VILNIUS UNIVERSITY

TATJANA LIOGIENĖ

HIERARCHICAL CLASSIFICATION OF SPEECH EMOTIONS

Summary of Doctoral Dissertation

Physical Sciences, Informatics (09P)

Vilnius, 2017

The dissertation work was carried out at Vilnius University from 2012 to 2016.

Scientific Supervisor

Assoc. Prof. Dr. Gintautas Tamulevičius (Vilnius University, Physical Sciences, Informatics – 09P).

The defense Council:

Chairman

Prof. Dr. Romas Baronas (Vilnius University, Physical sciences, Informatics – 09P).

Members:

Prof. Dr. Habil. Romualdas Baušys (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T),

Prof. Dr. Habil. Kazys Kazlauskas (Vilnius University, Physical Sciences, Informatics – 09P),

Prof. Dr. Audris Mockus (The University of Tennessee, USA, Physical Sciences, Informatics – 09P),

Assoc. Prof. Dr. Povilas Treigys (Vilnius University, Technological Sciences, Informatics Engineering – 07T).

The dissertation will be defended at the public meeting of the Council in the auditorium 203 of the Vilnius University Institute of Mathematics and Informatics on the 28th of September, 2017 at 14:00.

Address: Akademijos str. 4, LT-08412 Vilnius, Lithuania.

The summary of the dissertation was distributed on the 27th of August, 2017.

The dissertation is available at the library of Vilnius University.

VILNIAUS UNIVERSITETAS

TATJANA LIOGIENĖ

HIERARCHINIS ŠNEKOS EMOCIJŲ KLASIFIKAVIMAS

Daktaro disertacijos santrauka

Fiziniai mokslai, informatika (09P)

Vilnius, 2017

Disertacija rengta 2012–2016 m. Vilniaus universitete.

Mokslinis vadovas

doc. dr. Gintautas Tamulevičius (Vilniaus universitetas, fiziniai mokslai, informatika – 09P).

Disertacija ginama viešame Gynimo tarybos posėdyje:

Pirmininkas

prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika – 09P).

Nariai:

prof. habil. dr. Romualdas Baušys (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Kazys Kazlauskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09P),

prof. dr. Audris Mockus (Tenesio universitetas, JAV, fiziniai mokslai, informatika – 09P),

doc. dr. Povilas Treigys (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07T).

Disertacija bus ginama viešame Gynimo tarybos posėdyje 2017 m. rugsėjo 28 d. 14 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-04812 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2017 m. rugpjūčio 27 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: *www.vu.lt/lt/naujienos/ivykiu-kalendorius*.

1. Introduction

1.1. Relevance of the problem

During this decade, the voice interface has become more and more popular, the essence of which the interaction is between human and computer based on the verbal form. The idea of such an interface is based on the assertion that verbal communication is the most natural way of human communication that can increase the efficiency of interaction with a computer.

Emotions are an integral part of verbal communication. Emotion, like other non-verbal data (for example, facial expression, posture etc.), conveys some of the information on the basis of which we shape our reaction and respond to the message we receive. Thus, analysis of non-verbal information will only increase the effectiveness of the voice interface. For this purpose, there are ongoing researches on emotional identification of speech, hoping to create reliable, emotional recognition methods that are resistant to various factors, which will allow the human-computer interface to be more natural and informative. On the other hand, the analysis of speech emotions could be successfully applied in criminology, call centers, robot creation and other areas.

1.2. The object of research

The object of the dissertation research is the selection of features in the speech signal and the classification of emotional speech in order to recognize the emotional state of the spoken person.

1.3. The aim and tasks

The main objective of the work is to examine the task of classification emotions of speech and to offer solutions that increase the classification accuracy and reduce the set of necessary features.

To achieve the stated objective, the following tasks are as follows:

1. To propose a hierarchical classification scheme of emotions. This would increase the accuracy of the classification compared to the flat scheme (in which all emotions are classified in one step).

2. To formulate and apply feature selection methods for the hierarchical classification scheme, which allow increasing classification accuracy and reducing the set of necessary features.
3. To perform the experimental research of the proposed hierarchical classification scheme of speech emotions, to evaluate the obtained accuracy of emotion classification, the influence of feature selection methods on classification.

1.4. Scientific novelty

The dissertation examines the task of recognizing emotional speech. A hierarchical speech emotion classification scheme is proposed completely independent of the psychological, social and other factors that characterize emotions. It allows classifying emotional speech records efficiently by using only the acoustic features of a speech signal. The proposed hierarchical classification scheme was adapted to classify the emotional Lithuanian speech records of extremely large volume (5000 records).

1.5. Research methods

For the objective of the work to be achieved and the tasks to be solved, literature review, theoretical analysis and experimental exploratory research were performed. Algorithm theories, data mining, statistical analysis, recognition theory, digital signal processing knowledge were used in the work.

1.6. The defended statements

- The hierarchical speech emotion classification scheme is essentially more effective than flat (one stage) classification from the point of view of classification accuracy.
- Application of feature selection allows for a significant reduction in the amount of data in question, while at the same time increasing the accuracy of the classification compared to the full set of features.
- Different methods of feature selecting in the hierarchical classification scheme do not have a significant effect on the effectiveness of the whole scheme.

- As the number of emotions in question increases, the average accuracy of emotion classification decreases, and the number of features required for maximizing the accuracy of the classification increases.

1.7. Approbation and publications of the research

The main results of the dissertation were published in two peer-reviewed periodical publications, in four conference proceedings publications, and in two conference proceedings abstracts. The main results have been presented and discussed at 2 international conferences.

1.8. The scope of the scientific work

The dissertation is written in Lithuanian and consists of 4 sections. The volume of work is 100 pages and 37 figures and 7 tables are used in the text. The list of 90 references is listed in the dissertation.

2. The task of speech emotion recognition

The task of recognizing of emotional speech is not completely solved – the recognition accuracy of emotions is not yet very high, therefore there are no proposed systems of features and methods of recognition that would be unique and distinguished by their efficiency. Nevertheless, speech emotion recognition is a relevant task and has a great potential for application: identification of emotional state can be used in criminology, call centers (in assessing the emotional state of the caller), the creation of robots (in response to different emotional states) and in other fields, creating preconditions for more efficient human-computer interaction.

The task of speech emotion recognition is the classic recognition task with its content. Essentially, the process of recognizing emotions consists of 3 main stages: speech signal analysis, classifier training and the classification stage.

It is very important to select appropriate features in the classification of speech emotions. Prosodic and spectral features are the most popular and most often used for the speech emotion recognition. Also, features such as epoch parameters, voice quality characteristics, number of harmonics, Zipf features are also used [1]–[5].

The amount of distinctive features is uncertain and sometimes reaches several thousand features. Such a large set of features is bound to be reduced for two reasons. Firstly, for a well-trained classifier, a large set of features requires enormous amount of data for training (emotional speech records). Secondly, large-scale sets of features mean long training and classification process. The methods of feature selection are used for the reduction of the set of features (Sequential Forward Selection, Sequential Backward Selection, Promising First Selection, various genetic algorithms and others) and transformation of features (principal component analysis, linear discriminant analysis, multidimensional scaling, Lipschitz spacing method, Fisher discriminant analysis, processing with neuron networks, decision trees) [2], [3], [6], [7].

A classification of speech emotions is performed after formulation of the set of features. The following three types of classification are proposed in the literature: flat, local, global. Flat classification is attributed to one stage classification. Local and global classification is attributed to a hierarchical classification. Hierarchical classification takes place in several stages, distinguishing class or abstracted group in each stage. Such a classification scheme implies the use of several classifications and even the application of different feature subsets for each classifier. A tree structure or directional acyclic graph can be used to represent the hierarchy.

Various hierarchical speech emotion classification schemes are proposed as an alternative to one stage (flat) classification to improve the accuracy of emotion recognition:

1. Gender of the speaker based two-stage hierarchical classification was applied to identify two emotions (anger and neutral) in [8]. During the first stage, all utterances are classified into three groups: male or neutral, female or anger, and unknown group. During the second stage, the unknown group is classified into two more groups: anger or male and neutral state or female.
2. Gender information was also used in six emotion task [1]. Here female and male utterances were classified separately using different features.
3. Gender information with conjunction with arousal model of the emotion is used in [3]. All six emotions are arranged into three groups by arousal dimension:

active, median, and passive. Each of these groups is classified into two specific emotion classes using different classifiers.

4. Another dimensional emotion model was applied for hierarchical organization in [2]. During the first step, all emotions are classified into high and low activation classes. Furthermore, each class was divided into low and high potency. During last step each class utterances were classified into separate emotions.
5. A psychologically-inspired binary cascade classification scheme using dimensional descriptions of the emotions was presented in [7].
6. Arrangement of 5 emotions into two groups was presented in [5]. This hierarchical classification scheme was organized in hierarchical tree manner by splitting five-class task into six binary classification tasks.
7. Three-level pairwise classification scheme for six emotions was implemented five classifiers in [6]. Emotion pairs were composed in accordance with the highest Fisher rate of all features.
8. Two-stage hierarchical classification schemes were proposed in [4]. All emotions are grouped into two groups during first stage. Each emotional group is classified into separate emotions during second stage. Different feature vectors were used for each group.

To summarize, the hierarchical organization of classification of emotions can achieve an average accuracy of 50 % to 88 %. It is superior to the classification of one stage (flat) and results in a better recognition of emotions. The accuracy of speech emotion recognition is also enhanced by the use of additional information such as differences between male and female voices and psychological aspects of emotions.

3. Hierarchical speech emotions classification scheme

This work presents the hierarchical classification scheme of emotions of speech based on the classification of the tree principle. Such a classification organization allows the use of different sets of features at each stage of the classification, i.e. individually selected set of features for each emotion or group.

There are three basic assumptions of the hierarchical classification of emotional speech displayed below:

1. In the process of classifying emotions in one step, a problem such as the **overlapping** of the features of the emotions in question (acoustic, prosodic and other indications) occurs. This classification problem can be simplified by reducing the amount of emotions analyzed simultaneously. This can be done by classifying the emotions into stages, analyzing a limited amount of emotions in each of them.
2. Each emotion is characterized only by its characteristic acoustic and prosodic features. These features for different emotions may be different or identical. Therefore, increasing **the average of recognition of emotions will not necessarily increase** the individual classification accuracy of each emotion. The solution to this problem is to analyze each emotion or group of emotions, characterized by the same characteristics, separately. Then, the maximum accuracy of classification of such group would guarantee the accuracy of the classification of each that group individually.
3. Emotions can be divided into different classes of different abstractions depending on the selected features. The latter, according to other features, can be subdivided into lower-level classes and so on, up to individual emotion classes. For example, when classifying emotions according to the pitch frequency, low-tone and high-tone classes can be created. Sadness, boredom, neutral state can be attributed to the first one. Joy, anger would be attributable to the high-tone emotions. Each of these classes can be divided into separate emotions by using time, energy and other features for classification. A further distribution to the classes would be possible by prosodic features (for example, rising pitch, constant pitch, and falling pitch).

Thus, in view of the assumptions raised, it is possible to:

- Organize recognition of speech emotions on several levels;
- Use a different set of features for each emotion classification.

The predictable **advantage** is the ability for each level of classification to form the most effective set of features for this level.

3.1. Generalized hierarchical classification scheme

Taking into account the above assumptions for the classification of emotions in speech, the hierarchical classification scheme for emotional speech is proposed.

The basic idea of the hierarchical classification scheme is the classification of emotional speech records in several stages using different set of features in each stage (Figure 1). In the first stage, all emotional speech records are divided into N classes $\{C_1^{(1)}, \dots, C_{\mathrm{N}}^{(1)}\}$, which are determined by the set of first stage features $F_1^{(1)}$. This set of features is constructed in such a way that classification in the classes $\{C_1^{(1)}, \dots, C_{\mathrm{N}}^{(1)}\}$ would give the maximum accuracy. In the second stage, each class of emotions is divided into lower-level classes $\{C_1^{(2)}, \dots, C_{\mathrm{K}}^{(2)}\}$ or in separate individual emotions using sets of second stage features $F_k^{(2)}$, $k = 1, \dots, K$. Using this classification scheme can be any number of classification steps L with different sets of features for each class. The number of emotion classes in each stage is also unlimited. This can guarantee a proper set of features for each level of classification, i.e. for each emotion.

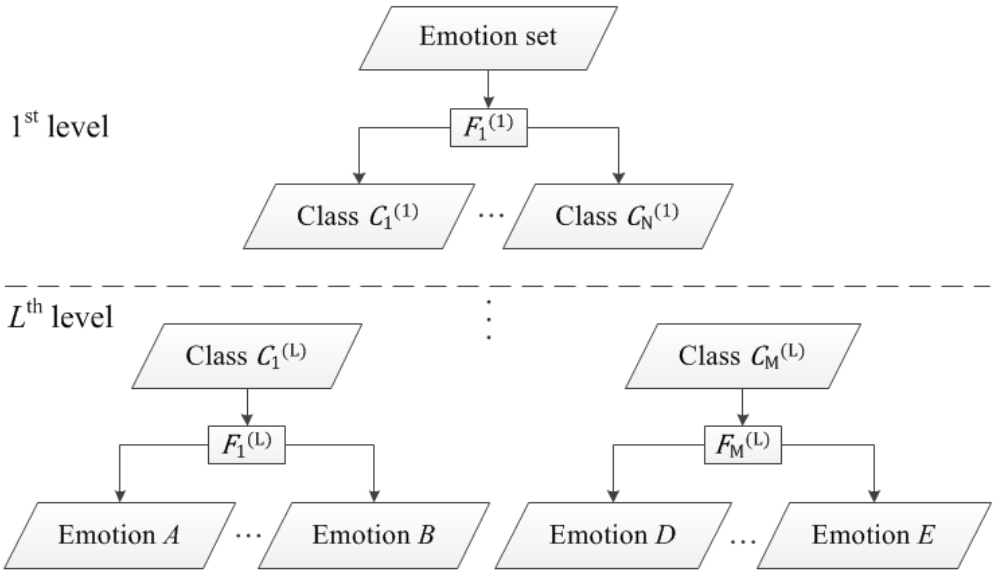

Figure 1: The generalized hierarchical classification scheme for speech emotion recognition

For example, when examining energy as features, in the first classification stage, classification will be made into low and high energy classes (examples of low energy emotions – boredom, neutral, high energy – joy, anger). In the second stage of classification, low and high energy classes would be classified into lower-level classes or individual emotions that would be defined by the features used (completely different from those used during the first stage classification).

Each partial set of features $F_m^{(l)}$ is used to classify a particular group of emotions or emotions by organizing the classification of emotions on several levels. The main principles of the proposed hierarchical classification scheme are as follows:

- Classification is organized on separate levels. In order to recognize emotions or class of emotion, an individual and most effective feature or set of features $F_m^{(l)}$ is used at each stage of the classification. Such sets of features can be made using the above-mentioned sequential or other feature selection techniques.
- The entire hierarchical classification process is described as a general combination of all classes and emotions.

$$F = \{F_m^{(l)}\}, \qquad m = 1, \dots, M;\ l = 1, \dots L. \qquad (1)$$

Here M is the number of emotion classes in a specific level of classification, L is the number of classification levels.

Generally speaking, a set of emotions (or a set of classes derived from higher-level classes) can be classified into any number of classes. The number of classes analyzed (one level) and the number of classification levels is defined by the total number of emotions analyzed. The simplest case for classifying hierarchical emotions is the classification into two classes of emotion or emotions.

The sets of features may be heterogeneous – they consist of different features (time, spectrum, energy, voice quality, etc.) using a variety of methods and criteria for the selection of features.

The main advantage of the proposed hierarchical classification of speech emotions method is that the processes for classifying different levels are independent of a common set of features. The classification process can be optimized by improving the accuracy of

the selected emotion class recognition without affecting the accuracy of other emotion classes.

3.2. Feature selection methods

Three feature selection techniques were implemented in hierarchical classification scheme: Maximal Efficiency criterion (ME), the criterion of the Minimal Cross-Correlation of features (MC), and the Sequential Forward Selection (SFS) based technique.

Maximal Efficiency Feature Selection Criterion

This criterion is proposed by making assumption on aggregate efficiency of features with maximal individual efficiency i.e. of features giving the lowest classification error. The formation of feature subset using ME selection criterion is performed.

$$f_m^{(l)} = \arg\min_j E(f_j^{(l)}),\ j = 1,...,J. \qquad (2)$$

Here $E(f_j^{(l)})$ is a classification error of the j-th feature in the l-th level $f_j^{(l)}$. J is a total number of features in the l-th classification level.

The feature subset is initialized once and extended with the most effective features $f_j^{(l)}$ repeatedly. The evaluation of every subset case is performed and the extension process is stopped when all J features-candidates are considered.

Minimal Cross-Correlation Criterion

In this case assumption on the efficiency of linearly independent features is made. Independent features make the set more effective than strongly correlated ones. Thus by selecting linearly independent features one seeks for more effective subset.

Minimal Cross-Correlation (MC) criterion based feature selection is initiated with the most efficient feature thus ensuring the discriminative power of the subset. The analyzed feature subset is expanded by adding features with the minimal cross-correlation value.

$$f_m^{(l)} = \arg\min_j |R(f_0^{(l)}, f_j^{(l)})|,\ j = 1,...,J. \qquad (3)$$

Here $f_0^{(l)}$ is the feature with highest classification accuracy for analyzed emotion group. $R(f_0^{(l)}, f_j^{(l)})$ is the cross-correlation of the $f_0^{(l)}$ and the new feature $f_j^{(l)}$.

Again, the expansion of the feature subset $\{F_m^l\}$ is stopped when J features having the least correlation with the most effective one are considered.

Sequential Forward Selection

Sequential Forward Selection technique (SFS) is one of the greedy search algorithms aiming to find the most significant subset of the features. Here the aggregate efficiency of the feature subset is considered rather than individual properties of the feature.

SFS of features starts from initialization of the empty feature subset F_0. The subset is extended with a feature $f_j^{(l)}$ making the new subset F_{i+1} more effective

$$f_m^{(l)} = \arg\max_j \left[E(F_i + f_j^{(l)}) - E(F_i) \right], j = 1, 2, ..., J. \qquad (4)$$

The feature set extension step is repeated until the efficiency of newly obtained feature set F_{i+1} increases or while $i < J$, J – total number of features.

4. Experimental study

Experimental research is taken in three stages: comparison of feature selection criteria, evaluation of efficiency of hierarchical emotion classification scheme, and comparison of alternative hierarchical schemes.

Recognition tasks of 3 emotions (anger, joy, neutral), 4 emotions (anger, joy, neutral, sadness), and 5 emotions (anger, joy, neutral, sadness, and fear) were chosen. 1000 examples (for Lithuanian spoken language emotion database [10]) and 60 examples (for Berlin emotional speech database [9]) of each emotion were analyzed during experiment.

The results of the thesis are averaged results of the 3-fold testing for Berlin emotional speech database and 10-fold testing for Lithuanian spoken language emotion database.

Non-parametric K-Nearest Neighbor classifier was selected for experimental testing. The value $K = 5$ (for Berlin emotional speech database) and $K = 7$ (for Lithuanian spoken language emotion database) were selected considering different size of data sets.

The initial full features set consisted of 6552 different speech emotion features. The features for the experiment were extracted using *OpenEAR* toolkit [11].

Comparison of feature selection criteria

Emotion classification accuracy has been compared measuring maximal efficiency, minimal correlation and using sequential feature selection methods. To evaluate accuracy, herewith, the results of classification are provided using full feature sets.

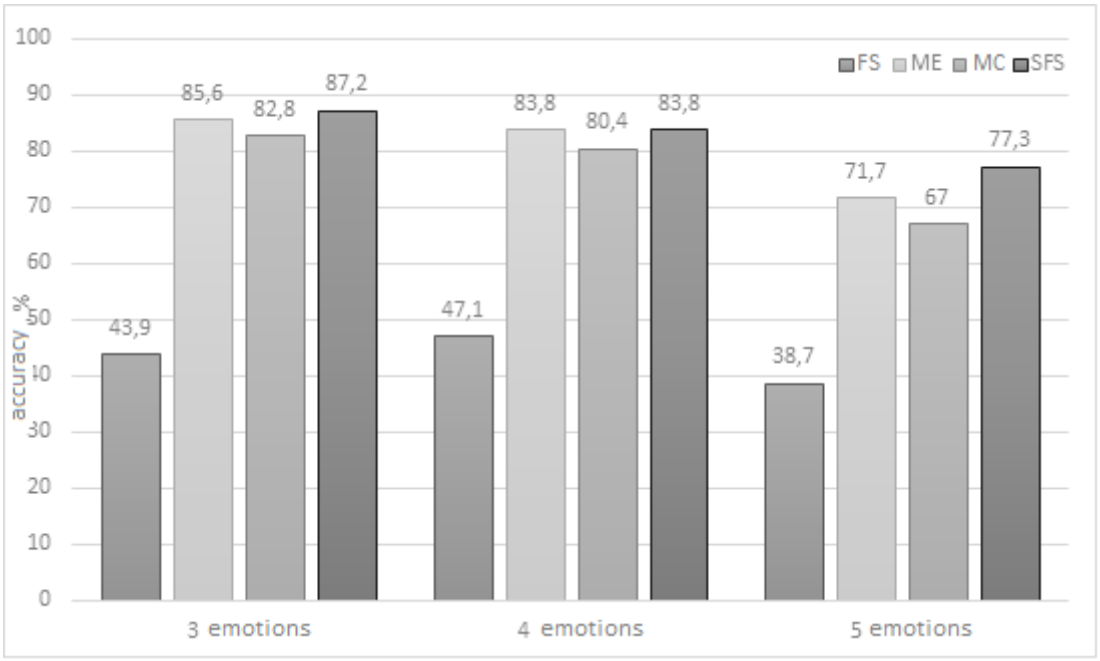

Figure 2: German emotion classification using full set (FS), Maximal efficiency (ME), minimal correlation (MC), Sequential forward selection (SFS).

As can be seen in Figure 2, using full feature set the minimal accuracy is under 50 %. Maximal accuracy is reached using sequential forward selection (SFS) method – the average accuracy in case of five emotions is 77.3 %, in case of three emotions – 87.2 %.

Looking at the results of research on Lithuanian spoken language emotions, it can be stated that there is not much of a difference between the German spoken language emotions (Figure 3).

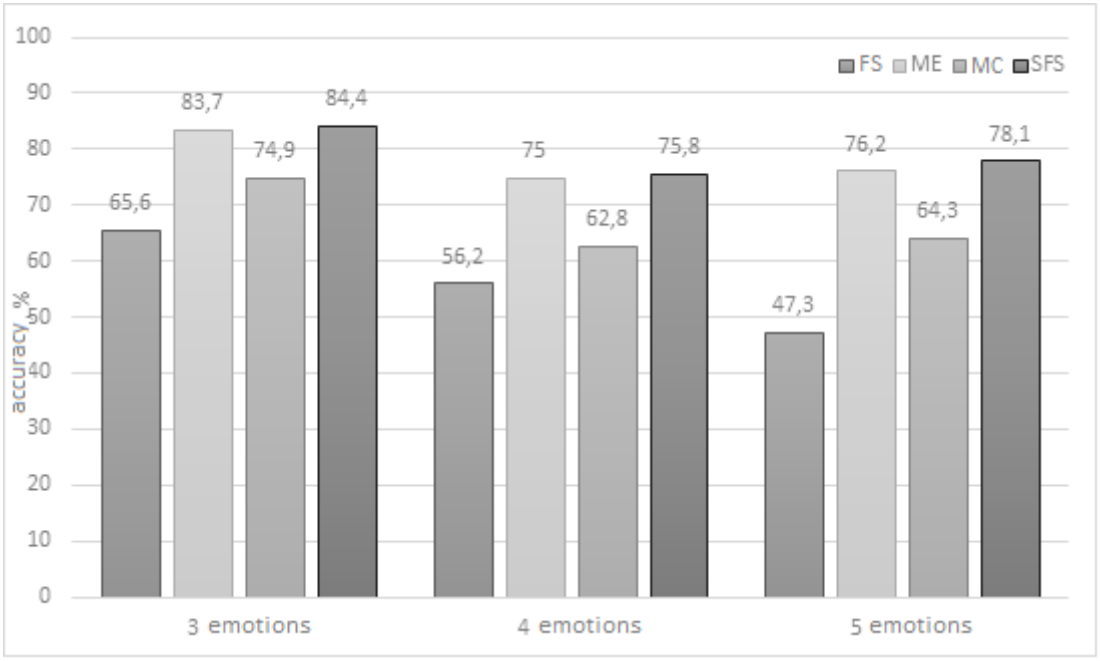

Figure 3: Lithuanian emotion classification using full set (FS), Maximal efficiency (ME), minimal correlation (MC), Sequential forward selection (SFS).

Sequential forward selection method is equally efficient in Lithuanian and in German cases.

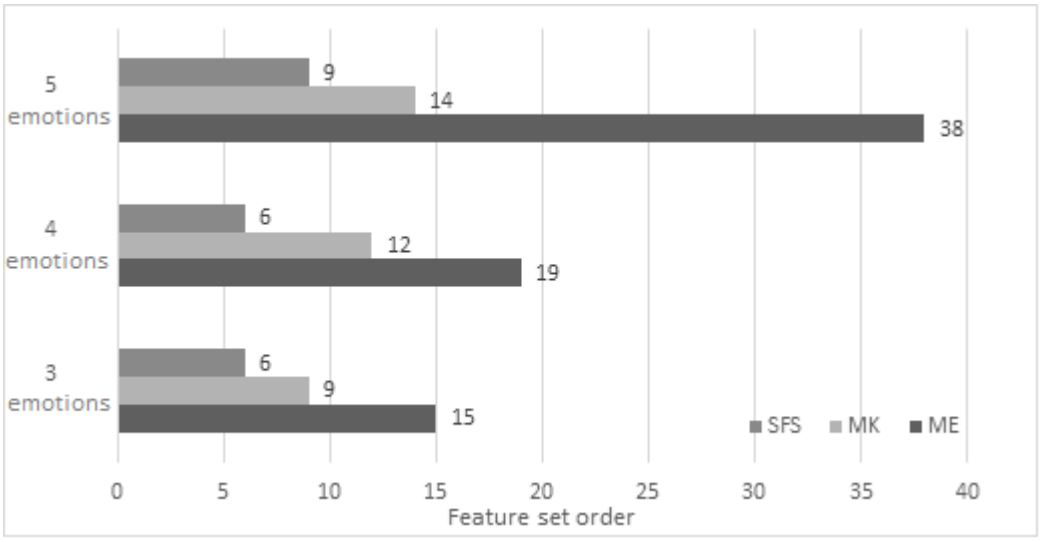

Figure 4: Feature set order using different selection methods for German emotion classification.

Figure 4 and Figure 5 listed feature set order combined using different selection methods.

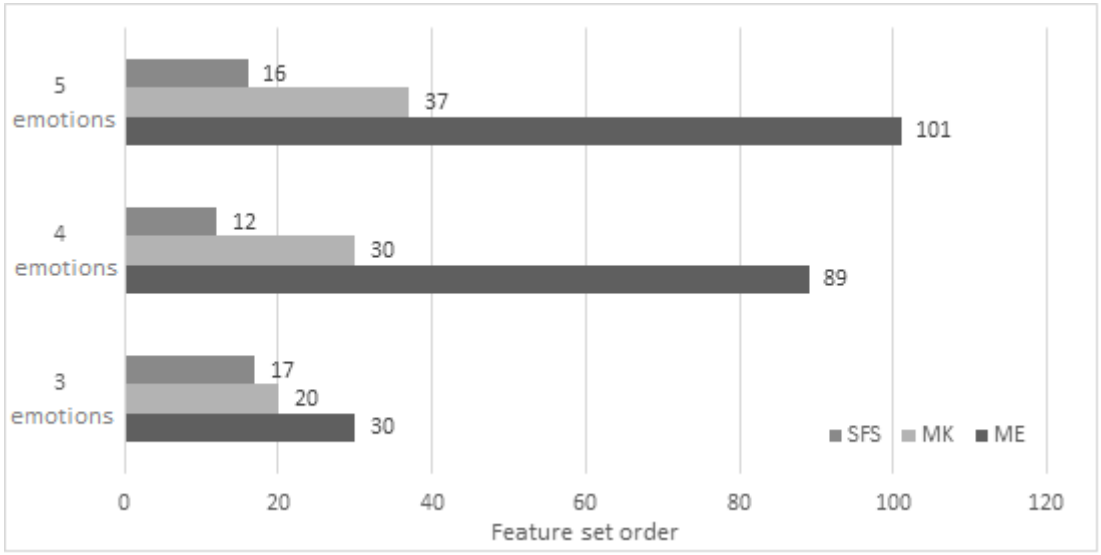

Figure 5: Feature set order using different selection methods for Lithuanian emotion classification.

In German and in Lithuanian language cases the minimal feature order is received using SFS method: 6 (3 emotions German), 17 (3 emotions Lithuanian). Maximal feature order received using ME: 15 (3 emotions), 101 (5 emotions Lithuanian). Having in mind, that number of German utterances was 5 times less comparing with Lithuanian, it can be stated that with increment of samples the feature set order is also increasing.

Experimental research on comparison of feature selection criteria shows that maximal efficiency, minimal cross-correlation, sequential forward selection provides similar results. No obvious advantage neither of methods can be stated after the research at the viewpoint of classification accuracy. Nevertheless, the Sequential forward selection technique enabled us to obtain up to 4 times smaller feature sets in comparison with other two techniques.

Hierarchical emotion classification

In this stage, efficiency of hierarchical classification scheme will be evaluated and compared with flat classification scheme.

A two-stage classification scheme was designed assuming low-pitch and high-pitch emotion classes in the first classification stage.

In case of **3 emotions** (joy, anger, and neutral state) task two-level classification will be held. The low-pitch class should contain neutral state patterns only so the second level will contain only one classification process using feature set $F_2^{(2)}$.

No obvious differences between selection methods were noticed – they all provided similar classification results. Comparing with flat scheme results, the advantage of hierarchical classification is obvious – 15–40 % higher accuracy (Table 1).

Table 1: Classification results of three emotions

Base	Criteria	Classification accuracy, %			
		Anger	Joy	Neutral	Average of three emotions
Berlin DB	ME	81.7	81.7	96.7	86.7
	MC	85	81.7	96.7	87.8
	SFS	86.7	75	96.7	86.1
LT DB	ME	56.6	67.3	65.5	63.1
	MC	54	62.6	70.3	62.3
	SFS	57.6	71.7	66	65.1

The most accurately classified emotion in both (German and Lithuanian) cases is neutral. The most number of errors received classifying joy in German case and anger in Lithuanian case. The possible reason might be different acting.

Table 2: Confusion matrix for Lithuanian emotions classification

Emotions	Classified, %		
	Anger	Joy	Neutral
Anger	57.6	32.1	10.3
Joy	15.1	71.7	13.2
Neutral	18.3	15.7	66

As one can see in Table 2, the most frequent classification error – anger classified as joy. These emotions are acoustically similar and that is the reason of false classification.

There were no signs of feature overlap between the separate classification levels, different feature subsets were formed at all levels. This demonstrates a well-defined hierarchical classification task by dividing emotions into low-pitch and high-pitch groups.

In case of **4 emotions** (joy, anger, sadness, and neutral state) two-level classification will occur. The low-pitch class should contain sadness and neutral state patterns and the high-pitch class should contain joy and anger patterns.

Again, all feature selection methods provided similar results (Table 3). In German case, results remained similar to results of classifying three emotions. While, in Lithuanian case, classification accuracy reduced and stayed at approx. 55 %, but still remained higher than in flat classification.

Table 3: Classification results of four emotions

Base	Criteria	Classification accuracy, %				
		Anger	Joy	Neutral	Sadness	Average of four emotions
Berlin DB	ME	80	76.7	91.7	96.7	86.3
	MC	85	81.7	93.3	95	88.8
	SFS	88.3	73.3	98.3	98.3	89.6
LT DB	ME	53.5	62.2	50.1	56.8	55.7
	MC	49.6	58.3	46.4	63.8	54.5
	SFS	52	67.6	48	57.8	56.4

One can see that in German case the most accuracy classified emotion is sadness, in Lithuanian case – joy and sadness.

Table 4: Confusion matrix for Lithuanian emotions classification

	Classified, %			
	Anger	Joy	Neutral	Sadness
Anger	52	36	6.2	5.8
Joy	19.1	67.5	11.1	2.2
Neutral	15.6	10.4	48	26
Sadness	3.7	7.1	31.4	57.8

Table 4 shows the most frequent classification error is classifying joy, in German case, and neutral emotion, in Lithuanian case.

No feature overlap between the separate classification levels was also obtained.

In case of **5 emotions** (joy, anger, sadness, fear and neutral state) two-level classification will occur. The low-pitch class should contain sadness and neutral state patterns and the high-pitch class should contain joy, anger and fear patterns.

In this case, SFS method provided the best results – 3–8 % better than the other two methods (Table 5). Accuracy is lower by 10 %, compared with four emotions. Therefore, it can be stated that classification complexity increases the accuracy reduces. The increasing complexity of classification task, pointed out the difference between German and Lithuanian databases – the German database is more resistant to increasing number of emotions.

It can be seen that in German case the most accuracy classified emotion is neutral and sadness, in Lithuanian case – sadness (Table 5).

Table 5: Classification results of five emotions

Base	Criteria	Classification accuracy, %					
		Anger	Joy	Neutral	Sadness	Fear	Average of five emotions
Berlin DB	ME	80	55	91.7	93.3	63.3	76.7
	MC	85	58.3	85	75	71.7	75
	SFS	81.7	63.3	90	90	73.3	79.7
LT DB	ME	42.7	48.1	45.1	49.3	45.2	46.1
	MC	42.4	38.1	40.3	52.8	34.6	41.6
	SFS	49.4	60.6	41.2	50.3	46.7	49.6

The highest overlapping is between joy and anger, neutral and sadness. The highest identification is between anger and sadness (Table 6).

Table 6: Confusion matrix for Lithuanian emotions classification

	Classified, %				
	Anger	Joy	Neutral	Sadness	Fear
Anger	49.4	29.6	12.4	6.9	1.7
Joy	16.4	60.6	8.5	6.1	8.4
Neutral	7.2	10.9	41.2	33.5	7.2
Sadness	1.1	5.5	35.7	50.3	7.4
Fear	4.7	5.1	33.4	10.1	48.7

Figure 6 and Figure 7 illustrate the dependency between emotion classification accuracy and number of emotions. Two tendencies can be noted:

- Increasing number of emotions reduce classification accuracy.

- The number of samples used in classification has impact on accuracy – the higher number of samples the lower accuracy of classification because of overlapping features.

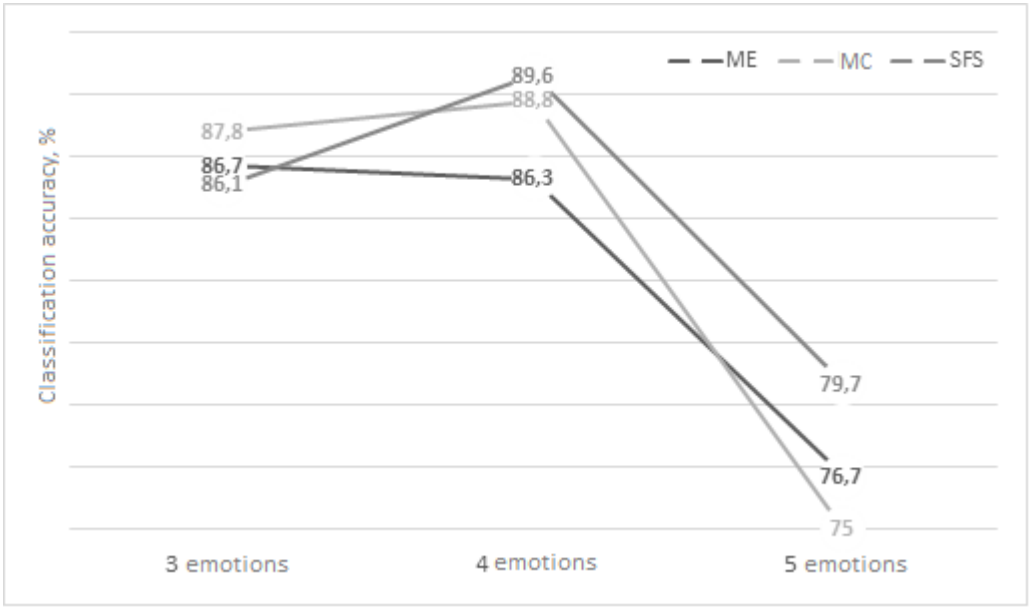

Figure 6: German emotion classification dependency on number of emotions using full set (FS), Maximal efficiency (ME), minimal correlation (MC), Sequential forward selection (SFS)

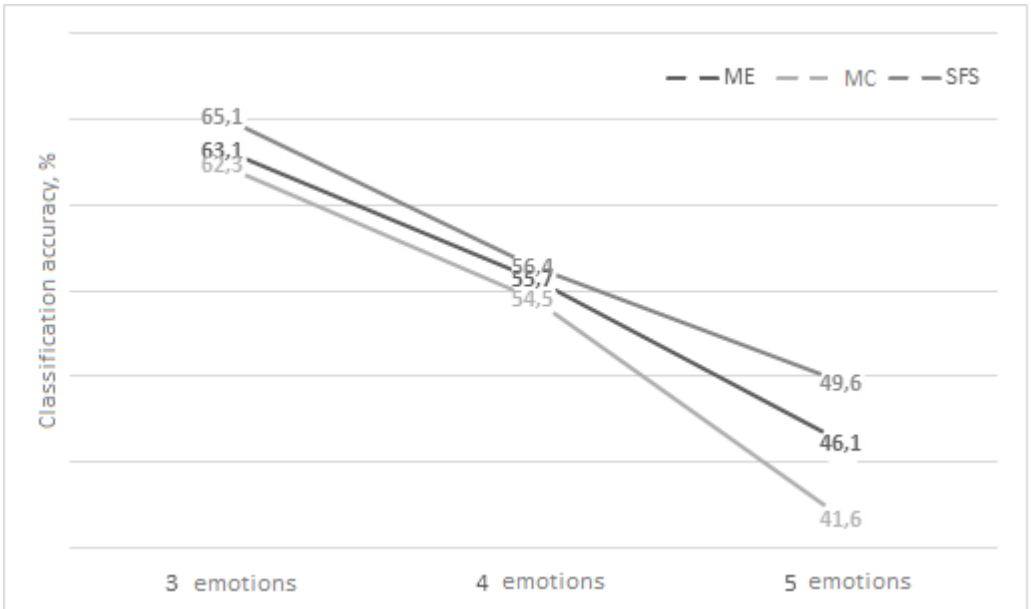

Figure 7: Lithuanian emotion classification dependency on number of emotions using full set (FS), Maximal efficiency (ME), minimal correlation (MC), Sequential forward selection (SFS)

A more complex classification task has led to a partial feature overlapping of different levels. In the case ME selection based analysis of Lithuanian language

emotions, 2 (out of 59) first-level features were selected on the next levels. This corresponds to 3.4 % overlap only, so this fact does not invalidate the above-stated statement of the well-defined levels of the hierarchical classification scheme.

Besides, in case of 5 emotions classification task more complex feature sets were obtained: the sets were larger and contained wider range of various features.

Alternative hierarchical schemes

For comparison, five emotions classification task was used from German database, as other authors also use the same database. For experimental research, work has been done on the schemes:

- Scheme #0. It is previously investigated 5 emotions classification hierarchical scheme.
- Scheme #1. It is previously investigated 5 emotions classification hierarchical scheme modification. The low-pitch class should contain sadness, fear, and neutral state patterns and the high-pitch class should contain joy and anger patterns. We will have two-level classification.
- Scheme #2. This scheme is based on work [7]. Neutral state and joy depends to non-negative valence class. Anger, sadness and fear – to negative valence class. The latter class during second level divided into positive activation (anger, fear) and negative activation (sadness) classes. We got two-level classification.
- Scheme #3. During the first step all 5 emotions are classified into high activation (anger, joy, fear) and low activation (sadness and neutral set) classes [2]. Further first class was divided into high (anger and joy are divided during third step) and low (fear) potency classes. Low activation emotions also are classified into high (neutral state) and low (sadness) emotions.

All schemes differ on emotions grouping. Comparing the classification results, it will be possible to evaluate proposed scheme resistance to grouping emotions to classes.

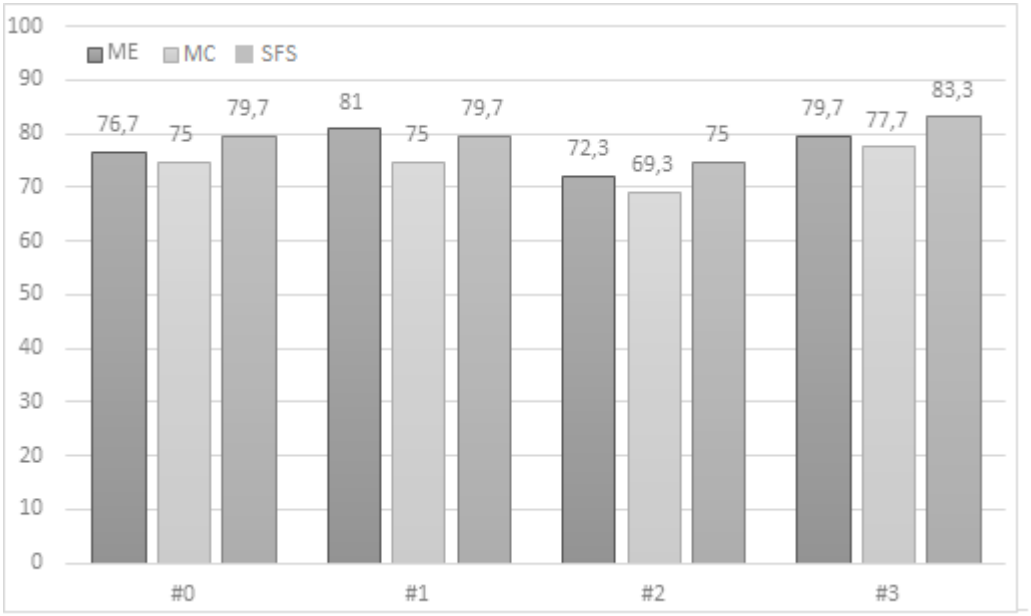

Figure 8: Alternative hierarchical schemes emotions classification accuracy mean values

Figure 8 illustrates classification results that points out that the results provided by different methods are similar – variation of accuracy is in 5–8 % interval.

Figure 9 illustrates the order of feature sets gathered by selection methods and used in alternative schemes. From the given illustration, one can see that the minimal order of feature set is received using SFS method (13–20 features), the maximal order provided by ME method (22–37 features).

By evaluating the dependence of the features on the number of analysed emotions, we have noticed the consistent increase in sets with increasing number of emotions. In the case of SFS based analysis of Germanic emotions, 57 % of the features selected for classification of 3 emotions were also included into the set for classification of 4 emotions. 27 % of the 4 emotions oriented features were included into the set for classification of 5 emotions.

In case of Lithuanian language, 62 % of the features selected for 3 emotions classification task have been included into the set for 4 emotion classification. 36 % of features of the 4 emotions task were incorporated into the set of 5 emotion classification.

Somewhat smaller, but similar results were obtained for other selection techniques also.

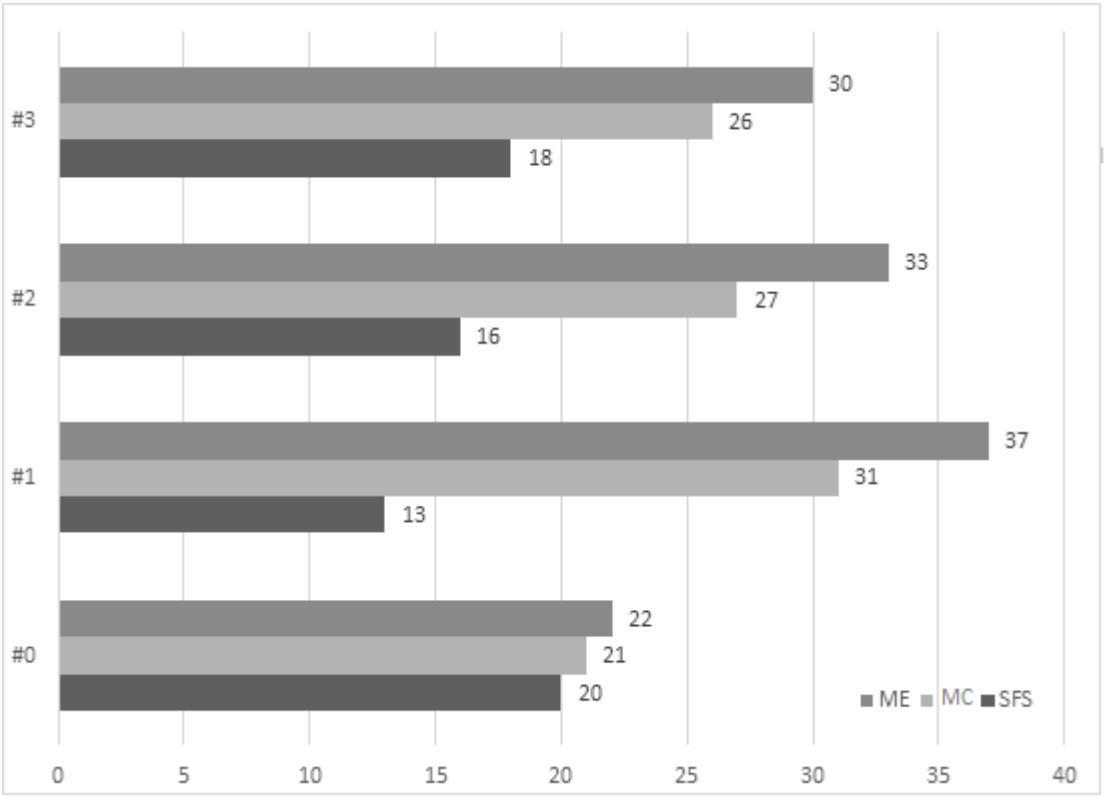

Figure 9: Feature set orders for alternative hierarchical schemes.

Conclusions

After experiments, the following conclusions can be made:

- Hierarchical emotions classification scheme is more efficient, with respect to classification accuracy, than flat (one stage) classification.
    - Experiments prove that proposed hierarchical scheme is more efficient against flat classification by 40 %.
    - The feature overlap between the different hierarchical classification levels varied from 0 % to 3.4 %.
- Feature selection methods enable greatly to reduce the amount of data used in classification process and, at the same time, to improve classification accuracy comparing with classification using the full sets.
    - The proposed and applied selection methods allowed reducing the feature sets order from 6552 features to 9–38 features in flat case, and 7–110 features in hierarchical scheme case. The most efficient selection method is sequential forward selection (SFS) method providing up to four times shorter feature sets.

- Different feature selection methods do not have great impact on scheme efficiency classifying emotions.
    - Experiments shows that tested selection methods give 1.5–5 % accuracy difference in German case, and 2–8 % in Lithuanian case.
- Classification accuracy is reducing with increasing number of emotions.
    - With increment of emotions from 3 to 5 the order of feature sets increased by 1.5–2 times. With increment of number of emotion samples by 5 times, feature set order increased by 2–6 times.
    - With the expansion of emotion set from 3 to 4, the feature set overlap went up to 62 %. Expansion up to 5 emotion set, the highest overlap was 36 %.

List of publications

The articles published in the peer-reviewed periodical publications:

- Liogienė T., Tamulevičius G., 2016. Multi-stage Recognition of Speech Emotion Using Sequential Forward Feature Selection. *Journal on Electrical, Control and Communication Engineering. Vol.10:* 35–41, ISSN 2255-9140, e-ISSN 2255-9159.
- Tamulevičius G., Liogienė T., 2015. Low-order Multi-level Features for Speech Emotion Recognition. *Baltic Journal of Modern Computing Vol. 3, No. 4*: 234–347. ISSN 2255-8950.

The articles published in the conference proceedings:

- Liogienė T., Tamulevičius G., 2016. Comparative Study of Multi-stage Classification Scheme for Recognition of Lithuanian Speech Emotions. *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems. ACSIS, Vol. 8,:* 483–486, ISSN 2300-5963.
- Liogienė T., Tamulevičius G., 2015. SFS Feature Selection Technique for Multistage Emotion Recognition. *Proceeding of the 2015 IEEE 3th workshop on Advances in Information, Electronic and Electrical Engineering*: 1–4, ISBN 978-1-5090-1201-5.
- Liogienė T., Tamulevičius G., 2015. Minimal Cross-correlation Criterion for Speech Emotion Multi-level Feature Selection. *Proceedings of the Open Conference of Electrical, Electronic and Information Sciences (eSTREAM). Washington, IEEE:* 1–4, ISBN 978-1-4673-7445-3.
- Liogienė T., 2014. Dinaminiai pagrindinio tono dažnio požymiai kalbos emocijoms atpažinti. Informacinės technologijos: *19-oji tarpuniversitetinė magistrantų ir doktorantų konferencija "Informacinė visuomenė ir universitetinės studijos"*: 161–166. ISSN 2029-4832.

Abstracts in the conference proceedings:

- Liogienė T., Tamulevičius G. Multistage Speech Emotion Recognition for Lithuanian: experimental study. *Data Analysis Methods for Software*

Systems: 7th International Workshop: [abstracts book], Druskininkai, Lithuania, ISBN 9789986680581, 3–5 of December, 2015, p. 34.

- Liogienė T., Tamulevičius G. Low-order Multi-level Features for Speech Emotions Recognition. *Data Analysis Methods for Software Systems: 6th International Workshop*: [abstracts book], Druskininkai, Lithuania, ISBN 978-9986-680-50-5, 4–6 of December, 2014, p. 35.

About the author

Tatjana Liogienė received the B.Sc and M.Sc. degree in informatics from Lithuanian University of Educational Sciences in 2003 and 2005, respectively. Since 2005, she is a Lecturer at the University of Applied Sciences. From 2012 till 2016 she was a Doctoral student of Vilnius University, Institute of Mathematics and Informatics.

References

[1] J. Liu, C. Chen, J. Bu, M. You, and J. Tao, "Speech Emotion Recognition using an Enhanced Co-Training Algorithm," *2007 IEEE International Conference on Multimedia and Expo,* pp. 999–1002, July 2007.

[2] M. Lugger, M.-E. Janoir, and B. Yang, "Combining classifiers with diverse feature sets for robust speaker independent emotion recognition," *17th European Signal Processing Conference,* pp. 1225–1229, 2009.

[3] Z. Xiao, E. Centrale, L. Chen, and W. Dou, "Recognition of emotions in speech by a hierarchical approach," *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops,* pp. 1–8, September 2009.

[4] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language,* pp. 556–570, 2011.

[5] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach," *Speech Communication,* pp. 1162–1171, 2011.

[6] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Processing,* pp. 1154–1160, 2012.

[7] M. Kotti and F. Paterno, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *Iternational Journal of Speech Technology,* pp. 131–150, 2012.

[8] W.-J. Yoon and K.-S. Park, "Building robust emotion recognition system on heterogeneous speech databases," *2011 IEEE International Conference on Consumer Electronics,* pp. 825–826, 2011.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *In: Proceedings of Interspeech,* pp. 1517–1520, 2005.

[10] J. Matuzas, T. Tišina, G. Drabavičius, and L. Markevičiūtė, "Lithuanian Spoken Language Emotions Database," Baltic Institute of Advanced Language, 2015. [Online]. Available: http://datasets.bpti.lt/lithuanian-spoken-language-emotions-database/.

[11] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR – Introducing the Munich open-source emotion and affect recognition toolkit," *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops,* pp. 1–6, September 2009.

HIERARCHINIS ŠNEKOS EMOCIJŲ KLASIFIKAVIMAS

Tyrimo sritis ir problemos aktualumas

Šiame dešimtmetyje itin pradėjo populiarėti balso sąsaja, kurios esmė – verbaline forma grįsta žmogaus ir kompiuterio sąveika. Tokios sąsajos idėja remiasi teiginiu, jog verbalinis bendravimas yra pats natūraliausias žmogaus bendravimo būdas, kuris gali padidinti sąveikos su kompiuteriu efektyvumą.

Neatsiejama verbalinio bendravimo dalis yra emocijos. Emocija, kaip ir kita neverbalinė raiška (pavyzdžiui, veido išraiška, laikysena), perduoda dalį informacijos, kuria remdamiesi mes formuojame savo reakciją ir atsaką į gaunamą žinią. Taigi neverbalinės informacijos analizė tik padidins balso sąsajos efektyvumą. Tuo tikslu ir yra vykdomi šnekos emocijų atpažinimo tyrimai, tikintis sukurti patikimus, įvairiems veiksniams atsparius šnekos emocijų atpažinimo metodus, kurie leis žmogaus ir kompiuterio sąsajai suteikti daugiau natūralumo, informatyvumo. Kita vertus, šnekos emocijų analizė galėtų būti sėkmingai pritaikoma kriminalistikoje, skambučių centruose, kuriant robotus ir kitose srityse.

Tyrimų objektas

Disertacijos tyrimų objektas – emocijų požymių šnekos signale atranka bei emocionalios šnekos klasifikavimas, siekiant atpažinti šnekančiojo emocinę būklę.

Darbo tikslas ir uždaviniai

Pagrindinis darbo tikslas yra išnagrinėti šnekos emocijų klasifikavimo uždavinį ir pasiūlyti sprendimus, leidžiančius padidinti klasifikavimo tikslumą bei sumažinti reikalingų požymių aibę.

Siekiant suformuluoto tikslo, buvo sprendžiami šie uždaviniai:

1. pasiūlyti hierarchinę emocijų klasifikavimo schemą, kuri leistų padidinti klasifikavimo tikslumą lyginant su tiesiogine schema (kurioje visos emocijos klasifikuojamos iš karto vienu žingsniu);
2. hierarchinei klasifikavimo schemai suformuluoti ir pritaikyti požymių atrankos metodus, leidžiančius padidinti klasifikavimo tikslumą bei sumažinti reikalingų požymių aibę;

3. atlikti pasiūlytosios hierarchinės emocijų klasifikavimo schemos eksperimentinį tyrimą, įvertinti gaunamą emocijų klasifikavimo tikslumą, požymių atrankos metodų įtaką klasifikavimui.

Mokslinis darbo naujumas

Disertacijoje yra nagrinėjamas emocionalios šnekos atpažinimo uždavinys. Yra pasiūlyta hierarchinė klasifikavimo schema, kuri yra visiškai nepriklausoma nuo psichologinių, socialinių ir kitų emocijas apibūdinančių veiksnių. Ji leidžia atlikti efektyvų emocionalios šnekos įrašų klasifikavimą naudojant tik akustinius šnekos signalo požymius. Pasiūlytoji hierarchinė klasifikavimo schema buvo pritaikyta itin didelės apimties (5 000 įrašų) emocionalios lietuvių šnekos įrašams klasifikuoti.

Tyrimo metodika

Darbo tikslui pasiekti ir uždaviniams spręsti buvo atliekama literatūros apžvalga, teorinė analizė, įvykdytas eksperimentinis žvalgomojo (angl. *exploratory*) pobūdžio tyrimas. Darbe buvo panaudotos algoritmų teorijos, duomenų gavybos, statistinės analizės, atpažinimo teorijos, skaitmeninio signalo apdorojimo žinios.

Ginamieji teiginiai

- Hierarchinė emocionalios šnekos klasifikavimo schema tikslumo požiūriu yra iš esmės efektyvesnė už tiesioginį (vieno etapo) klasifikavimą.
- Požymių atrankos taikymas leidžia reikšmingai sumažinti nagrinėjamų duomenų kiekį ir kartu padidinti klasifikavimo efektyvumą, palyginti su pilnais požymių rinkiniais.
- Hierarchinėje klasifikavimo schemoje skirtingi požymių atrankos metodai neturi esminės įtakos visos schemos efektyvumui.
- Didėjant nagrinėjamų emocijų skaičiui, vidutinis emocijų klasifikavimo tikslumas mažėja, o klasifikavimo tikslumui maksimizuoti reikalingas požymių kiekis didėja.

Darbo rezultatų aprobavimas

Publikacijos periodiniuose recenzuojamuose leidiniuose:

- Liogienė T., Tamulevičius G., 2016. Multi-stage Recognition of Speech Emotion using Sequential Forward Feature Selection. *Journal on Electrical, Control and Communication Engineering. Vol.10:* 35–41, ISSN 2255-9140, e-ISSN 2255-9159.
- Tamulevičius G., Liogienė T., 2015. Low-order Multi-level Features for Speech Emotion Recognition. *Baltic Journal of Modern Computing Vol. 3, No. 4*: 234–347. ISSN 2255-8950.

Publikacijos recenzuojamuose konferencijų leidiniuose:

- Liogienė T., Tamulevičius G., 2016. Comparative Study of Multi-stage Classification Scheme for Recognition of Lithuanian Speech Emotions. *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems. ACSIS, Vol. 8,:* 483–486, ISSN 2300-5963.
- Liogienė T., Tamulevičius G., 2015. SFS Feature Selection Technique for Multistage Emotion Recognition. *Proceeding of the 2015 IEEE 3th Workshop on Advances in Information, Electronic and Electrical Engineering*: 1–4, ISBN 978-1-5090-1201-5.
- Liogienė T., Tamulevičius G., 2015. Minimal Cross-correlation Criterion for Speech Emotion Multi-level Feature Selection. *Proceedings of the Open Conference of Electrical, Electronic and Information Sciences (eSTREAM). Washington, IEEE:* 1–4, ISBN 978-1-4673-7445-3.
- Liogienė T., 2014. Dinaminiai pagrindinio tono dažnio požymiai kalbos emocijoms atpažinti. Informacinės technologijos: *19-oji tarpuniversitetinė magistrantų ir doktorantų konferencija "Informacinė visuomenė ir universitetinės studijos"*: 161–166. ISSN 2029-4832.

Santraukos konferencijų leidiniuose:

- Liogienė T., Tamulevičius G. Multistage Speech Emotion Recognition for Lithuanian: experimental study. *Data Analysis Methods for Software Systems: 7th International Workshop*: [abstracts book], Druskininkai, Lietuva, ISBN 9789986680581, gruodžio 3–5 d., 2015, p. 34.
- Liogienė T., Tamulevičius G. Low-order Multi-level Features for Speech Emotions Recognition. *Data Analysis Methods for Software Systems: 6th*

International Workshop: [abstracts book], Druskininkai, Lietuva, ISBN 978-9986-680-50-5, gruodžio 4–6 d., 2014, p. 35.

Disertacijos struktūra

Disertaciją sudaro 4 skyriai: Įvadas, Emocijų klasifikavimo uždavinys, Hierarchinė emocijų klasifikavimo schema, Eksperimentinis tyrimas bei bendrosios išvados su literatūros sąrašu. Disertacijos apimtis: 100 puslapių, 7 lentelės, 37 iliustracijos. Disertacijoje remtasi 90 literatūros šaltinių.

Bendrosios išvados

Disertacijoje išnagrinėtas emocijų atpažinimo šnekoje uždavinys, išanalizuotos hierarchinio emocijų klasifikavimo schemos. Analitinėje dalyje emocijoms klasifikuoti pritaikyta sprendimų medžio struktūra ir pasiūlyta hierarchinio klasifikavimo schema.

Pagrindiniai rezultatai:

- Remiantis sprendimų medžių idėja, suformuluota hierarchinio klasifikavimo schema, leidžianti nagrinėti pageidaujamą šnekos emocijų kiekį. Pasiūlytos schemos struktūra priklauso nuo pasirinktojo emocijų grupavimo.
- Požymių rinkiniams sudaryti suformuluoti maksimalaus efektyvumo ir minimalios koreliacijos atrankos metodai, pritaikytas nuoseklaus aibės didinimo metodas. Visi atrankos metodai pritaikyti hierarchinei klasifikavimo schemai.
- Atliktas pasiūlytos klasifikavimo schemos eksperimentinis tyrimas su emocijų vokiečių kalbos įrašais. Atlikti didelės apimties eksperimentai su suvaidintų emocijų lietuvių kalba įrašais.

Atlikus eksperimentinius tyrimus, suformuluotos tokios išvados bei jas patvirtinantys eksperimentiniai rezultatai:

- Hierarchinė emocionalios šnekos klasifikavimo schema yra iš esmės efektyvesnė už tiesioginį (vieno etapo) klasifikavimą tikslumo požiūriu.
    - Eksperimentų metu gautas hierarchinės schemos vidutinis klasifikavimo tikslumas iki 40 proc. didesnis nei tiesioginės vieno etapo klasifikavimo schemos.

- Požymių persidengimas tarp atskirų hierarchinių lygmenų kito nuo 0 proc. iki 3,4 proc.

- Požymių atrankos taikymas leidžia reikšmingai sumažinti nagrinėjamų duomenų kiekį ir kartu padidinti klasifikavimo tikslumą, lyginant su pilnais požymių rinkiniais.
    - Pasiūlytieji ir pritaikytieji atrankos metodai leido sumažinti požymių skaičių nuo 6 552 (pilno požymių rinkinio atveju) iki 9–38 požymių tiesioginės schemos atveju bei 7–110 požymių hierarchinės schemos atveju. Požymių rinkinių apimties požiūriu efektyviausias –nuoseklaus aibės didinimo metodas, kuris leido gauti iki 4 kartų mažesnės apimties požymių rinkinius, palyginti su kitais dviem atrankos metodais.
- Hierarchinėje klasifikavimo schemoje skirtingi požymių atrankos metodai neturi esminės įtakos visos schemos efektyvumui.
    - Tyrime skirtingi atrankos metodai lėmė 1,5–5 proc. tikslumo skirtumus vokiečių kalbos atveju ir 2–8 proc. skirtumus lietuvių kalbos atveju.
- Didėjant nagrinėjamų emocijų skaičiui, vidutinis emocijų klasifikavimo tikslumas mažėja, o klasifikavimo tikslumui maksimizuoti reikalingas požymių rinkinio sudėtingumas bei dydis auga.
    - Nagrinėjamų emocijų skaičiui pakitus nuo 3 iki 5, požymių kiekis rinkiniuose padidėjo 1,5–2 kartus. Nagrinėjamam vienos emocijos pavyzdžių skaičiui padidėjus 5 kartus, požymių skaičius rinkiniuose eksperimentų metu padidėjo 2–6 kartus.
    - Nagrinėjamų emocijų skaičius pakitus nuo 3 iki 4, gautųjų požymių rinkinių persidengimas siekė 62 proc., emocijų skaičiui pakitus nuo 4 iki 5 – 36 proc.

5. Trumpai apie autorių

Tatjana Liogienė Vilniaus edukologijos universitete 2003 metais įgijo matematikos bakalauro laipsnį ir 2005 metais – informatikos magistro laipsnį. Nuo 2012 iki 2016 metų buvo Vilniaus universiteto Matematikos ir informatikos instituto doktorante. Nuo 2005 metų dirba lektore Vilniaus kolegijoje.

Tatjana Liogienė

HIERARCHICAL CLASSIFICATION OF SPEECH EMOTIONS

Summary of Doctoral Dissertation
Physical Sciences (P000)
Informatics (09P)
Editor Ieva Vitonytė

Tatjana Liogienė

HIERARCHINIS ŠNEKOS EMOCIJŲ KLASIFIKAVIMAS

Daktaro disertacijos santrauka
Fiziniai mokslai (P000)
Informatika (09P)
Redaktorė Inesa Didikaitė