Baltic J. Modern Computing, Vol. 3 (2015), No. 4, 234-247

Low-Order Multi-Level Features for Speech Emotion Recognition

Gintautas TAMULEVIČIUS, Tatjana LIOGIENĖ

Institute of Mathematics and Informatics, Vilnius University, Akademijos 4, Vilnius, Lithuania

gintautas.tamulevicius@mii.vu.lt, tatjana.liogiene@mii.vu.lt

Abstract. Various feature selection and classification schemes were proposed to improve efficiency of speech emotion classification and recognition. In this paper we propose multi-level organization of classification process and features. The main idea is to perform classification of speech emotions in step-by-step manner using different feature subsets for every step. We applied the maximal efficiency feature selection criterion for composition of feature subsets in different classification levels. The proposed multi-level organization of classification and features was tested experimentally in two emotions, three emotions, and four emotions recognition tasks and was compared with conventional feature combination techniques. Using the maximal efficiency feature selection criterion 2nd and 16th order multi-level feature sets were composed for three and four emotions recognition tasks respectively. Experimental results show the superiority of proposed multi-level classification scheme by 6,3–25,6 % against straightforward classification and conventional feature combination schemes.

Keywords: speech emotion recognition, features, feature selection, classification.

1. Introduction

The main aim of speech emotion recognition task is to identify the emotion state of the speaking person analyzing his speech. Speech emotion recognition emerged as separate area of research in the 9th decade of the last century. Despite the intensive research there is no definitive solution giving accurate and reliable speech emotion identification. Still, the accuracy is not very high, no efficient methods have been proposed for acoustical analysis of the speech signal and classification of the emotional features. Accurate and reliable speech emotion recognition would find its application in criminalistics, call centers, robotics, and enhancement of human-computer interaction (Dellaert et al., 1996; Casale and Russo, 2007; Xiao et al., 2009).

Speech emotion recognition task is a typical classification task. In general, the recognition process can be separated into three stages: acoustical signal analysis, training of classifier and classification based decision process. During the speech signal analysis stage the emotional features of the speech are extracted. Hundreds of various acoustical features are proposed for evaluation speech emotion. Often this variety gives sets of a few thousands features. Unfortunately, high order features cannot guarantee efficient

speech emotion recognition. Thus, efficient speech emotion recognition requires some more decisions regarding feature extraction, training and classification processes.

In this paper a novel multi-stage classification of speech emotions using multi-level features is proposed. The idea of the multi-level features is to use separate particular emotion groups and classify them into particular emotions using their specific feature sets. This allows us to reduce the feature sets and to improve recognition rate of speech emotions.

2. Speech emotion features

There is no general consensus in selection of feature set for speech emotion recognition (Origlia et al., 2010). The researchers use wide variety of features expecting to improve efficiency of the recognition process.

The most popular and frequently used speech emotion features can be grouped into prosodic and spectral features (Rong et al., 2009; Koolagudi et al., 2010). Prosodic features are obtained from pitch frequency, formant frequency values, vocal intensity, energy, pauses, speech duration and rate, voice quality characteristics (Koolagudi et al., 2012; Koolagudi and Rao, 2012). Spectral features are based on short-time signal spectrum properties like linear prediction coefficients, one-sided autocorrelation linear prediction, mel scale cepstral coefficients (Ayadi et al., 2011; Koolagudi and Rao, 2012).

The extracted features are supplemented with derivative statistics. Statistical data of prosodic and spectral feature values such as average, median, standard deviation, dispersion, minimum and maximum values, quantiles and other are used very often as extension of extracted feature sets. Epoch (instant of glottal closure) parameters like strength of epoch, instantaneous frequency, sharpness of epoch, epoch slope strength are also used together with statistical data for speech emotion recognition (Koolagudi et al., 2010). Besides, voice quality features such as excitation signal properties, articulation method, and voice timbre are used also as emotional features of speech (Ayadi et al., 2011).Another important feature proposed is the number of harmonics, caused by nonlinear voice tract properties (Origlia et al., 2010).

Vast majority of speech emotion recognition researches tend to explore huge feature sets up to a few thousand different features. This causes the "curse of dimensionality" problem, when the dimension of feature set is too high to train classifiers properly (because of the insufficient amount of training data). This problem can be solved by enlarging speech data amount or reducing predefined feature sets (Origlia et al., 2010). Feature set reduction methods can be classified into two groups: feature selection methods and feature transform methods (Rong et al., 2009).

Feature selection methods allow to select feature subsets by choosing most effective features or rejecting less significant ones. The most popular approaches are sequential forward selection (Casale and Russo, 2007), sequential backward selection (Casale and Russo, 2007), promising first selection (Dellaert et al., 1996), genetic algorithms (Origlia et al., 2010), maximum relevance – minimum redundancy approach (Peng et al., 2005), and others.

Using sequential forward selection procedure feature set is initialized with the most efficient feature and is cyclically appended with a new one making more efficient feature set. Sequential backward selection, on the contrary, reduces the dimension of initial set by rejecting features to make the set more efficient. Considering the variety of proposed features these procedures can cause a time consuming feature selection process as every

possible variant of feature set should be evaluated in speech emotion identification separately.

Promising first selection approach is based on individual efficiency of every feature. Features are sorted in descending order by their efficiency and the feature set is formed by choosing best features sequentially. The final feature set version is the one giving the lowest classification error.

Maximum relevance – minimum redundancy approach selects features with maximum relevance to analysed emotion class. Relevance is characterized by mutual information between features (Peng et al., 2005; Giannoulis and Potamianos, 2012). Genetic algorithms were proposed for generation and optimization of feature sets also (Casale and Russo, 2007).

The main idea of feature transform methods is optimization of feature sets by transforming dimensionality of feature sets. Various standard mathematical techniques are used for feature transform – principal components analysis (Chiou and Chen, 2013), linear discriminant analysis (You et al., 2006), multidimensional scaling (Rong et al., 2009), Lipschitz spacing method (You et al., 2007), Fisher discriminant analysis (Zhang et al., 2010), neural networks (Gharavian et al., 2012), decision trees (Rong et al., 2009). The main weakness of feature transform approach is pure mathematical operation and defiance of acoustical content of the features.

3. Classification schemes

Additional efficiency of speech emotion recognition can be obtained using different classification schemes. Straightforward usage of conventional classifiers makes speech emotion recognition process dependent on feature set. Unique organization of the classification process can improve emotion recognition even for same feature sets. The examples of such classification scheme can be parallel classification, various hierarchical and multi-stage classification schemes. We will introduce a few proposed classification schemes.

Enhanced co-training algorithm was proposed in order to increase emotions recognitions accuracy during classification step (Liu et al., 2007). Two different feature sets for two different classifiers were used for classification of the six emotions. First 20-dimensional feature set included means, standard deviations, maximums and minimums of fundamental frequency (F0), delta F0, log energy, first and second linear prediction cepstral coefficients features and was used for SVM classifier training. Second feature set included 12 mel-frequency cepstral coefficients and was used for HMM classifier training. Training was repeated up to 18 times for both classifiers using the same labelled data. Each classifier was fed with unlabeled training utterances. Utterances which both classifiers labeled identically were assigned to temporal collection. Further the temporal collection utterances were examined and added into labelled training utterances set. Both trained classifiers were rebuilt and training was repeated up to 18 times again with both classifiers using the updated labelled training utterances. The process is repeated until unlabeled training utterance set will become empty. Gender information in this research was used too and female and male utterances were classified separately. The obtained emotion recognition accuracy was 75,9 % for females and 80,9 % for males.

Fusion among different classifiers was also proposed for recognition improvement (Zhang et al., 2010). Fusion principle was implemented by using queuing voting

algorithm. Three kinds of classifiers with different feature set (obtained by using promising first selection method) were used. Majority voting principle was extended with confidence weights and the final decision on emotion is obtained considering these weights.

Fusion approach requires individually efficient classifiers and feature sets as the performance of a separate classifier affects efficiency of the whole scheme. Thus, this fusion classification scheme also requires careful selection of classifiers and features.

Two-stage hierarchical classification scheme based on gender separation was proposed in (Yoon and Park, 2011). During the first step all emotional speech utterances are classified using gender specific pitch feature into three emotion groups: male (or neutral), female (or anger), and unknown group (Fig. 1). The number and type of classes is determined by range of pitch feature values. The goal of the second step is to classify utterances of unknown group into two more classes: anger or male and neutral state or female. The second step classification is performed using additional energy, cepstral, and delta features. The total order of features in this scheme was 56 and was fairly low in comparison with a few hundreds or even thousands in straightforward classification. The average emotion recognition rate was 80,7 %.

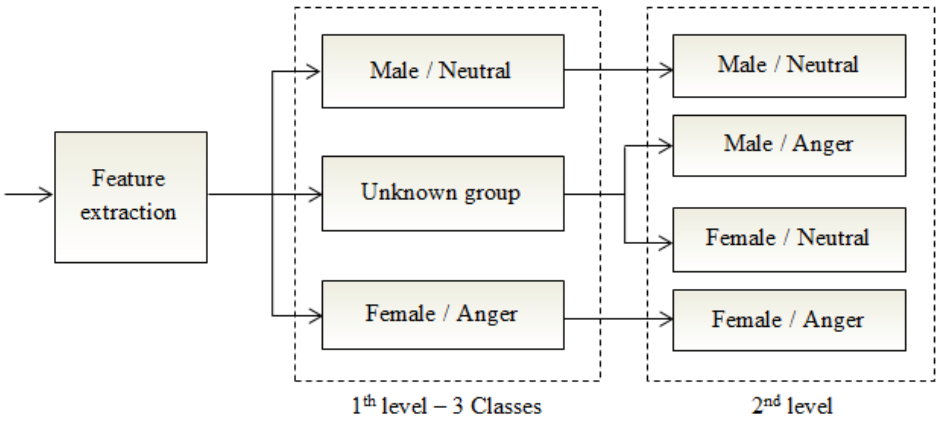

Fig. 1. Two-step hierarchical classification of two emotions (Yoon and Park, 2011).

Another proposed hierarchical two-stage classification scheme (Xiao et al., 2009) is given in Figure 2. During the first step all utterances are classified by arousal dimension into active and not-active emotions (the last ones are classified into median and passive emotions additionally). During the next stage every emotion group is classified into two specific emotions. The active emotion group is classified into anger and gladness (joy), median emotion group is classified into fear and neutral state, and passive group is classified into sadness and boredom. Different feature sets and classifiers were used in each classification stage. Overall feature set consisted of 68 features obtained using sequential forward selection method. This scheme gave average recognition accuracy of 76,4 %.

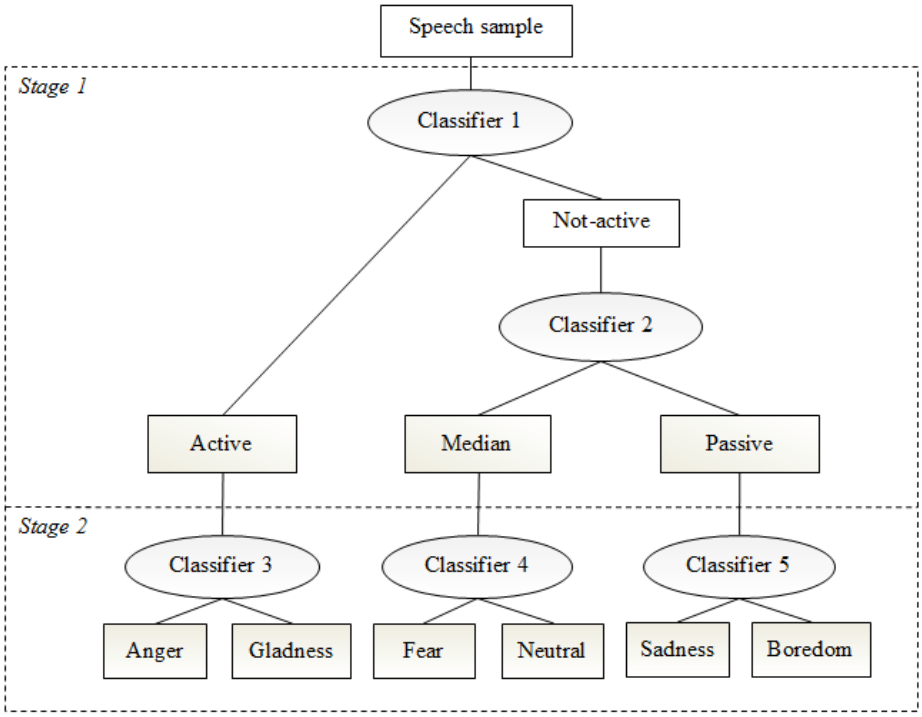

Fig. 2. Two-stage hierarchical classification scheme driven by dimensional emotion model (Xiao et al., 2009).

The main idea of sub-system based hierarchical classification scheme (Giannoulis and Potamianos, 2012) is emotion specific training (Fig. 3). Six emotions were analyzed: anger, joy, sadness, fair, boredom and neutral. All these emotions were grouped into 15 emotion pairs (called sub-systems). Every sub-system is analyzed using particular feature set for the emotion recognition. These feature sets were obtained from general feature set of 112 different features applying sequential backward selection and maximum relevance – minimum redundancy approaches. The overall emotion recognition accuracy was 85,2 % in gender dependent experiment and 80,1 % in gender independent case.

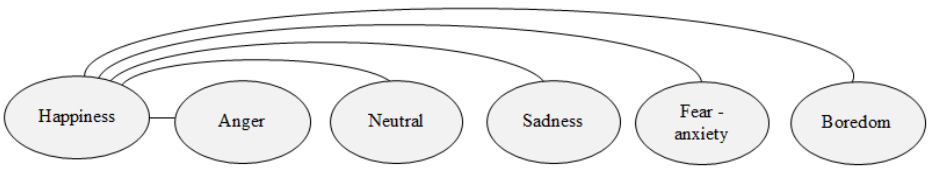

Fig. 3. Five sub-systems for recognizing “happiness”, denoted by the five lines connecting the ovals (Giannoulis and Potamianos, 2012).

Another hierarchical approach for emotion classification (Lugger et al., 2009) is introduced in Figure 4. Six emotions (anger, happiness, anxiety, neutral, boredom, sadness) were classified in three stages by using multiple Bayesian classifiers. 25 features from 333 were selected by using the sequential floating forward selection algorithm. This classification scheme based on emotion model covering the 3 dimensions of emotion: activation, potency and evaluation. The patterns are classified in binary mode (2 particular classes are separated in every step) using 3-dimensional emotion model based features. Therefore the features in every classification level are limited by this model. The first classification level is intended for separation of high activation and low activation emotion classes. Low potency and high potency classes are separated during the second classification stage. In the third stage all classes are classified into separate emotions. This hierarchical scheme of emotion classification gives classification rate of 59,7 % .

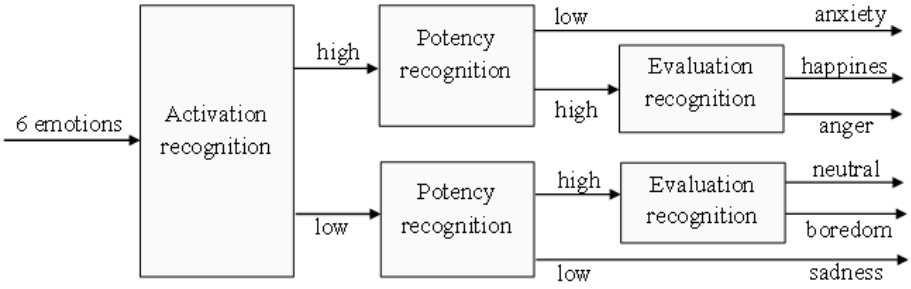

Fig. 4. Design of a 3-stage hierarchical combination classifier (Lugger et al., 2009).

The genetic algorithms were used for feature subset selection for multistyle classification of emotional speech (Casale and Russo, 2007). Five feature subsets were used for pairwise and multistyle classification. Feature subsets of 16 to 48 features were analyzed for classification of four speaking styles: neutral, angry, loud and Lombard. The highest classification performance of 82,74 % was obtained using multistyle classification with 48 features.

4. Multi-level approach

In this paper we present multi-stage speech emotion classification scheme using multi-level features. The main idea of this scheme is to perform classification of speech emotions in step-by-step manner using different feature sets in every step.

Let us formulate three main presumptions on multi-stage classification of speech emotions:

- Recognition of all emotions in one step is still a complicated process because of overlapping acoustic, prosodic and other features of the emotions. Classification problem can be simplified by reducing the number of analysed

emotions at a time. This could be done by organizing emotion classification process in stages with limited number of analysed emotions in every stage.

- Each emotion is characterized by its own acoustic and prosodic features. These features for various emotions can be different or the same. Composing a feature set by maximizing average classification rate for the entire set of the speech emotions we cannot ensure the maximal classification accuracy for individual emotion. This can be achieved by analysing every emotion (or a group of emotions characterized by the same feature) separately.
- Emotions, depending on the selected feature or features set, can be classified into various classes. The classes themselves can be decomposed to the lower level classes and etc., until the single emotion class is obtained. For example, classification by pitch frequency can give us high-pitch (happiness, anger) and low-tone (neutral, sadness, boredom) emotions. Each of these classes can be decomposed to separate emotion using duration, energy and other features as classification feature.

In accordance with these assumptions the new multi-stage speech emotion classification scheme was proposed. The visual generalization of the proposed scheme is given in Figure 5.

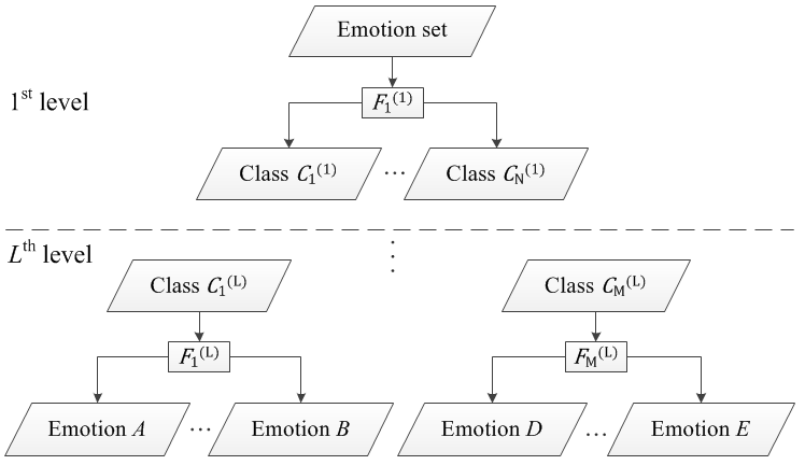

Fig. 5. The generalized scheme of emotion classification using multi-level features

First of all we perform the first stage (we will call them classification level or level simply) classification. The whole set of unknown speech emotion patterns are classified into N classes $\{C_1^{(1)}, \dots, C_{\mathrm{N}}^{(1)}\}$ using first level feature set $F_1^{(1)}$. This feature set should be selected to maximize the accuracy of classification into $\{C_1^{(1)}, \dots, C_{\mathrm{N}}^{(1)}\}$. On the second level every class of the $\{C_1^{(1)}, \dots, C_{\mathrm{N}}^{(1)}\}$ is classified into lower level classes $\{C_1^{(2)}, \dots, C_{\mathrm{K}}^{(2)}\}$ using its specific second level feature set $F_k^{(2)}$, $k = 1, \dots, \mathrm{K}$. The classification process is repeated as long as we get separated particular emotions. The main idea of multi-level classification is to use specific and most powerful feature set in every level for every class. Thus we can guarantee the appropriate feature set for every classification level i.e. every emotion.

The proposed classification uses so called multi-level feature set. Every partial feature set $F_m^{(l)}$ is applied for particular emotion or emotion group classification thus organizing emotion classification into separate levels. The main principles of classification using multi-level features are following:

- The classification is organized in separate levels. In order to identify emotions or emotion groups, the particular feature or a set of these features is used in every level. These features sets can be composed using aforementioned sequential forward selection, sequential floating forward selection, sequential backward selection, maximum relevance – minimum redundancy based selection or any other feature selection technique.
  In this paper we applied the maximal efficiency feature selection criterion enabling us to employ features with the lowest classification error
  $$F_m^{(l)} = \left\{ arg \min_j E\left(F_j^{(l)}\right) \right\}, \qquad j = 1, \dots, J.$$
  here $E\left(F_j^{(l)}\right)$ – classification error in the l-th level using j-th feature subset $F_j^{(l)}$. J denotes the total number of features in the l-th classification level.
  Most efficient features $F_j^{(l)}$ are added to the feature set $F_m^{(l)}$ repeatedly. The expansion of feature set $F_m^{(l)}$ is stopped when the extended feature set does not show any improvement in classification rate.
- The set of features of the speech emotion recognition problem is formed as combination of all the employed subsets $F_m^{(l)}$.
  $$F = \left\{ F_m^{(l)} \right\}, \qquad m = 1, \dots, M;\ l = 1, \dots L.$$
  Here M is the number of emotions classes in particular classification level, L is the number of classification levels.

In general, the set of emotions (or the set of classes derived from higher level class) can be classified to any number of classes. The number of analysed classes (in one level) and the number of classification levels are defined by the overall number of emotions and the selected feature subsets. The simplest case of multi-level classification is the classification into two classes (emotions).

Feature subsets can be heterogeneous – composed of various features (time, spectral, cepstral, energy, voice quality and etc.). The user is free to choose any feature subsets using predefined feature selection criterion.

The main advantage of our proposed multi-level feature organization is as follows. Different level classification processes are independent from the feature viewpoint. Thus we can optimize classification process of any selected emotion group without affecting others.

5. Experimental research

The proposed multi-level classification scheme was experimentally tested in different speech emotion recognition tasks. In this study we analysed 2 emotions (joy and anger), 3 emotions (joy, anger, and neutral states), and 4 emotions (joy, anger, neutral state, and sadness) recognition cases.

Recordings of the freely accessible Berlin emotional speech database (Burkhardt et al., 2005) were used for recognition experiments. To ensure homogenous experimental

conditions the equal number of each emotion patterns was selected for classification. As the number 60 is quite low for reliable classification estimation 3-fold testing methodology was applied in this study. All the results in this paper are averaged results of the 3-fold testing.

Considering the data amount we have chosen non-parametric *k*-Nearest neighbours (kNN) classifier. As the classifier is not the goal of our investigation we will use the same kNN classifier ($k = 7$) for all classification levels of our scheme. In general case any type of classifier can be implemented in different levels.

We have decided to restrict our recognition experiment to a fundamental frequency (F0) based features. Six groups of F0 features were analysed in our experiment (Eyben et al., 2009):

- Smoothed static low-level and functional F0 features;
- 1st order low-level and functional F0 delta features;
- 2nd order low-level and functional F0 delta features;
- Envelope features of the smoothed F0 contour;
- 1st order delta envelope features;
- 2nd order delta envelope features.

Each group contained 39 distinct features: the absolute and the arithmetic means of the F0 contour, positions of the minimal and maximal F0 values, various order statistical moments and quartiles, and others. We understand, the set of these features is not sufficient for reliable speech emotion recognition and they should be appended with more various acoustical features. Even so, we think F0 based feature set will be competent to illustrate the principle of the multi-level features and multi-level classification of speech emotions.

For deeper understanding of multi-level classification and features let us elaborate the case of 4 emotions recognition. The intermediate classes are determined by the chosen feature set. For example, speech rate feature set will divide speech patterns into high tempo and low tempo classes. Energy based feature set will give us high energy and low energy classes. In our case using fundamental frequency based features all emotions can be divided into low-pitch (sadness, neutral state) and high-pitch (anger and joy) classes. Thus the first level of classification will be based on highness of fundamental frequency and will result in two emotion classes: low-pitch and high-pitch. On the next level these two classes can be classified into 4 above mentioned emotions. Thus, we will have two-level classification and two-level features (Fig. 6).

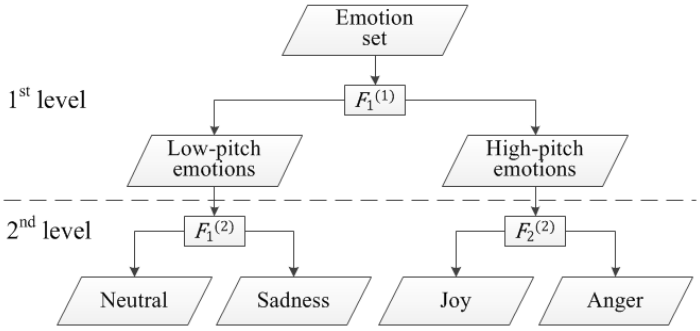

Fig. 6. Organization of 4 emotion recognition process

In case of 3 emotions (joy, anger, and neutral state) task we will have two-level classification also. The low-pitch class should contain neutral state patterns only so the second level will contain only one classification process using feature set $F_2^{(2)}$ (Fig. 6).

In case of 2 emotions (joy and anger) multi-level classification scheme becomes a simple classification of 2 emotions and the principle of multi-level features disappears.

In order to define the most efficient features for separate classification levels we carried out experimental testing of separate F0 features firstly. Each of 234 features was individually tested in 2 emotions, 3 emotions, and 4 emotions classification tasks – 702 classification tests were carried out totally. The feature set $F_m^{(l)}$ was initialized with the most efficient feature and expanded recurrently using next most efficient features. The expansion of the feature set is stopped when the classification error using this set achieves minimal value. Table 1 presents all the formed feature subsets.

Table 1. Selected feature sets

Classification task	Level	Feature subset	Features
2 emotions	1st level	$F_1^{(1)}$	F0env_sma_iqr1-3
	2nd level	–	–
3 emotions	1st level	$F_1^{(1)}$	F0env_sma_variance
	2nd level	$F_1^{(2)}*$	–
		$F_2^{(2)}**$	F0env_sma_iqr1-3
4 emotions	1st level	$F_1^{(1)}$	F0_sma_de_de_iqr1-3, F0_sma_iqr2-3, F0_sma_quartile3, F0_sma_iqr1-3, F0_sma_qregc2, F0env_sma_de_de_quartile1, F0env_sma_de_de_iqr1-2, F0_sma_de_nzabsmean.
	2nd level	$F_1^{(2)}*$	F0_sma_peakMeanMeanDist, F0_sma_de_de_meanPeakDist, F0_sma_qregc1, F0_sma_de_nzgmean, F0_sma_meanPeakDist, F0_sma_de_nzabsmean, F0env_sma_de_de_qregc2, F0_sma_peakMean, F0env_sma_maxPos.
		$F_2^{(2)}**$	F0env_sma_iqr1-3

* for classification of neutral state and sadness;
** for classification of joy and anger.

We can see that most efficient feature subsets are different for high-pitch and low-pitch emotions. Besides, feature sets for different tasks differ too (for example, 1st level feature subsets for three emotions and four emotions tasks differ significantly). Hence our presumption about specific acoustic features for the particular emotion group was correct.

In case of two emotions task there is only one classification level, thus the 2nd level and the 2nd level feature subset $F_2^{(2)}$ are absent. In this case the goal of classification is to separate joy and anger.

Separate results of the first and the second classifications levels are given in Tables 2 and 3.

Table 2. First level classification results

Classification task	Classification rate		
	Low-pitch emotions	High-pitch emotions	Average
2 emotions	–	–	–
3 emotions	81,7 %	87,5 %	84,6 %
4 emotions	74,2 %	85,8 %	80 %

In case of two emotions task the average classification rate was 65,8 % (60 % rate for anger and 71,7 % rate for joy). The rate is quite satisfactory considering the 1st order feature set (see Table 1, please) and complexity of the task (acoustical properties of joy and anger overlap in fundamental frequency domain heavily). In general, classification rate decreases with the growing number of analysed emotions.

Table 3. Second level classification results

Classification task	Classification rate				
	Neutral	Sadness	Anger	Joy	Average
3 emotions	100 %	–	55,2 %	73,1 %	76,1 %
4 emotions	86,9 %	81 %	56,9 %	73,2 %	74,5 %

Table 3 shows 100 % classification rate in case of 3 emotions task. This value should be interpreted as absence of classification. The entire low-pitch emotion group was labelled as neutral state. Again, in all cases separation of anger and joy was the most complicated part of the task.

Analysing classification results (Tables 1, 2, and 3) we can notice that most efficient feature sets include static, first (the feature title includes suffix *de*) and second (the feature title include suffix *de_de*) order delta features. For example, delta features dominate in sets for 4 emotions classification task. Thus delta features are very important for accurate speech emotion classification.

Having first and second level feature subsets we can implement multi-level recognition of speech emotions. Table 4 gives averaged results of the entire speech emotion recognition process.

As we can see average speech emotion recognition rates vary from 65,8 % (for two emotions) to 59,6 % (in four emotions case). Results are satisfactory considering the feature set order. We obtained 1st, 2nd, and 16th order features for two, three, and four emotions recognition tasks respectively. These values are extremely low in comparison with widely published speech emotion recognition results therefore we can denote our proposed multi-level features as low-order.

Table 4. Speech emotion recognition results

Recognition task	Recognition rate				
	Neutral	Sadness	Anger	Joy	Average
2 emotions	–	–	60 %	71,7 %	65,8 %
3 emotions	80 %	–	48,3 %	66,7 %	65 %
4 emotions	55 %	68,3 %	50 %	65 %	59,6 %

Obtained average recognition results are lower in comparison with above given alternative classification scheme results (59,7–85,2 %). This could be explained with lower order of multi-level feature sets, different number of emotions and with restricted feature space in our experiment. In aim to increase obtained recognition rate the used F0 feature sets should be extended with more various features. This would give order feature sets and definitely higher classification.

For comparison purposes the proposed multi-level features were compared with various feature sets. Usually feature sets are composed by joining various features. Often these features are chosen without any selection procedure thus giving high order sets. These feature sets as the rule are used for straight classification schemes, where all utterances are classified into emotions in one step. To imitate this feature set composition and classification techniques in this study we tested the feature set F_{234} including all 234 features (used in above experiments) and the set F_{B} including all the features used in multi-level scheme. Recognition was performed under the same circumstances: the same speech data partition and the same kNN classifier were used. Averaged recognition results using these feature sets are presented in Figure 7.

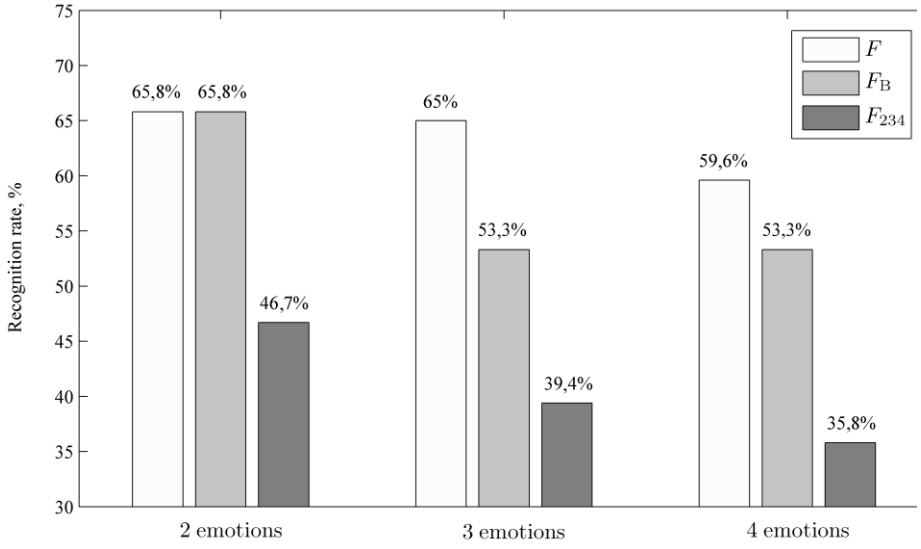

Fig. 7. Averaged speech emotion results for different feature sets

We can see that multi-level organization of classification (feature set F) gave higher speech emotion recognition rate in comparison with straight classification scheme using conventional feature sets (F_{B} and F_{234}). Superiority of multi-level feature organization ran from 6,3 % (in 4 emotions classification task) up to 25,6 % (3 emotion case). The main reason for this is application of specific acoustic features for particular emotion (or emotion group) classification in multi-level scheme. Results of feature sets F and F_{B} for two emotions recognition coincided because multi-level classification becomes identical to straight classification scheme in this case.

Superiority of feature set F against F_{B} by 6,3–11,7 % shows the superiority of multi-level classification scheme against straightforward classification as the features in these sets were the same. Superiority of feature set F against F_{234} by 23,8–25,6 % proves the superiority of multi-level features. Low-order multi-level features enable us to recognize speech emotions more accurately than 234th order feature set. Besides, in case of feature set F_{234} we have got the lowest recognition rate. This proves the necessity of feature selection process as the full set of features cannot give high classification accuracy.

6. Conclusions

The multi-level organization of features was proposed for speech emotion recognition. The main idea is to organize speech emotion recognition in levels, where every level of classification uses specific features for particular emotion group. The advantage of multi-level organization is independent feature subsets for emotion groups. This enables us to maximize classification rate of any selected emotion group without affecting another.

The proposed classification scheme and feature sets were applied for two emotions, three emotions, and four emotions recognition tasks and were compared with conventional feature combination techniques. Multi-level classification scheme enabled us to increase speech emotion recognition rate by 6,3–25,6 % in comparison with straightforward classification and conventional feature combination schemes. We obtained low-level 1st, 2nd, and 16th order features for two, three, and four emotions recognition tasks respectively.

With reference to obtained experimental results we state:

- Multi-level feature organization enables us to apply specific features for particular emotion thus improving recognition rate of separate emotions without affecting other ones.
- Multi-level organization of classification and features improves speech emotion recognition rate in comparison with straight organization of recognition process.
- Multi-level organization of features gives lower order feature sets in comparison with conventional feature combination techniques without selection. The combination of feature sets without selection is inexpedient.

References

Ayadi, M., Kamel, M., & Karray, F. (2011, March). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, pp. 572–587.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech. *In: Proceedings of Interspeech*, pp. 1517-1520.

Casale, S., & Russo, A. (2007). Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. *Speech Communication*, pp. 801-810.

Chiou, B.-C., Chen, C.-P. (2013). Feature Space Dimension Reduction in speech emotion recognition using Support Vector Machine. *Signal and Information Processing Association Annual Summit and Conference*, pp. 1 - 6.

Dellaert, F., Polzin, T., Waibel, A. (1996, October). Recognizing emotion in speech. *Fourth International Conference on Spoken Language, 3*, pp. 1970-1973.

Eyben, F., Wollmer, M., Schuller, B. (2009, September 10-12). openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *Affective Computing and Intelligent Interaction and Workshops*, pp. 1-6.

Gharavian, D., Sheikhan, M., Nazerieh, A., Garoucy, S. (2012, November). Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. *Neural Computing and Applications*, pp. 2115-2126.

Giannoulis, P., Potamianos, G. (2012, May). A hierarchical approach with feature selection for emotion recognition from speech. *In: Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pp. 1203-1206.

Koolagudi, S., Rao, K. (2012, June). Emotion recognition from speech: a review. *International Journal of Speech Technology*, pp. 99-117.

Koolagudi, S., Reddy, R., Rao, K. (2010, July 18-21). Emotion recognition from speech signal using epoch parameters. *International Conference on Signal Processing and Communications*, pp. 1 - 5.

Liu, J., Chen, C., Bu, J., You, M., Tao, J. (2007 m. July 2-5 d.). Speech Emotion Recognition using an Enhanced Co-Training Algorithm. *2007 IEEE International Conference on Multimedia and Expo*, p. 999 - 1002.

Lugger, M., Janoir, M.-E., Bin Yang. (2009). Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. *17th European Signal Processing Conference*, pp. 1225–1229.

Origlia, A., Galata, V., Ludusan, B. (2010). Automatic classification of emotions via global and local prosodic features on a multilingual emotional database. *In: Proceedings of Speech Prosody*.

Peng, H., Long, F., Ding, C. (2005, August). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1226-1238.

Rong, J., Li, G., Chen, Y.-P. P. (2009, May). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing and Managemen*, pp. 315–328.

Xiao, Z., Centrale, E., Chen, L., Dou, W. (2009, September 10-12). Recognition of emotions in speech by a hierarchical approach. *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1 - 8.

Yoon, W.-J., Park, K.-S. (2011). Building Robust Emotion Recognition System on Heterogeneous Speech Databases. *2011 IEEE International Conference on Consumer Electronics*, pp. 825-826.

You, M., Chen, C., Bu, J., Liu, J., Tao, J. (2006, July 9-12). Emotion Recognition from Noisy Speech. *IEEE International Conference on Multimedia and Expo*, pp. 1653-1656.

You, M., Chen, C., Bu, J., Liu, J., Tao, J. (2007, March). Manifolds based emotion recognition in speech. *International Journal of Computational Linguistics and Chinese Language Processing*, pp. 49-64.

Zhang, S., Lei, B., Chen, A., Chen, C., Chen, Y. (2010, October 24-28). Spoken emotion recognition using local Fisher discriminant analysis. *IEEE 10th International Conference on Signal Processing*, pp. 538 - 540.

Zhang, Y., Wang, C., Fu, L. (2010, October 29-31). Classifier fusion for speech emotion recognition. *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on, 3* , pp. 407 - 410.

Received June 23, 2015, revised September 15, 2015, accepted October 11, 2015