposed features for recognition of speech under stress. Speech Audio Process IEEE Trans 2000;8(4):429–42.

[4] Loizou P. Colea: A MATLAB software-tool for Speech Analysis. Univ. Ark; 2003. May.

[5] Boersma P, Weenink D. Praat, a system for doing phonetics by computer. Glot Int 2002;5(9):341–5.

[6] Eyben F, Wöllmer M, Schuller B. Opensmile: the munich versatile and fast open-source audio feature extractor. Proc Int Conf Multimedia. 2010. p. 1459–62.

[7] Eyben F, Wollmer M, Schuller B. OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. 2009. p. 1–6.

[8] Computerized Speech Lab, Kay Elemetrics, Kay Elemetrics . [Online]. Available: < http://www.kaypentax.com/index.php?option=com_product&controller=product&Itemid=3&cid%5B%5D=11&task=pro_details > . [Accessed: 02-Apr-2015].

[9] M. Brookes et al. Voicebox: Speech processing toolbox for matlab, Softw. Available Mar 2011 Www Ee Ic Ac Ukhpstaffdmbvoiceboxvoicebox Html, 1997.

[10] Hansen JH, Bou-Ghazale SE, Sarikaya R, Pellom B. Getting started with SUSAS: a speech under simulated and actual stress database. Eurospeech 1997;97:1743–6.

[11] Li X. SPEech Feature Toolbox (SPEFT) Design and Emotional Speech Feature Extraction Doctoral dissertation Faculty of Graduate School, Marquette University; 2007.

[12] M. Slaney. Auditory toolbox: a Matlab toolbox for auditory modelling work. In: Tech Rep 45, Apple Technical Report, Apple Computer Inc, 1994.

[13] lingWAVES. *lingWAVES*. [Online]. Available: < https://www.wevosys.com/products/lingwaves/lingwaves.html > . [Accessed: 17-Jan-2016].

[14] Speech Analyzer. *SIL*. [Online]. Available: < http://www-01.sil.org/computing/sa/index.htm > . [Accessed: 17-Jan-2016].

[15] Welcome to talkbox documentation — talkbox v0.1 documentation. *Talkbox*. [Online]. Available: < http://www.ar.media.kyoto-u.ac.jp/members/david/softwares/talkbox/talkbox_doc/index.html > . [Accessed: 17-Jan-2016].

[16] Rabiner LR, Schafer RW. Digital processing of speech signals. Prentice Hall; 1978.

[17] Kane J, Gobl C. Evaluation of glottal closure instant detection in a range of voice qualities. Speech Commun 2013;55(2):295–314.

[18] McLoughlin I. Applied speech and audio processing: with Matlab examples. Cambridge University Press; 2009.

[19] Vaseghi SV. Advanced digital signal processing and noise reduction. John Wiley & Sons; 2008.

[20] Virtanen T, Singh R, Raj B. Techniques for noise robustness in automatic speech recognition. John Wiley & Sons; 2012.

[21] Deshmukh O, Espy-Wilson CY, Salomon A, Singh J. Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech. Speech Audio Process IEEE Trans 2005;13(5):776–86.

[22] Sethu V. Automatic emotion recognition: an investigation of acoustic and prosodic parameters Doctoral dissertation The University of New South Wales; 2009.

[23] Lokhande NN, Nehe DS, Vikhe PS. Voice activity detection algorithm for speech recognition applications. In: IJCA Proceedings on International Conference in Computational Intelligence (ICCIA2012), vol. iccia, 2012, p. 1–4.

[24] Cannizzaro M, Harel B, Reilly N, Chappell P, Snyder PJ. Voice acoustical measurement of the severity of major depression. Brain Cogn 2004;56(1):30–5.

[25] El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit Mar. 2011;44(3):572–87.

[26] Schuller B, Batliner A, Steidl S, Seppi D. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun Nov. 2011;53(9–10):1062–87.

[27] Duda RO, Hart PE. Pattern classification and science analysis. New York: Wiley-Interscience; 1973.

[28] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. Interspeech 2005;5:1517–20.

[29] Costantini G, Iaderola I, Paoloni A, Todisco M. EMOVO Corpus: an Italian Emotional Speech Database. LREC, 2014, p. 3501–4.

[30] Martin O, Kotsia I, Macq B, Pitas I. The eNTERFACE'05 audio-visual emotion database. In: Proceedings 22nd International Conference on Data Engineering Workshops, 2006, p. 8–8.

[31] Jackson P, Haq S, Edge JD. Audio-visual feature selection and reduction for emotion classification. Proc Int'l Conf Auditory-Visual Speech Process. 2008. p. 185–90.