PAN
ARCHIVES OF ACOUSTICS
Vol. **44**, No. 2, pp. 277–286 (2019)
DOI: 10.24425/aoa.2019.128491

Speech Emotion Recognition Based on Voice Fundamental Frequency

Teodora DIMITROVA-GREKOW[(1)*], Aneta KLIS[(1)], Magdalena IGRAS-CYBULSKA[(2)]

[(1)] *Faculty of Computer Science*
Bialystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
[*]Corresponding Author e-mail: t.grekow@pb.edu.pl

[(2)] *Faculty of Humanities*
AGH University of Science and Technology
Mickiewicza 30, 30-059 Kraków, Poland

(*received July 9, 2018; accepted January 7, 2019*)

The human voice is one of the basic means of communication, thanks to which one also can easily convey the emotional state. This paper presents experiments on emotion recognition in human speech based on the fundamental frequency. AGH Emotional Speech Corpus was used. This database consists of audio samples of seven emotions acted by 12 different speakers (6 female and 6 male). We explored phrases of all the emotions – all together and in various combinations. Fast Fourier Transformation and magnitude spectrum analysis were applied to extract the fundamental tone out of the speech audio samples. After extraction of several statistical features of the fundamental frequency, we studied if they carry information on the emotional state of the speaker applying different AI methods. Analysis of the outcome data was conducted with classifiers: K-Nearest Neighbours with local induction, Random Forest, Bagging, JRip, and Random Subspace Method from algorithms collection for data mining WEKA. The results prove that the fundamental frequency is a prospective choice for further experiments.

Keywords: emotion recognition; speech signal analysis; voice analysis; fundamental frequency; speech corpora.

1. Introduction

Nowadays, the trend of using computer speech processing in human-computer interaction has been developing very dynamically. There is a huge number of applications and information systems on the market which are voice controlled. Examples include: voice search in the Google browser (Google Now), which uses among other methods, deep machine learning (Yu, Deng, 2014), the Siri application for iPhone users, i.e. intelligent assistant and knowledge navigator. Programs of this type can control phone functions based on voice commands, e.g. save a note in the schedule, set a reminder, send a message, make reservations in restaurants, and many more. They also have the functionality of conducting conversations with people (Haleem, 2008; Igras, Ziółko, 2013). Hence the question: how would all these systems improve with the inclusion of an emotion recognition module – an extra feature bringing the machine closer to the man. The addition of the emotion recognition module turned out to be very useful in systems searching musical compositions on databases (Grekow, Raś, 2010).

The subject of this work is to examine whether the laryngeal tone is a sufficient attribute to identify the emotional state of the speaker and how efficient emotion classification can be achieved on the basis of laryngeal tone. We try to answer the question: is it possible to determine the emotional state of the speaker based mainly on changes of one feature?

This paper presents an investigation of emotion recognition in human speech using only fundamental frequency of voice (F0). Fast Fourier Transformation (FFT) and magnitude spectrum analysis were applied to extract the F0 out of the audio samples from the database of emotional speech recordings collected in AGH University of Science and Technology. The statistic functionals, describing datasets of the calculated F0 series were selected based on observations, calculations and considerations. Feature extraction was done using

a Java application written for this purpose. The outcome data were analysed with classifiers from a collection of machine learning algorithms for data mining (Hall *et al.*, 2009).

In order to analyse the impact of the emotions on the fundamental frequency behaviour various types of artificial intelligence methods have been used. There are implementations of Convolutional Neural Networks – CNN (Bertero, Fung, 2017), Artificial Neural Networks – ANN (Yashaswi *et al.*, 2015) and Tree Grammars – TGI (Bertero, Fung, 2017; Yashaswi *et al.*, 2015). Classifiers are frequently used approach, e.g. Hidden Markov Model – HMM (Bertero, Fung, 2017), K-Nearest Neighbour – KNN or Support Vector Machines – SVM (Bertero, Fung, 2017; Khan *et al.*, 2011). We verified several classifiers in this work, considering simplicity and accuracy as main improvement factors. We considered all seven emotions together and in several less numerous combinations. The test results are described and discussed at the end of this paper.

The next section shortly reviews some of the most interesting points in the voice-based emotion recognition attempts. Then our method is introduced – the idea, realization and the experimental results. At the end of the article conclusions are made and further plans are described.

2. Speech emotion recognition

Despite the fact that Aristotle claimed that a particular emotional state was associated with a particular tone of voice, it was only in the early 1990's that the topic started to noticeably draw the attention of scientists. These papers describe emotional communication models based on neural networks (Yamada *et al.*, 1995), emotion recognition based on time domain analysis for natural and synthesized speech (Heuft *et al.*, 1996) and many others. Further analysis of the published research of voice-based emotion recognition shows that there are many different approaches. The authors classify the affective state of speech in terms of various parameters characterizing the voice, e.g. basic tone, formants, intensity, vibrations, MFCC (Mel Frequency Cepstral Coefficients) and many others (Ananthakrishnan *et al.*, 2011; Kamińska, 2014). Some of the papers compare emotions between themselves and reference states, examine emotions in singing, spoken or spontaneous speech (Chua *et al.*, 2015). Other investigations analyze changes in the basal tone in infant crying (Maulida *et al.*, 2016). There are also studies comparing the recognition of emotions in quiet and loud environments (Kim *et al.*, 2007) and also issues focused on many attributes of the human voice or on individual features of speech.

A lot of research has already been done on the fundamental voice frequency. This attribute appears in most of the works on the recognition of emotions in the voice, however, it is not the only parameter and stands equally with factors like MFCC, LPC (Linear Predictive Codes), BFCC (Basilar-membrane Frequency-band Cepstral Coefficient), HFCC (Human Factor Cepstral Coefficients), PLP (Perceptual Linear Prediction), F1–F3 formants and signal energy (Kamińska, 2014). Depending on the parameters used, the methods of classification, language, age and sex of the speakers and many other factors achieved an accuracy of emotion recognition oscillating around 83% for 7 emotions (Emerich, Lupu, 2011), 79.5% for 8 states (Savargiv, Bastanfard, 2015), 76.66% for 9 (Firoz, Babu, 2017) or 77.1% in the case of 6 emotions (Soltani, Ainon, 2007). All the examples cited above are based on many parameters describing the human voice.

This paper focuses on the analysis of only one voice trait to check if it is a sufficient information medium describing a given emotion. F0 can be a promising parameter for several reasons. First of all, this feature can be easily extracted from speech signal with relatively high independence from environmental noise which makes it universal and possible to use in variety of different applications. What is more, its extraction is computationally cheap, which can help in applying it in mobile devices or real-time computing applications. Next, F0 is considered one of the main correlates of emotions, among energy, speech rate and voice quality, starting from works of Scherer and Banse (Scherer, 1986; Banse, Scherer, 1996; Scherer *et al.*, 2001; 2003; Scherer, 2003). Its correlations with emotions and also some other paralinguistic features (gender, age, identity) have been well investigated. It was found that pitch-related features are more specific for acted emotions than for natural emotional speech (Vogt, André, 2005), which makes F0 a good starting point for future comparative research. Also, using only this parameter we aim to construct a benchmark for further experiments with more features, if needed.

3. Methodology

The data processing starts from audio records, goes through initial and secondary treatment and thus, obtained parameters are finally evaluated by an artificial intelligence module. The entire process is shown in Fig. 1. In this section all the stages are discussed sequentially.

3.1. Emotional Speech Corpus

In the case of speech analysis, database choice greatly affects the trust and dependability results. The database of emotional speech recordings *Emotional Speech Corpus* developed in AGH University of Science and Technology Krakow, was chosen because of

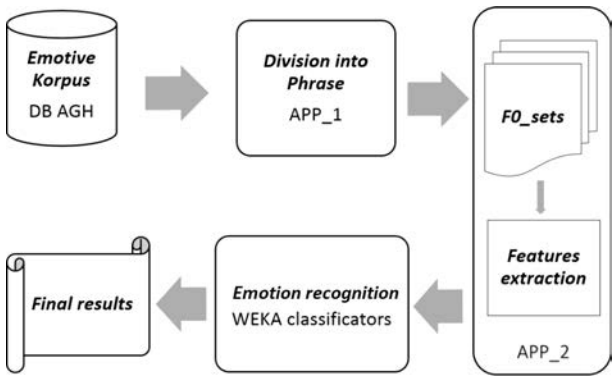

Fig. 1. Data processing – from the DB records to the emotion recognition.

its size and availability. Emotional Speech Corpus is available on an open source license for the aims of non-commercial research. The database contains audio records with the following moods:

- five basic emotions: joy, sadness, fear, anger and surprise,
- additional emotion: irony,
- neutral state.

All samples are performed by twelve persons aged 20–30, where half are female and half male. Also, half of the speakers are artists and the other half non-artists. The database includes 280 recordings, 40 per each emotion. They are divided into thematic groups (Kamińska, 2014). Statistically, each category has 214 words, which means that there are 2568 samples for each emotional set. However, as we believe that short, one-syllable words would be hardly helpful or even confusing, out of context, our experiment was conducted on phrases, considering simple words only if it was an independent expression, such as numbers, orders, etc.

3.2. Parameter extraction

Searching for emotional evidence or traces in the human voice supposes further from words specific, better global conclusions. In this study, we focused on the derivative information of voice melody: intonation and its changes over time. This was an important assumption to attain the results we intended. Primarily, speech carries pure verbal content, but there are also non-linguistic data in the human voice. Any data connected to verbal information could be regarded as a potential noise source. Thus, to achieve proper characteristics of the audio data, we focused on the laryngeal fundamental tone. By choosing such a base, the specifics of particular phonemes could be minimized or even excluded.

There are different methods to calculate fundamental frequency F0 at a given time duration: autocorrelation, zero crossing, complex cepstrum, etc. However, if we need to obtain how exactly F0 changes over the time of a phrase, it would be very time-consuming to determine the duration of each pitch period. It is much easier to use mathematical algorithms that calculate how F0 changes over time. To determine F0 we implemented Fast Fourier Transformation (FFT), finding the frequency of the largest peak. Although this method is not very precise, its outcomes meet our expectations. Using certain time-frames and overlapping (20 ms frames with 10 ms overlapping in this experiment), fundamental frequency series (FFS) were built for each phrase. The length of these FFS varies because of the different durations of the input records. For this reason, we respected the relative characteristics. The FFS parameters we calculated for each set had been primarily based of statistical functions.

The statistical measures we used to characterize the laryngeal tone changes in each audio file are: arithmetic mean, geometric mean, median, skewness, variance, the harmonic mean, standard deviation, kurtosis.

In the end, we added the decision attribute to each sample-record description. It had a value from 0 to 6, respective to the emotion in the file. The processed data was saved into a file, which was the input information for the machine learning system. These parameter collections were finally passed for analysis by WEKA.

3.3. Decision module

Data classification is part of a comprehensive field called data mining, the aim of which is to discover automatically unknown rules and dependencies in a set. Classification is a method determining the affiliation of an object to one of the predefined classes (Fatyga, Podraza, 2010). The purpose of this process is to predict the value of the decision attribute based on a set of features that describe the object. One of the most common approaches to classification is the construction of models, or classifiers, which on the basis of descriptors (features) determine the value of the decision attribute. The classification process is divided into two stages. At the beginning, a classifier describing a defined set of object classes is built. Then the classifier is used to predict the value of the decision attribute of objects for which assignment to a class is unknown. The first step is also divided into two stages. The set of sample objects (samples) is divided into two groups: training and test.

In the initial phase, known as teaching or training, a classifier is constructed on the basis of the training set, after which the accuracy of the classifier is calculated in the second phase, called testing, using the test set. In order to calculate the quality of the classifier, the accuracy coefficient should be calculated as a percentage of correctly classified objects from the test set. Classifiers are usually presented in the form of trees and decision tables, logical formulas and classification rules (Morzy, 2013). The classifiers we applied

were: K-Nearest Neighbours with local induction, Random Forest, Bagging, JRip, Random Subspace Method (Ho, 1998).

KNN (K-Nearest Neighbours) assigns an object to a class, based on a fixed distance measure, to which the largest number of its (object's) nearest neighbours belongs. The most commonly used distance metric are Manhattan, Euclidean, Czybyszew and Mahalanobis. The classifier learning process consists of selecting the parameter k, usually based on cross-validation. The biggest disadvantage of KNN is the very large computational cost ($O(n^2)$), caused by assigning the entire training set for each classified sample. An advantage of the method is the very good results in many applications. In this work, we implemented an extended version of the method: KNN with local induction, which adds an extra step, in which the classifier calculates the local metric units of each object (Skowron, Wojna, 2004).

Random Forest is a combination of tree predictors in which each tree depends on a random vector (Breiman, 2001). The algorithm generates many decision trees based on a random data set. Initially n-elements called pseudo-test are randomly chosen from the training set. Based on them, a tree is built, in which each node is subdivided by independent drawing of a subset of attributes. From the generated subset, the feature that is used to divide the subsample of a given node is selected. In the Random Forest method, the trees are built without cropping, which results in the appearance of homogeneous leaves, i.e. belonging to one decision class (Ho, 1995).

Bagging is a method that involves downloading from a teaching set a large number of subsets for which separate classification models are built. They are called a committee. For each new example, the prediction of each classifier is determined, and the final class is the one that was most often chosen by the classifiers (Andruszkiewicz, 2009). Possible irregularities in the training files have no significant impact on the classification process, as they constitute a small part of all randomly drawn examples (Adamczak, 2001).

Random Subspace Method RSM is a classifier based on team learning. Its function is to reduce the correlation between the estimators in the team by training them on a random subset of attributes and not on the whole set. Each team acting in the testing and training process uses only a subset of attributes, the number of which is specified. Groups select features independently and these subsets are randomly selected from all available data. This method is a parallel learning algorithm, i.e. an independent generation of decision trees. It is a random selection of a set of features and the solution of the initial problem on a reduced subset. In each iteration of the algorithm, weights are assigned to individual features that indicate their validity in the created model. The obtained values are the basis for assessing the validity of attributes in the final model (Ho, 1998).

JRip consists of a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) proposed by William (1995) and proceeds by treating all the examples of a particular judgment in the training data as a class. Thus, it finds a set of rules that covers all the members of that class. An initial set of rules is generated for each class. It repeats for all classes. Classes are analysed in increasing size.

3.4. Evaluation of the classification quality

The effectiveness of the classifier is determined on the basis of tests carried out on two sets: training and test. On the first of these collections, the classifier undergoes the correct classification learning process, while on the second the class examines the quality of the tested classifier in relation to the learning that it achieved during the training. The most well-known method of evaluating the classification quality is K-fold cross-validation. This method consists of a random division of the objects set into K with relatively evenly distributed subsets, so-called folds. During validation, the classifier is K-fold trained on a set consisting of K-1 samples and tested on the K-th part, which was not used during the learning process. For each iteration, another part of the set is tested, and each object is tested exactly once in the whole validation process (Skowron, Wojna, 2004).

3.5. Measures of the quality of classification methods

The foundation for assessing the quality of classification methods is the confusion matrix, which specifies in how many cases the test data were correctly classified by the model and how many errors occurred (Kołodziej *et al.*, 2011). In the case of a binary problem, the purpose of the classifier is to assign the object to a positive class or its rejection and to classify it in a negative class. Thus, the classifier has the option of taking one of four decisions, that is:

- TP (true positive) – correct indication of a positive class,
- FP (false positive) – incorrect indication of a positive class,
- TN (true negative) – correct indication of the negative class,
- FN (false negative) – incorrect indication of a negative class, i.e. a negative decision, while the object is in fact positive.

The most common measures to assess the quality of the classifier in this article are:

- Accuracy, or Correctly Classified Items (CCI): the probability of correct classification, defined by the formula:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}.$$

- Sensitivity or Recall (TP_{rate}) specifies the probability that the classification will be correct for a positive sample. It is defined by the formula:

$$\text{TP}_{\text{rate}} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

- Precision, or Positive Predictive Value (PPV), answers the question: what is the probability that the sample is negative if the result is positive? The PPV indicator is calculated on the basis of the formula:

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

- Balance assessment between sensitivity and precision, called F-measure or F1-score being the harmonic mean of PPV and TP_{rate} indicators is represented by the formula:

$$\text{F-measure} = \frac{2 \cdot \text{PPV} \cdot \text{TP}_{\text{rate}}}{\text{PPV} + \text{TP}_{\text{rate}}}.$$

4. Application presentation

Figure 1 shows a screen of the application's main window: F0_sets/Features extraction (APP). It consists of two separate units: F0_sets and Features extraction. Both modules were originally developed as parts of an application for the aims of this study. Figure 2 presents the GUI at the moment of choosing the frame and overlapping for the calculation of the F0 series. The Speech Emotion Recognition System is a young method collection having been developed at the Bialystok University of Technology for a year. The application for processing research data was written in the JAVA language using the Eclipse programming environment in version Neon 4.6.0 and JavaFX Scene Builder tools. Graphical User Interface (GUI) was implemented using the JavaFx platform included in the standard Java 8 package, which is the successor of the Swing library. JTransforms was used – the first, multithreaded and open FFT library written in the JAVA language. The code comes from the General Purpose FFT package created by Takuyo Ouro and from the Java FFT Pack by Boose Zhang.

Its main functions are to:

- get the F0 contours, or all sequences of F0 for a chosen group of files: main menu item *F0_sets*,
- extract appointed parameters from the calculated above F0 contours and to save them in a form suited to the further process of AI evaluation: button in the low right corner *Save_an_ARFF_file*.

In addition, the system also proposes:

- parameter adjustment for the F0 extraction, i.e. manually choosing the frame length and overlapping values for F0 calculation,
- attribute selection for the generated *arff* file: *FFS_Attributes*.

Under construction:

- automatic division of the audio records into utterances suitable for the research: main menu item *Split*,
- parameter adjustment for the splitting process, i.e. manually choosing the input and output level of the audio signal: *Split_settings*.

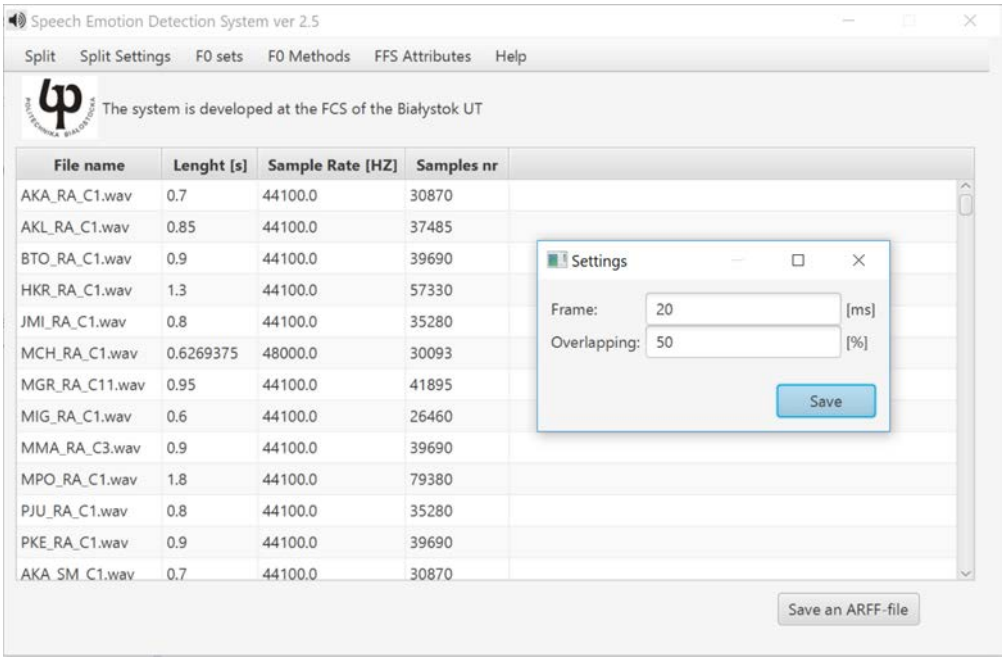

Fig. 2. Application supporting the speech data evaluation.

5. Experimental results

Each record had a sampling rate of 44100 Hz, a resolution of 16 bits, Signal to Noise Ratio (SNR) of approximately 40 dB. The tracks were recorded in a mono system, with the use of microphones: the capacitive AKG C5 VOCAL the dynamic AKG Shotgun C568 and the sound recorder Zoom H4N (Igras, Ziółko, 2013).

The first question that arose was: "How do the classifiers work on the total set of emotions?". So far, the answer is not satisfying. However, it provided us information about the workable ways of thinking. The initial results confirmed the classical psychological statement: the more opposite the emotions are, the better you distinguish them. We also got some hints about how each emotion relays to the others and thus which emotions could be better and which discriminate less. Hence our further consideration. The classifiers we applied were: K-Nearest Neighbours with local induction, Random Forest, Bagging, JRip and Random Subspace methods. We added two parameters to the statistical attributes of the F0 sequences, connected with the raw signal, which increased the primary results.

5.1. Preparation of input files and test conditions

The system for data generating (Fig. 1, *Division into Phrases block*) requires proper preparation of the input files. Each sample should contain an utterance consisting of individual words or a phrase. Examples of expressions are "Good morning," "Nice to see you", "Stop", "One", "Undo", "Thank you for your help". The average length of the recordings used in this work was 0.64 s, the minimum was 0.28 s and the maximum was 1.95 s.

For each of the emotions, 40 audio files were prepared with words or phrases which in total gives 280 elements for all the studied emotions. The choice of the audio entities and the actor was random, however it assured:

- at least four different persons,
- half of the recordings should be spoken by a male and half by a female actor,
- half of the recordings should be spoken by a professional and half by a non-professional actor.

The name of each audio file contains abbreviation of the: emotion implicit in, entity (type and number), actor. Based on the last, we applied an additional attribute – which group does the actor belong to: actor-male, actor-female, non-actor-male or non-actor female. The name consists of the words: *joy*, *sadness*, *anger*, *fear*, *irony*, *surprise* or *neutral*, the size of the characters does not matter. The decision attribute is generated on the basis of the title constructed in this way. The experiments are divided into three groups consisting of: two, three, and four emotions, respectively. In all tests, 40 records per emotion were considered.

We also used a number of abbreviations to show as much information as possible in the tables, while maintaining intelligibility and clarity:

J – joy,
S – sadness,
A – anger,
F – fear,
I – irony,
Sp – surprise,
N – neutral,
C1 – Local KNN,
C2 – Random Forest,
C3 – Bagging,
C4 – JRip,
C5 – Random Subspaces,
AvgE – Average accuracy for an emotion,
AvgC – Average accuracy for a classificator,
TP_{rate} – Sensitivity, Recall,
PPV – Precision,
Fm – F-measure,
WA – Weighted Average,
CCI – Correctly Classified Instances.

Fundamental frequency is calculated on frames of 20 ms with an overlapping of 50%.

5.2. Two emotions exploration

Opposed elements are always most likely to be distinguished. Applying this presumption to the emotions, we started our tests from exploring the most opposing *joy* and *sadness*. In fact, this has been also confirmed by the initial research, maintained in the introduction of this section. The common confusion table appeared especially interesting, but because of the unambiguous information we did not show it here.

All results obtained with the five methods of classification are shown in Tables 1 and 2. The first data set presents the accuracy on the studied methods, while Table 2 allows to compare the sensitivity, precision and the F-measure for the emotion couples in all classification methods.

Table 1. Accuracy of the classification for selected emotions couples.

	C1	C2	C3	C4	C5	AvgE
J_S	68.750	78.750	78.750	73.750	77.50	75.50
S_A	75.000	83.750	86.250	85.000	82.50	82.50
J_Sp	73.420	86.080	86.080	83.540	87.34	83.29
A_F	79.310	89.660	81.610	83.910	82.76	83.45
A_I	84.610	**89.740**	89.740	87.180	85.90	87.43
AvgC	76.218	85.596	84.486	82.676	83.20	

Table 2. Confusion matrix for selected emotions couples.

		E1 = J, E2 = S			E1 = S, E2 = A			E1 = J, E2 = Sp			E1 = A, E2 = F			E1 = A, E2 = I		
		TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm
C1	E1	0.650	0.700	0.680	0.700	0.780	0.740	0.700	0.760	0.730	0.800	0.760	0.780	0.780	0.910	0.840
	E2	0.730	0.670	0.700	0.800	0.720	0.760	0.770	0.710	0.740	0.790	0.820	0.800	0.920	0.790	0.850
C2	E1	0.800	0.780	0.790	0.880	0.810	0.840	0.880	0.850	0.860	0.930	0.860	0.890	0.890	0.920	0.900
	E2	0.780	0.790	0.790	0.800	0.870	0.830	0.850	0.870	0.860	0.870	**0.930**	0.900	**0.920**	0.890	0.900
C3	E1	0.830	0.770	0.800	0.880	0.850	0.860	0.880	0.850	0.860	0.800	0.800	0.800	0.900	0.900	0.900
	E2	0.750	0.810	0.780	0.850	0.870	0.860	0.850	0.870	0.860	0.830	0.830	0.830	0.900	0.900	0.900
C4	E1	0.880	0.690	0.760	0.850	0.850	0.850	0.850	0.830	0.840	0.880	0.800	0.830	0.900	0.860	0.880
	E2	0.600	0.830	0.700	0.850	0.850	0.850	0.820	0.850	0.830	0.810	0.880	0.840	0.840	0.890	0.870
C5	E1	0.800	0.760	0.780	0.830	0.830	0.830	0.900	0.850	0.880	0.830	0.800	0.800	0.830	0.890	0.860
	E2	0.750	0.790	0.770	0.830	0.830	0.830	0.850	0.890	0.870	0.830	0.850	0.840	0.900	0.830	0.860
Average	E1	0.792	0.740	0.762	0.828	0.824	0.824	0.842	0.828	0.834	0.848	0.804	0.820	0.860	0.896	0.876
	E2	0.722	0.778	0.748	0.826	0.828	0.826	0.828	0.838	0.832	0.826	0.862	0.842	0.896	0.860	0.876

The best result was an accuracy of 89.74%, achieved with the Random Forest and Bagging classifiers for the *Anger_Irony* couple. The average best results were achieved from the Random Forest. This method shows the best outcomes for each emotion couple tested in our research. Amazingly, *Joy_Sadness* were discriminated the worst. The best distinguishing was between *Anger_Irony* and *Anger_Fear*. Several more combinations, such as *Joy_Surprise*, *Fear_Sadness*, also had very high scores.

All outcomes pointed to a recognition. The lowest number was TPrate for *Joy* (vs. *Sadness*), determined from KNN classification. Only 0.3% were less than 0.7, and 22% were more than 0.88. The highest recall and precision were for *Anger_Irony*, calculated by Random Forest and KNN. Up to 0.90 are more than 10% of the results, mainly to the emotional couple.

The last data group, in the bottom of the table presents average precision and recall for each emotion in a particular couple. The best recognition was registered for *Anger* and *Irony* (the last case). Our expectations for good distinguishing of *Joy* vs. *Sadness* was not confirmed. However, surprisingly *Joy* was well distinguished from surprise (third case).

5.3. Exploration of three emotions

After the satisfactory results obtained in the preliminary tests, a third emotional state was involved. The research was conducted on several sets of three emotions. To achieve good distinguishment, we chose the most promising emotions from the database – these which showed a high specificity by the first stage of the conducted experiments. The most successful combinations were:

- anger, fear, irony,
- joy, anger, neutral state,
- sadness, anger, fear,
- joy, anger, surprise,
- joy, irony, surprise.

The total number of examined records for a set was 120 (40 per emotion). Tables 3 and 4 demonstrate the best combinations.

Table 3. Accuracy of the classification for selected three emotion sets.

3E_set	C1	C2	C3	C4	C5	AvgE
A_F_I	61.60	69.60	68.80	64.00	72.00	67.20
J_A_N	62.07	72.41	**76.14**	64.66	71.55	69.37
S_A_F	63.78	71.65	65.34	65.35	62.99	65.82
J_A_Sp	66.39	72.27	69.75	70.59	67.23	69.25
J_Sp_I	54.70	68.38	63.25	59.83	64.10	62.05
AvgC	64.08	71.48	70.01	66.15	68.44	

The best results achieved were with the combinations *Joy_Anger_Neutral* and *Joy_Anger_Surprise*. Also *Fear*, *Irony*, *Sadness* appeared in well distinguished sets. Again the *Random Forest* is the favourite classifier. Obviously, the accuracy was reduced, which is quite natural.

The recall and precision of the results are still promising. Generally, the best particular and average outcomes were achieved by *Random Forest*, *Bagging* and *Random Subsets*. The weakest was *KNN*.

5.4. Exploration of four emotions

The last experiment was with four emotion sets. The total number of examined records for a set was 160. Table 5 demonstrates the best accuracy we obtained for three combinations.

Although the results are barely positive, still some successful recognition might be denoted. Again, the

Table 4. Confusion matrix for three emotions sets.

		E1 = A, E2 = F, E3 = I			E1 = J, E2 = A, E3 = N			E1 = S, E2 = A, E3 = F			E1 = J, E2 = A, E3 = Sp			E1 = J, E2 = I, E3 = Sp		
		TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm	TP_{rate}	PPV	Fm
C1	E1	0.730	0.760	0.740	0.480	0.590	0.530	0.550	0.560	0.560	0.530	0.660	0.580	0.600	0.630	0.620
	E2	0.510	0.600	0.550	0.700	0.620	0.660	0.750	0.730	0.740	0.830	0.650	0.730	0.580	0.500	0.540
	E3	0.630	0.510	0.570	0.690	0.640	0.670	0.620	0.620	0.620	0.640	0.900	0.670	0.460	0.510	0.490
C2	E1	0.800	0.780	0.790	0.580	0.700	0.630	0.550	0.690	0.610	0.530	0.680	0.590	0.800	0.800	0.800
	E2	0.680	0.670	0.670	0.730	0.630	0.670	0.850	0.790	0.820	0.800	0.670	0.730	0.630	0.630	0.630
	E3	0.620	0.740	0.620	**0.890**	**0.870**	**0.880**	0.750	0.670	0.700	0.850	0.830	0.840	0.620	0.620	0.620
C3	E1	0.800	0.800	0.800	0.580	0.680	0.620	0.480	0.680	0.560	0.650	0.650	0.650	0.830	0.770	0.800
	E2	0.640	0.670	0.650	0.800	0.730	0.760	0.800	0.700	0.740	0.630	0.640	0.630	0.580	0.540	0.560
	E3	0.630	0.600	0.620	0.860	0.820	0.840	0.680	0.600	0.640	0.810	0.800	0.810	0.490	0.580	0.530
C4	E1	0.680	0.750	0.710	0.500	0.540	0.520	0.400	0.590	0.480	0.730	0.630	0.670	0.830	0.670	0.750
	E2	0.680	0.590	0.630	0.630	0.580	0.600	0.830	0.830	0.830	0.680	0.780	0.700	0.450	0.530	0.490
	E3	0.550	0.600	0.580	0.830	0.830	0.830	0.720	0.570	0.640	0.720	0.620	0.750	0.510	0.540	0.530
C5	E1	0.880	0.800	0.830	0.530	0.660	0.580	0.380	0.560	0.490	0.630	0.660	0.640	0.830	0.720	0.770
	E2	0.660	0.720	0.690	0.750	0.680	0.710	0.730	0.670	0.700	0.630	0.610	0.620	0.580	0.560	0.570
	E3	0.630	0.630	0.630	**0.890**	0.800	0.840	0.770	0.630	0.690	0.770	0.750	0.760	0.510	0.630	0.560
Average	E1	0.778	0.778	0.774	0.534	0.634	0.576	0.472	0.616	0.540	0.614	0.656	0.626	0.778	0.718	0.748
	E2	0.634	0.650	0.638	0.722	0.648	0.680	0.792	0.744	0.766	0.714	0.67	0.682	0.564	0.552	0.558
	E3	0.612	0.616	0.604	0.832	0.792	0.812	0.708	0.618	0.658	0.758	0.780	0.766	0.518	0.576	0.546

Table 5. Accuracy of the classification for selected four emotion sets.

4E_set	C1	C2	C3	C4	C5	AvgE
S_A_F_I	49.09	60.00	54.55	47.27	53.94	52.97
J_A_F_N	48.47	59.51	57.06	57.06	60.12	56.44
S_A_I_N	52.60	**62.99**	61.04	53.25	60.39	58.05
AvgC	50.053	60.833	57.55	52.527	58.15	

best working method was Random Forest, with 60.83% correctly classified samples. None of the emotion sets crossed the 60% threshold, as average accuracy. However, the methods Bagging and Random Subsets also showed a distinguishment. The most well differed set was *Sadness_Anger_Irony_Neutral*. Despite this we did not get any result greater than 60% for bigger combinations.

6. Conclusions

The aim of this study was to prove the significance of F0 parameters to classify speech with emotion labels. The authors of the work focused on checking whether the laryngeal tone is sufficient information on the basis of which to detect emotions in speech.

The results reported in this paper are a representative part of the investigation that was done in order to 'test a direction' or to improve the irrelevance of this. Test results prove that the fundamental frequency is a prospective choice for further experiments.

The conducted tests gave positive results, which compared to the current state of knowledge obtained from the literature analysis included in the second section, are satisfactory and achieved an accuracy equal to 89.74% in the case of 2 emotions, 76.14% for 3 emotion sets, and 62.99% for four-emotion sets.

To summarize, laryngeal tone is a very promising basis for emotion recognition. The process of emotion recognition can be improved by using further features extraction.

References

1. Adamczak R. (2001), *Application of neural networks for the classification of experimental data* [in Polish: *Zastosowanie sieci neuronowych do klasyfikacji danych doświadczalnych*], Ph.D. Thesis, Department of Computer Science Methods, Nicolaus Kopernikus University in Toruń.

2. Ananthakrishnan S., Vembu N.A., Prasad R. (2011), *Model-based parametric features for emotion recognition from speech*, Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 529–534, Big Island, USA.

3. Andruszkiewicz P. (2009), *Metalearning and the possibility of improving the efficiency of classification* [in

Polish: *Metauczenie a możliwość poprawy skuteczności klasyfikacji*], Metody Informatyki Stosowanej, **3**, 5–18.

4. Banse R., Scherer K.R. (1996), *Acoustic profiles in vocal emotion expression*, Journal of Personality and Social Psychology, **70**(3), 614–636.

5. Bertero D., Fung P. (2017), *A first look into a Convolutional Neural Network for speech emotion detection*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5115–5119, New Orleans, USA.

6. Breiman L. (2001), *Random Forests*, Machine Learning, **45**, 48–156.

7. Chua G., Chang Q., Park Y., Chan P., Dong M., Li H. (2015), *The Expression of Singing Emotion – Contradicting the Constraints of Song*, Proceedings of 19th International Conference on Asian Language Processing, pp. 98–102, Soochow, China.

8. Emerich S., Lupu E. (2011), *Improving speech emotion recognition using frequency and time domain acoustic features*, Proceedings of Signal Processing and Applied Mathematics for Electronics and Communications Workshop, pp. 85–88, Cluj Napoca, Romania.

9. Fatyga P., Podraza R. (2010), *Data classification – an overview of selected methods* [in Polish: *Klasyfikacja danych – przegląd wybranych metod*], Zeszyty Naukowe Wydziału ETI Politechniki Gdańskiej. Technologie Informacyjne, **19**, 55–60.

10. Firoz Shah A., Babu Anto P. (2017), *Wavelet Packets for Speech Emotion Recognition*, Proceedings of 3th International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics, pp. 479–481, Chennai, India.

11. Grekow J., Raś Z.W. (2010), *Emotion based midi files retrieval system*, Advances in Music Information Retrieval, Studies in Computational Intelligence, vol. 274, pp. 261–284 Springer, Berlin, Heidelberg.

12. Haleem M.S. (2008), *Voice controlled automation system*, Proceedings of 12th IEEE International Multitopic Conference, pp. 508–512, Karachi, Pakistan.

13. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I.H. (2009), *The WEKA data mining software: an update*, SIGKDD: The community for data mining, data science and analytics, Explorations, **11**(1), 10–18.

14. Heuft B., Portele T., Rauth M. (1996), *Emotion in time domain synthesis*, Proceedings of 4th IEEE International Conference on Spoken Language Processing, **3**, pp. 1974–1977.

15. Ho T.K. (1995), *Random Decision Forest*, Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 278–282, Montreal.

16. Ho T.K. (1998), *The Random Subspace Method for Constructing Decision Forests*, Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence, **20**(8), 832–844.

17. Igras M., Ziółko B. (2013), *Database of emotional speech recordings* [in Polish: *Baza danych nagrań mowy emocjonalnej*], Studia Informatica, **34**, 67–77.

18. Kamińska D. (2014), *Emotion recognition based on natural speech* [in Polish: *Rozpoznanie emocji na podstawie mowy naturalnej*], Ph.D. Thesis, Department of Faculty of Electrical, Electronic, Computer and Control Engineering, Lodz University of Technology.

19. Khan M., Goskula T., Nasiruddin M., Quazi R. (2011), *Comparison between k-nn and svm method for speech emotion recognition*, International Journal on Computer Science and Engineering, **3**(2), 607–612.

20. Kim E., Hyun K., Kim S., Kwak Y. (2007), *Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments*, Proceedings of 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 689–694, Jeju Island, South Korea.

21. Kołodziej M., Majkowski A., Rak R. (2011), *The use of a support vector machine (SVM) to classify the EEG signal for the brain-computer interface* [in Polish: *Wykorzystanie maszyny wektorów wspierających (SVM) do klasyfikacji sygnału EEG na użytek interfejsu mózg-komputer*], Pomiary Automatyka Kontrola, **12**, 1546–1548.

22. Maulida N., Alfiah W., Pawestri D., Susanto H., Zaman M., Aritianto D. (2016), *Fundamental Frequency Evaluation of Infant Crying*, Proceedings of IEEE International Seminar on Intelligent Technology and Its Application, pp. 61–66, Mataram, Indonesia.

23. Morzy T. (2013), *Data mining* [in Polish: *Eksploracja danych*], PWN, Warszawa, pp. 83–104.

24. Savargiv M., Bastanfard A. (2015), *Persian speech emotion recognition*, Proceedings of 7th International Conference on Information and Knowledge Technology, pp. 1–5, Urmia.

25. Scherer K.R. (1986), *Vocal affect expression: A review and a model for future research*, Psychological Bulletin, **99**(2), 143.

26. Scherer K.R., Banse R., Wallbott H.G. (2001), *Emotion inferences from vocal expression correlate across languages and cultures*, Journal of Cross-Cultural Psychology, **32**(1), 76–92.

27. Scherer K.R., Johnstone T., Klasmeyer G. (2003), *Vocal expression of emotion*, [in:] Handbook of Affective Sciences, pp. 433–456.

28. Scherer K.R. (2003), *Vocal communication of emotion: A review of research paradigms. Speech communication*, **40**(1–2), 227–256.

29. Sidorova J. (2009), *Speech emotion recognition with TGI+.2 classifier*, Proceedings of the Student Research Workshop at EACL, pp. 54–60, Athens, Greece.

30. Skowron A., Wojna A. (2004), *K Nearest Neighbor Classification with Local Induction of the Simple Value Difference Metric*, [in:] J.F. Peters, A. Skowron [Eds.], Rough Sets and Current Trends in Computing, LNCS, 3066, pp. 229–234, Springer, Berlin, Heidelberg.

31. Soltani K., Ainon R. (2007), *Speech emotion detection based on neural networks*, Proceedings of 9th International Symposium on Signal Processing and Its Applications, pp. 1–3, Sharjah.

32. William W.C. (1995), *Fast Effective Rule Induction*, Proceedings of 12th International Conference on Machine Learning, pp. 115–123, Edinburg.

33. Vogt T., André E. (2005), *Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition*, Proceedings of IEEE International Conference on Multimedia and Expo, pp. 474–477, Amsterdam.

34. Yamada T., Hashimoto H., Tosa N. (1995), *Pattern recognition of emotion with Neural Network*, Proceedings of 21st International Conference on Industrial Electronics, Control, and Instrumentation, **1**, 183–187.

35. Yashaswi A.M., Nachamai M., Joy P. (2015), *A Comprehensive Survey on Features and Methods for Speech Emotion Detection*, Proceedings of International Conference on Electrical, Computer and Communication Technologies, pp. 1–6, Coimbatore, India.

36. Yu D., Deng L. (2014), *Automatic Speech Recognition: A Deep Learning Approach*, Springer.