Proceedings of the Federated Conference on Computer Science and Information Systems pp. 483–486
DOI: 10.15439/2016F316
ACSIS, Vol. 8. ISSN 2300-5963

Comparative Study of Multi-Stage Classification Scheme for Recognition of Lithuanian Speech Emotions

Tatjana Liogienė, Gintautas Tamulevičius
Vilnius University Institute of Mathematics and Informatics,
Akademijos str. 4, Vilnius, Lithuania
Email: {tatjana.liogiene, gintautas.tamulevicius}@mii.vu.lt

***Abstract*—This paper presents the experimental study of multi-stage classification based recognition of Lithuanian speech emotions. Three different criteria for feature selection were compared for this purpose: Maximal Efficiency, Minimal Cross-Correlation feature criterions, and the Sequential Feature Selection. A large database of spoken emotional Lithuanian language was used in this experiment – each of 5 emotions was represented by 1000 utterances. The results of the speaker-independent emotion recognition experiment show the superiority of multi-stage classification using the SFS technique by 0.7-8 %. This classification scheme gave the highest recognition accuracy and the smallest feature set.**

I. Introduction

SPEECH emotion recognition is a classical task of pattern classification including feature extraction, training and classification (decision making). The feature extraction step is a crucial for the successful speech emotion identification process: appropriate and relevant feature set is a key component of any valid and efficient recognition system.

Various feature sets have been proposed for speech emotion recognition [1]-[5]. In straightforward manner composed feature sets often contain a few hundred or even thousand features and this can become problematic in case of limited datasets. Thus various feature selection or transformation techniques are applied for reduction purposes [2], [3], [6], [7]. Various parallel, serial, and hierarchical classification schemes have been proposed and proved to be more effective for speech emotion recognition [2], [4], [8], [9] also.

In this paper we present multi-stage classification based recognition of Lithuanian speech emotions. Section II contains the review of multi-stage classification of speech emotions. The multi-stage classification scheme using three different feature selection criteria is presented in next section. The results of the experimental study are given in Section IV and concluded in Section V.

II. Multi-stage Classification of Speech Emotions

The classification of speech emotion can be implemented in two ways. The simplest is to classify emotions in one step using one general feature set for all emotions. Usually this means a very large but not optimal feature set.

The interest in sophisticated classification schemes has been noticeable in last few years. Variations of classification scheme include multi-stage classification (when the whole recognition process is implemented in a few steps), multiple classifier schemes (different classifiers are dedicated to separate emotions or emotion groups), pair-wise classification and others. All these classification schemes can be arranged into three groups: serial, parallel, and hierarchical (Table I).

The serial combination of classifiers considers the speech emotion classification process as the consecutive identification of one or more separate emotions during one classification step. *N-1* separate classifiers will therefore be needed to identify *N* emotions [2]. The parallel scheme is based on the concurrent identification of separate emotions – all the emotions are analyzed by a set of classifiers during one step [2], [9]. The third and the biggest group of multiple classifier systems is based on the hierarchical organization of the classification process according to some criterion.

The speaker's gender, three dimensional emotion model based groups and other criteria are used for hierarchical organization of the classification process. In general, the hierarchical group contains attributes of both the serial and parallel schemes [1]-[9].

As we can see, multi-stage organization of the speech emotion classification process results in a complicated process and the accuracy obtained varies from 50 % up to 88 %. Nevertheless, the above-mentioned multi-stage classification schemes outperform single-step schemes and provide an opportunity to modify the feature set for a particular emotion or emotion group without affecting another. This should be considered as the main advantage of multi-stage classification of speech emotion.

III. Feature Selection Based Multi-stage Classification

Considering the above-mentioned advantages, we proposed a multi-stage classification scheme for speech emotion recognition [10]. The main idea of the proposed scheme is the grouping of emotions for different classification stages. All groups of emotional speech

978-83-60810-90-3/$25.00

TABLE I
EMOTIONAL SPEECH CLASSIFICATION SCHEMES

No	Authors	Classification Schemes	Number of Emotions	Language	Accuracy
1.	W.-J. Yoon, K.-S. Park [8]	Two-step hierarchical classification	2	Chinese	80.7%
2.	J. Liu, et al. [1]	Enhanced co-training algorithm	6	Chinese	75.9% male, 80.9% female
3.	Z. Xiao, et al. [3]	Hierarchical classification	6	German	76.4%.
4.	M. Lugger, etal. [2]	Hierarchical combination of classifiers	6	German	88.8%.
5.	M. Kotti, F. Paterno [7]	Psychologically-inspired binary cascade classification scheme	6	German	87.7%
6.	C.-C. Lee, et al. [5]	Hierarchical binary decision tree approach	5	German	48.27%
7.	L. Chen, X. Mao, Y. Xue, and L. L. Cheng [6]	Three-level classification model	6	Chinese (Mandarin)	86.5%, 68.5%, and 50% (for each level)
8.	E. M. Albornoz, D. H. Milone, and H. L. Rufiner [4]	Two-stage hierarchical classification	7	German	71.75%
9.	M. Lugger, M.-E. Janoir, and B. Yang (2009) [2]	Serial combination of classifiers	6	German	96.5%
10.	M. Lugger, M.-E. Janoir, and B. Yang [2]	Parallel combination of classifiers	6	German	92.6%
11.	A. Milton and S. Tamil Selvi [9]	Class-specific multiple classifiers scheme	7	German	80.6%

utterances are labeled in several stages. During the first stage all utterances are classified into predefined groups. During successive stages, these groups are divided into subgroups or separate emotions. This classification scheme enables us to use different (more effective, we suppose) feature sets per classification node, thereby improving the overall recognition rate. The feature set for every node (we will call this set as subset) is formed individually according to performance on the emotional group analyzed.

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

A. Maximal Efficiency Feature Selection Criterion

This criterion is applied by making an assumption about the aggregate efficiency of features with maximal individual efficiency i.e. of features giving the lowest classification error. The formation of a feature subset using the ME selection criterion is carried out

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

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 repeatedly extended with the most effective features $f_j^{(l)}$. The evaluation of every subset case is carried out and the extension process is stopped when the overall efficiency of the subset is not improved. Thus the selection procedure of J features from M feature set will require analysis of J+M-1 feature subsets.

B. Minimal Cross-Correlation Criterion

In this case an assumption is made as to the efficiency of linearly independent features. Independent features make the set more effective than strongly correlated ones. Thus by selecting linearly independent features we seek for a more effective subset.

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

$$f_m^{(l)} = \arg\min_j \left| R\left(f_0^{(l)}, f_j^{(l)}\right)\right|, j = 1,...,J. \qquad (2)$$

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 is stopped when the efficiency of the feature subset begins to diminish. Similar to ME criterion the selection procedure of J features from M feature set using MC criterion will result in analysis of J+M-1 subsets.

Considering the unknown distribution of emotion feature values, the Spearman coefficient was selected to evaluate the correlation of the features. Moreover, the Spearman correlation is hypothesized as being more robust to data outliers, an aspect we find important in the case of speech emotion features.

C. Sequential Forward Selection

The SFS technique is one of the acquisitive search algorithms aiming to find the most significant subset of the features, and the aggregate efficiency of the feature subset is considered rather than individual properties of the features.

The selection 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

$$f_m^{(l)} = \arg\max_j \left[E\left(F_i + f_j^{(l)}\right) - E\left(F_i\right) \right], j = 1,...,J. \quad (3)$$

The feature set extension step is repeated until the efficiency of newly obtained feature set F_{i+1} increases or while $j \le J$. $J \times M$ different feature subsets should be analyzed to select J-th order feature subset using this procedure. Therefore, the SFS will require number analyzed feature subsets grows significantly in comparison with aforementioned criteria.

The applied Maximal Efficiency and Minimal Cross-Correlation feature selection criteria, and the Sequential Forward Selection Technique make locally optimal choices and will thus give suboptimal feature subsets.

IV. Experimental Study

In this study we decided to perform a thorough comparison of the aforementioned feature selection criteria for the recognition of Lithuanian speech emotions. Three versions of the multi-stage classification scheme were implemented using different feature selection criteria and applied to the Lithuanian speech emotion identification task.

We have chosen recognition tasks for 3 emotions (anger, joy, neutral), 4 emotions (anger, joy, neutral, sadness), and 5 emotions (anger, joy, neutral, sadness, and fear). 1000 examples of each emotion (recorded by 5 females and 5 males) were analyzed during the experiment [11].

The initial full set consisted of 6552 different speech emotion features including time and frequency domain features, mel scale features, probabilities of voicing in speech, and their various derivatives (first and second order differentials, statistics, distribution data) [12].

A non-parametric K-Nearest Neighbor classifier was chosen for experimental testing. The value $K = 5$ was selected considering the large size of the data sets.

A two-stage classification scheme was designed assuming low-pitch and high-pitch emotion classes in the first classification stage. These two groups are classified into separate emotions during the second stage using a group-specific feature subset.

Considering the number of examples for each emotion, a 10-fold cross-validation scheme was selected to obtain more robust results. As every speaker pronounced 100 emotional sentences, the speech emotion recognition experiment was performed in speaker-independent mode.

The average recognition results are given in Fig. 1, detailed results for every emotion are given in Table II (the Maximal Efficiency criterion is denoted as *ME*, Minimal Cross-Correlation criterion denoted as *MC*, and the SFS technique denoted as *SFS*).

As we can see in Fig. 1, the SFS based multi-stage recognition of speech emotions showed the highest accuracy in all cases. Its superiority over other selection criteria based schemes was 0.7-8 %. Two things should be pointed about these results. To begin with, the average results are much lower than our previously obtained results. In the case of 5 emotions, the recognition rate was 50 % approximately (not impressive in comparison with results from other studies). Besides, the superiority of the SFS based multi-stage scheme is much smaller in comparison with previous results. There are two possible reasons for the lower results:

- Non-professional actors: the valence of the emotions is much lower in comparison with emotions expressed by professional actors. Consequently, the classification of these patterns is a more challenging task.
- The size of the database: the much larger set of emotional utterances contains more different verbal expressions, thus the variability of the utterances is much wider and more confusing.

The MC criterion produced the lowest recognition rates for emotional cases. In case of 5 emotions, the recognition rate was 41.6 %. Nevertheless, the single-stage recognition of 5 emotions (using the entire set of features) has shown an accuracy of 28.4 % only. Thus multi-stage classification has obvious superiority over single-stage classification.

Analysis of the recognition results for particular emotions reveals that anger and the neutral state are the most difficult emotions to recognize among the others.

Figure 2 shows the dependence of the obtained feature set size (order) on the number of recognized emotions. Again, the SFS based multi-stage scheme demonstrated the highest efficiency. The size of the feature sets was 2.5-4 times smaller in comparison with the cases of ME and MC selection criteria. The highest order was obtained for the MC criterion; the total size of the feature set was 90-110. In general, the order of the feature set increases with the number of analyzed emotions. This could be caused by suboptimal feature selection techniques.

Analyzing individual results from every speaker, we have noticed a fluctuation in recognition rate amongst speakers. For example, the average recognition results of speaker #9

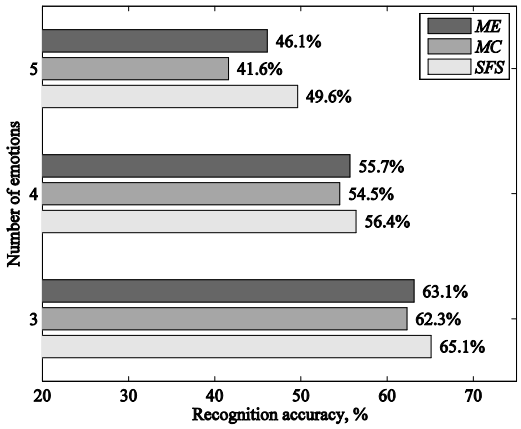

Fig 1. Average emotion recognition results

TABLE II
RECOGNITION RATES FOR PARTICULAR EMOTIONS

Criterion	Number of emotions	Recognition accuracy, %				
		Anger	Joy	Neutral	Sadness	Fear
ME	3	56.6	67.3	65.5	–	–
	4	53.5	62.2	50.1	56.8	–
	5	42.7	48.1	45.1	49.3	45.2
MC	3	54	62.6	70.3	–	–
	4	49.6	58.3	46.4	63.8	–
	5	42.4	38.1	40.3	52.8	34.6
SFS	3	57.6	71.7	66	–	–
	4	52	67.6	48	57.8	–
	5	49.4	60.6	41.2	50.3	46.7

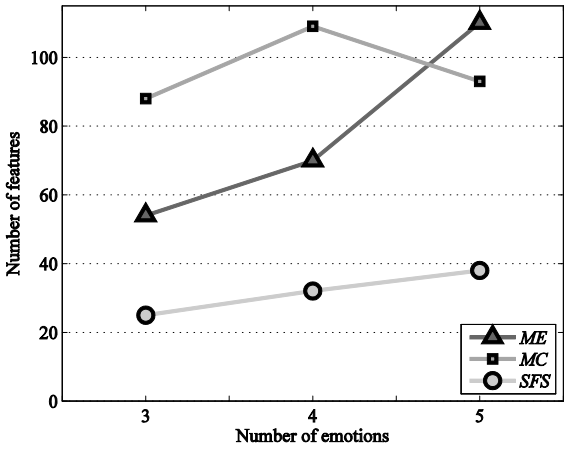

Fig 2. Feature order dependence on number of emotion

were 1.5-2.5 times lower than average. The results of speaker #10 were 1.2-1.3 times higher than the average results. The reason is the suboptimal feature selection aiming for the highest average recognition rate not for the individual one.

V. CONCLUSIONS

In this paper the results of a comparative study of a multi-stage classification scheme for Lithuanian speech emotion recognition are presented. Three different feature selection criteria were applied for recognition purposes: Maximal Efficiency, Minimal Cross-Correlation of features and Sequential Forward Selection. The following conclusions can be drawn from the results:

- The average recognition rate was 62-65 % for the 3 emotion set (anger, joy, and neutral), 55-57 % for the 4 emotion set (anger, joy, sadness, and neutral), and 42-50 % for the 5 emotion set (anger, joy, sadness, fear, and neutral). The results are not impressive in comparison with results from other studies, but the large set of emotional utterances should be considered as the main factor for the accuracy obtained.
- Sequential Forward Selection based scheme shows higher performance in comparison with individual feature properties based selection criteria (Maximal Efficiency and Minimal Cross-Correlation in our case). The superiority was 0.7-8 %.
- The recognition of a large number of emotional utterances requires large feature sets. Increasing the number of recognized emotions also expands the size of the required feature set. Consequently, speaker and text-independent speech emotion recognition would require for huge feature sets.

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, http://dx.doi.org/10.1109/ICME.2007.4284821.

[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, http://dx.doi.org/10.5281/zenodo.41415.

[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, http://dx.doi.org/10.1109/ACII.2009.5349587.

[4] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language,* pp. 556–570, 2011, http://dx.doi.org/10.1016/j.csl.2010.10.001.

[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, http://dx.doi.org/10.1016/j.specom.2011.06.004.

[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, http://dx.doi.org/doi:10.1016/j.dsp.2012.05.007.

[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, http://dx.doi.org/10.1007/s10772-012-9127-7.

[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, http://dx.doi.org/10.1109/TCE.2011.5955217.

[9] A. Milton and S. Tamil Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," *Computer Speech and Language,* pp. 727–742, 2014, http://dx.doi.org/10.1016/j.csl.2013.08.004.

[10] G. Tamulevicius and T. Liogiene, "Low-order multi-level features for speech emotion recognition," *Baltic Journal of Modern Computing,* pp. 234–247, 2015.

[11] 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/.

[12] 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, http://dx.doi.org/10.1109/ACII.2009.5349350.