COMBINING CLASSIFIERS WITH DIVERSE FEATURE SETS FOR ROBUST SPEAKER INDEPENDENT EMOTION RECOGNITION

Marko Lugger, Marie-Elise Janoir, Bin Yang

Chair of system theory and signal processing, University of Stuttgart
Pfaffenwaldring 47, 70550, Stuttgart, Germany
phone: +49 711 68567355, fax: +49 711 68567322, email: marko.lugger@lss.uni-stuttgart.de

ABSTRACT

In this paper, we consider two ways of combining classifiers for speaker independent emotion recognition: serial and parallel combination. In contrast to methods like bagging or boosting, our combination is based on different feature sets, having maximum diversity, instead of different training pattern sets. For that purpose, ensemble feature selection methods are presented for both combination types. For the parallel combination, we propose a novel method that has, to our knowledge, never been considered in the literature. The evaluation is performed on a well-known German emotional database [1]. Both new methods outperform the single stage and the hierarchical classifier presented in [2],[3] on the same database. Moreover, we examine the generalization capability of these classifiers when their feature subsets are not optimized directly on the test set. Here, the parallel combination proved to have the best generalization capability among all studied methods with a benefit of about 10%.

1. INTRODUCTION

Feature selection methods for a single classifier have to cope with the well-known problem of the "curse of dimensionality": When the available feature set is too large and the size of the training set is limited, keeping all features for classification will give a poor performance. An optimal feature subset has thus to be selected and many features have to be discarded in the selection process, although they might contain useful information. By combining classifiers with different feature subsets, more information contained in the complete feature set can be recovered and the recognition performance can be improved.

Three main methods of combining classifiers have been discussed in the literature: hierarchical, serial, and parallel combination. Hierarchical combination starts with a coarse and ends up with a fine classification. Classifiers are ordered in a tree structure and the classification process becomes more precise at each node of the tree. This method has already been proved to be successful in emotion recognition [3]. In serial combination, classifiers are ordered in a queue. Each classifier recognizes only a subset of the received patterns it is capable of classifying with a high detection rate, filters it out, and passes the remaining patterns to the next classifier, hoping it is competent to classify them. The last combination method is parallel combination. Here, each classifier classifies all patterns independently from the other classifiers. The final decision is a fusion of all single decisions which is done, for instance, by majority voting.

Up to now, combination methods have not often been considered in the application of acoustic emotion recognition. Research rather concentrated on finding complex single stage classifiers. In [4], the authors used a single support vector machine to achieve a recognition rate of 86.7%. [3] showed that similar recognition rates could also be obtained on the same database by combining Bayesian classifiers in a hierarchical structure. In this paper, we investigate two other methods of combining multiple classifiers which differ only in their feature subsets. Both serial and parallel combination outperform the hierarchical and the single stage classifier.

Moreover, we also study the issue of overfitting in the feature selection process and how to design robust combination methods. Up to now, we used feature selection in order to reduce the number of features for classification. Unfortunately, the feature selection used the same validation patterns as for later testing [3],[4]. Some theoretical papers [5], [6] on feature selection have warned against that method. There is a risk of finding classifiers which achieve remarkable recognition rates on the specific evaluation set over which the feature selection has been optimized, while having a poor generalization on other test patterns. We avoid this by dividing the complete pattern set into 3 parts, one for training (training set), one for feature selection (validation set), and one for testing (test set).

The paper is organized as follows: After we specify the different feature groups used in this study in section 2, we briefly introduce the 3-stage hierarchical combination in section 3. Sections 4 and 5 present our serial and parallel combination design methods, respectively. In sections 6, the simulation setup is explained and our methods are evaluated and compared with a single stage classifier and the hierarchical combination. Finally, we compare the recognition performance of 3 different base classifiers, the Bayesian, a second order polynomial and an artificial neural network classifier.

2. FEATURE GROUPS

In the field of emotion recognition, mainly suprasegmental prosodic features are used. Sometimes segmental spectral parameters as mel frequency ceptral coefficients (MFCC) are added. In our approach, the common prosodic features are combined with a set of so called voice quality parameters (VQP).

2.1 Prosodic features

There are three main classes of prosodic features: pitch, energy, and duration. Two more classes that do not belong directly to prosody are articulation (formants and bandwidths) and zero crossing rate. The individual features are obtained by measuring statistical values of the corresponding extracted contours. Mean, median, minimum, maximum, range, and variance are the most used measurements. All together we extracted 201 prosodic features from the speech signal.

2.2 Mel frequency cepstral coefficients

The cepstrum of a signal is the inverse Fourier transform of the logarithm of the Fourier transform. In contrast to the standard cepstrum, MFCC uses frequency bands which are positioned logarithmically based on the mel scale motivated by the human auditory system. MFCC is a standard spectral parameter set in automatic speech recognition. For this study the mean values as well as the 2nd to the 5th central moments of 13 MFCC are calculated. The total number of MFCC features is thus 65. The implementation we use was first published in [7].

2.3 Voice quality parameters

In contrast to other spectral feature sets, the voice quality parameters describe the properties of the glottal excitation. Phonation is one important process besides articulation and prosody in generating emotional coloured speech. By inverse filtering, the influence of the vocal tract is compensated to a great extent. The feature set we use is a parameterization of the voice quality in the frequency domain by spectral gradients, see Figure 1. The detailed computation of the basic speech features, the vocal tract compensation, and the voice quality parameters is given in [3]. All together we extracted 67 voice quality parameters.

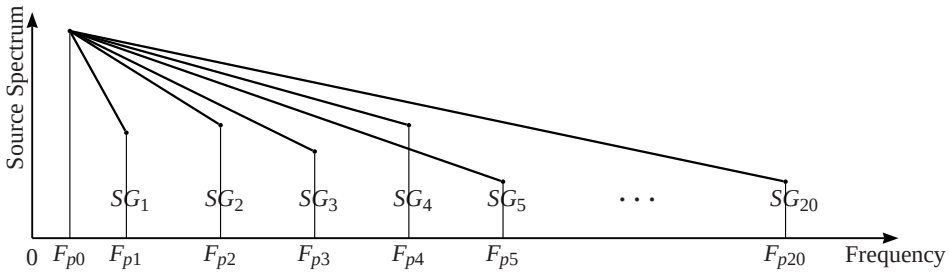

Figure 1: Spectral gradients at fixed frequencies in the glottal excitation spectrum

2.4 Feature selection

There are two main reasons for reducing the number of features from the original set. First, the number of training patterns had to be enormous if we want to use all features. Second, the feature extraction and the training would take a long time when using the whole feature set. So for all classifications, the original number of 333 features is reduced by using an iterative selection algorithm. After the selection process, the final feature number is reduced to 25 because for this feature number a local maximum in the classification rate was observed. We used the sequential floating forward selection algorithm (SFFS). It is an iterative method to find a subset of features that is near the optimal one. It was first proposed in [8]. In each iteration, a new feature is added to the subset of selected features and afterwards the conditionally least significant features are removed. As selection criterion, the speaker independent recognition rate is used. This process is repeated until the final dimension is obtained.

3. HIERARCHICAL COMBINATION OF CLASSIFIERS

Motivated by the psychological emotion model [3], we found out that one can improve the emotion classification performance by using multiple Bayesian classifiers. In the hierarchical combination, we perform 5 binary classifications in 3 stages as shown in Figure 2. Every frame corresponds to one Bayesian subclassification whose best 25 features are optimized by SFFS separately. In the first

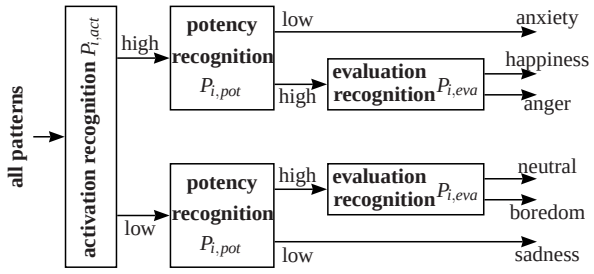

Figure 2: Design of a 3-stage hierarchical combination classifier

stage, we classify two different activation levels. One class including anger, happiness, and anxiety has a high activation level while the second class including neutral, boredom, and sadness has a low activation level. In the second stage, we classify two potency levels within each activation class. That means, all patterns that were classified to high activation in the first stage are classified to one class containing happiness and anger or to a second class only containing anxiety. Similarly, all patterns that were classified to low activation in the first stage are classified to one class containing neutral and boredom or to sadness. In the third stage, we distinguish between the emotions that only differ in the evaluation dimension: happiness vs. anger as well as neutral vs. boredom.

In hierarchical combination, a pattern is correctly classified, only if every single subclassification is correct. That means, the recognition rate P_i is the product of all subrecognition rates $P_{i,k}$, where i is the class index and k the label of the classification stage.

$$P_i = P_{i,act} \cdot P_{i,pot} \cdot P_{i,eva} \tag{1}$$

4. SERIAL COMBINATION OF CLASSIFIERS

4.1 Binary classifiers for serial combination

In the literature, it is quite usual to use a cascade of asymmetrical binary classifiers whose costs of misclassification for the two classes are not equal [9]: There should be almost no false alarm detections assigning a wrong pattern to one class c_i, i.e. $P_{i,FA} \approx 0$. On the other hand, some missing detections rejecting a correct pattern of class c_i are tolerated, see Figure 3.

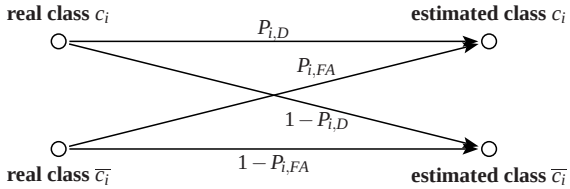

Figure 3: Model of a binary classification problem

4.2 Design of the serial classifier experts

According to [10], our algorithm applies the serial combination to multi-class problems: each classifier of the cascade is responsible for recognizing only one part of the patterns for one specific class, the so called target class. For these patterns, the classifier is called expert and filters them out. In general, there are several experts $C_{i,j},\ j = 1,\cdots,E_i$ for the same target class c_i containing N_i patterns. Each expert is characterized by its target class, the detection rate of this class $P_{i,j}$, and the number of patterns which have been filtered out $n_{i,j}$, see Figure 4. The different classes $c_i,\ i = 1,\cdots,k$ are filtered out iteratively, until almost all patterns have been classified. It is crucial that every expert has to have a very low false alarm rate $P_{i,FA}$, as errors at one stage can never be recovered at a later stage. On the other hand, we wish that each classifier filters out as many patterns as possible in order to reduce the length of the cascade. At the end of the cascade, a multi-class default classifier C_{def} with recognition rate $P_{i,def}$ is employed to classify the $n_{i,def} = N_i - \sum_{j=1}^{E_i} n_{i,j}$ remaining patterns for which no expert could be found. Since in serial combination a pattern is classified and filtered out by exactly one classifier of the cascade, the recognition rate P_i is the weighted sum of the recognition rates of the experts for this target class and the default classifier:

$$P_i = \frac{1}{N_i}\left[\sum_{j=1}^{E_i}\left(n_{i,j}\cdot P_{ij}\right) + n_{i,def}\cdot P_{i,def}\right] \tag{2}$$

The major problem in designing cascaded combinations is building "asymmetrical" classifiers which have a very low false alarm rate, without being too bad in detecting patterns of the target class c_i. To achieve that, we use a modified version of SFFS. In that version, the algorithm chooses the features in order to maximize the detection rate of the target class $P_{i,j} = P_{i,D}$, under the constraint that the false alarm rate $P_{i,FA}$ is below a given threshold θ. θ must be chosen such that a good trade-off is reached between the detection rate $P_{i,D}$ and the false alarm rate $P_{i,FA}$. The lower θ, the higher the

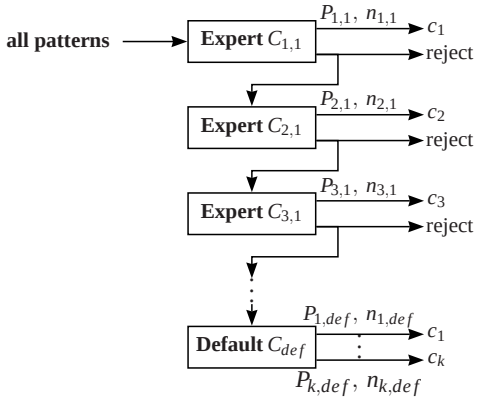

Figure 4: Design of a serial combination classifier

selectivity of the corresponding classifier (fewer false alarm detections), but the lower its sensitivity (fewer patterns filtered out). The whole feature selection algorithm contains the following steps:

a) Features for the first expert are selected due to the criteria: $\max P_{i,D}$ subject to $P_{i,FA} < \theta$. The selection process for the first expert ends, when the desired number of features is reached. The patterns this classifier is expert of are filtered out. The remaining patterns are rejected and sent to the next expert.

b) Based only on the rejected patterns, an expert for the next target class is sought. The selection is performed consecutively for all the target classes several times. If for one target class the constraint cannot be fulfilled, the current expert is cancelled and we proceed directly with selecting features for an expert of the next target class. An expert for the current class can be found a lot easier in the next iteration, when some patterns of the other classes have already been filtered out.

c) The whole design algorithm ends when the maximum number of experts E_i for each class in reached, in order to avoid adding too many classifiers which filter out only a few patterns.

5. PARALLEL COMBINATION OF CLASSIFIERS

5.1 Overview of parallel combination methods

It is well known that parallel combination is only efficient when the pool of classifiers is very diverse and negatively correlated [11]. The problem of building a good parallel combination amounts to finding an ensemble of single classifiers that are quite good and produce their classification errors on different patterns. In contrast to bagging or boosting, where the diversity is based on different training patterns, the diversity can also be achieved by choosing different feature subsets for each member of the ensemble. The problem of finding feature subsets that give the optimal recognition rate for the parallel combination has been called "ensemble feature selection" and has received much attention in theoretical pattern recognition literature [12], [13], [14], but not in emotion recognition.

In parallel combination, a pattern is correctly classified if the majority of the N subclassifiers are correct. If we assume an odd number of statistical independent classifiers, and the recognition rate for every subclassifier $P_{i,k} = p$ is constant, then the overall recognition rate P_i for class c_i can be calculated using the binomial formula. If $p > 0.5$, P_i is monotonically increasing with the number of subclassifiers N [15].

$$P_i = \sum_{m=\frac{N+1}{2}}^{N} \binom{N}{m} p^m (1-p)^{N-m} \tag{3}$$

5.2 Design of the parallel classifier ensemble

To select the feature subsets for our ensemble, we use an approach which has, to our knowledge, never been considered before. Subclassifiers are added one by one to the ensemble. For each new subclassifier which is added to the ensemble, features are selected by using SFFS. But instead of choosing the features which perform best on the whole validation set, we choose those which perform best on the reduced subset of the validation set which contains only those patterns difficult to classify by the ensemble of all subclassifiers up to the current stage. This subset alters as new classifiers are added to the ensemble and it is recomputed at each stage. It will be called "difficult validation set" in the following. The design process is depicted in Figure 5 and the feature selection algorithm consists of following steps:

a) We initialize the process by selecting features for the first subclassifier C_1. Those features are optimized to give the best recognition rate over the whole validation set, which is congruent with the "difficult validation subset". Note that the first classifier in the parallel combination is equal to the overall best single classifier which is used for the single stage method.

b) At each new stage k, we define the "difficult validation subset" as the subset of all patterns of the validation set which have not been correctly classified by the ensemble of k subclassifiers up to now. Except for the initialization of the first subclassifier C_1, this is done by counting the votes of all subclassifiers. A pattern belongs to the "difficult validation subset" if the right class has at most one vote more than the wrong class which has been most voted for. Features for the next subclassifier C_{k+1} are optimized only on the current "difficult validation subset".

c) The algorithm ends up when the size of the "difficult validation subset" can not be further reduced by adding more subclassifiers to the ensemble.

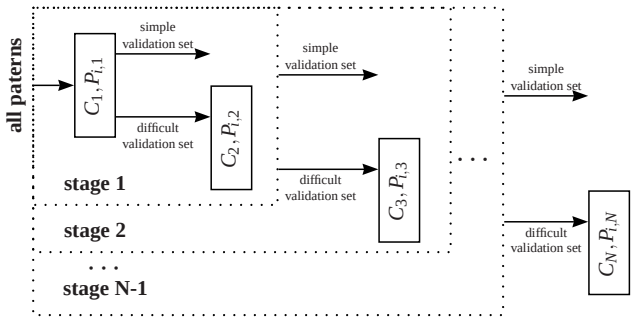

Figure 5: Design of a parallel combination classifier

Note that the underlying principle of this algorithm has much in common with boosting [16]. But in boosting, classifiers use different training subsets, which contain increasingly difficult patterns. Here, we train all classifiers on the whole training set and choose different feature subsets that perform well on increasingly difficult patterns of the validation set.

6. SIMULATION RESULTS

6.1 Database and evaluation

All simulations were performed on the Berlin emotional database [1] using six emotions: happiness, boredom, neutral, sadness, anger, and anxiety. The 690 short utterances are spoken by 10 actors. Every pattern is of 2-5 seconds length.

Two methods are used to evaluate our classifiers, see Figure 6. These two methods significantly differ in the patterns that were used for feature selection. First, a leave-one-speaker-out cross-validation is performed which was used in previous works on the same database. By one cross validation loop over all 10 speakers, the training is performed by using 9 speakers and the 10th speaker is used for both the feature selection and the testing. It will be called "evaluation method with optimized feature set". Here, the feature selection is optimized to the test set implying an overfitting. If the selected features are applied to some other test patterns which have not been used in the feature selection process, we expect a significant performance drop indicating a poor generalization capability.

Second, a leave-one-speaker-out cross-validation with inner and outer loop is used as presented in [17]. It will be called "evaluation method with realistic feature set". In this case, 8 speakers are used for training, one speaker for feature selection (validation set), and one speaker for testing. The training and feature selection

are made in an inner cross-validation over 9 speakers, and recognition rates are computed on the remaining 10th speaker (test set). So, the feature set is not optimized on the test data because the 10th speaker was neither used in the training nor in the feature selection process. By calculating the weighted overall recognition rate over all 10 speakers, we obtain a quite good measure for the generalization capability and consequently for the robustness of a classification method.

evaluation with optimized feature set

training set	validation set

evaluation with realistic feature set

training set	validation set	test set

Figure 6: Two methods for speaker independent evaluation of the classifiers with "leave-one-speaker-out cross-validation"

6.2 Single stage classifier and hierarchical combination

By using the "evaluation with optimized feature set", the hierarchical combination clearly outperforms the single Bayesian classifier as stated in [3], see Table 1. For the "evaluation method with realistic feature set", the recognition rate of the single Bayesian classifier and for the hierarchical combination decreases considerably to 58.6% respectively 59.7%, see Table 2. With this evaluation method, the hierarchical combination is only slightly better than the single Bayesian classifier. For both methods, happiness and neutral are particularly badly recognized as can be seen in Table 3.

6.3 Serial combination

The results in the optimized case are presented in Table 1. The serial combination has the highest overall recognition rate with 96.5%. The recognition rate and the percentage of classified patterns in each stage are depicted in Figure 7. Only 4% of all patterns have to be classified by the default classifier. Most of the patterns are classified in early stages. The feature selection algorithm failed to find an expert fulfilling the constraint in the first stage, so this emotion had to be skipped at the beginning. This is not surprising, as it has already been noticed in [3] that happiness is very difficult to separate from other emotions. However, good detectors for happiness could be found at later stages, so happy patterns could be filtered out more efficiently at the end. A basic assumption of serial combination proves to be true: difficult patterns become much easier to classify once the simplest patterns have been filtered out.

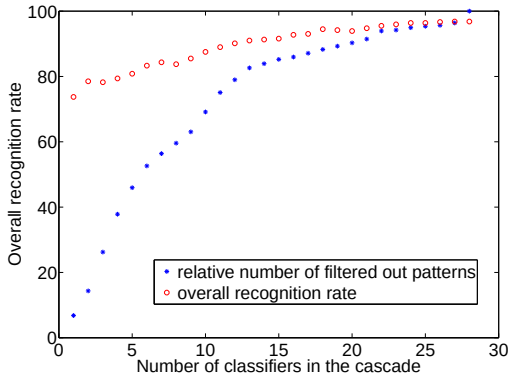

Figure 7: Overall recognition rate and number of classified patterns using serial combination and optimized feature set

When the feature subsets are realistically optimized, the average recognition rate is not much higher than that obtained with a single Bayesian classifier, see Table 2. That means, this method is prone to overfitting and does not have a good generalization capability. Interestingly, speakers who are difficult to recognize (3, 4, and 10) are improved, whereas easy speakers (5, 8, and 9) have a slightly lower recognition rate. So the range in recognition rate between the best ant the worst speaker has significantly decreased from 30% to 19%. The same phenomenon can be observed for the recognition rate of the different emotions, see Table 3. The recognition rates of difficult emotions (happiness and neutral) are improved compared to the single Bayesian classifier, whereas the recognition rates of "easy" emotions remain almost unchanged.

method	single stage	hierarchical	serial	parallel
rec. rate	74.6%	88.8%	96.5%	92.6%

Table 1: Overall recognition rates for a Bayesian classifier and optimized feature set

6.4 Parallel combination

In the optimized feature set case, the parallel combination performs slightly better than the hierarchical combination, but worse than the serial combination, see Table 1. Here, the overall recognition rate is given for an ensemble of 25 classifiers. The evolution of the overall recognition rate up to 25 classifiers is depicted in Figure 8. However, it is almost constant when the number of classifiers is higher than 20, as the size of the "difficult classification subset" does not vary much after that number of classifiers.

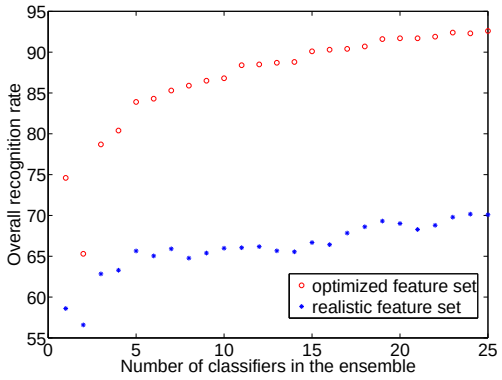

Figure 8: Overall recognition rate as a function of the number of classifiers using parallel combination

For the ensemble in stage 1, the two classifiers have been chosen to maximize the amount of patterns over which they disagree. Thus, the majority voting gives very poor results. However, the ensemble with 3 classifiers already performs better than the single Bayesian classifier and the overall recognition rate further increases after that stage. A similar increase can also be observed when using the realistic feature subset. The overall performance with 25 classifiers is 70.1%, see Table 2. This is by far the highest recognition rate in the realistic case. It is improved by 11.5% compared to the single Bayesian classifier and 10.4% compared to the hierarchical combination. So the parallel combination has a good generalization capability. It performs better than the other combination methods on all speakers, see Table 2, and all emotions, see Table 3. In particular, the performance of the emotions difficult to classify (happy and neutral) is improved by at least 15% compared to the single stage Bayesian classifier.

A possible explanation for the good generalization performance of that combination method could be found in [18]. By optimizing the feature subsets of each single classifier on a small part of the training set in cross-validation, we create overfitted classifiers which have a very high variance and a very poor generalization capability. However, the bad quality of single classifiers is compensated for by diversity gain, and the high variance of ensemble members is diminished by aggregation. In particular, the majority voting process lowers the influence of a few very bad members.

method / speaker	1	2	3	4	5	6	7	8	9	10	overall
single stage	61.6%	59.4%	50.7%	50.8%	68.7%	60.7%	60.9%	64.5%	69.4%	39.4%	**58.6%**
hierarchical comb.	56.1%	75.0%	56.5%	56.7%	67.2%	53.2%	62.3%	64.5%	59.7%	47.9%	**59.7%**
serial comb.	69.9%	60.9%	55.1%	59.7%	67.2%	62.3%	62.3%	63.2%	66.7%	50.7%	**61.8%**
parallel comb.	76.7%	78.1%	68.1%	62.7%	73.1%	62.3%	72.5%	80.3%	72.2%	54.9%	**70.1%**

Table 2: Recognition rates for realistic feature selection averaged over all emotions

emotion	happy	bored	neutral	sad	angry	anxious	overall
single stage	36.5%	65.8%	46.1%	70.8%	65.4%	62.8%	**58.6%**
hierarchical	35.5%	73.9%	48.5%	62.5%	71.3%	62.0%	**59.7%**
serial comb.	41.1%	67.6%	55.9%	71.7%	67.7%	63.7%	**61.8%**
parallel comb.	51.9%	73.9%	64.7%	86.0%	77.9%	63.7%	**70.1%**

Table 3: Recognition rates for realistic feature selection averaged over all speakers

6.5 Comparison of different base classifiers

Until now, all experiments were performed using simple Bayesian classifiers. Since the parallel combination showed the best results, we compare different base classifier in parallel combination. As we can see from Table 4, the parallel combination of other base classifiers than the Bayesian does not result in significantly different results. For all 3 applied base classifiers, the parallel combination improved the classification rate of the corresponding single stage classifier by more than 5%. The highest gain was achieved for the Bayesian classifier with an improvement of 11.5%. The best absolute result, with an overall recognition rate of 73.0%, was obtained by using a parallel combination of the neural network classifier, with 6 nodes in the hidden layer.

method	Bayesian	polynomial	neural net
single stage	58.6%	61.9%	62.2%
parallel comb.	70.1%	67.2%	73.0%

Table 4: Comparison of different base classifiers using parallel combination and realistic feature set

7. CONCLUSION

In this paper, we presented two ways of combining classifiers for emotion recognition: a serial and a parallel combination. The latter uses an ensemble feature selection process that has, to our knowledge, never been considered in the literature. We evaluated the results when feature sets are optimized on the test set or on a separate validation set which is disjoint to the test set. The first method suffers from overfitting in the feature selection process but it shows the theoretical performance of the methods if we assume knowledge about the test set. The latter was recommended in [6]. It is more relevant for practical applications as it gives a measure of the generalization capability of the classification method. The serial combination proved to be the best method when optimized on the evaluation set; however, it does not have a good generalization capability, because it can not afford to have a single bad member in the cascade. The parallel combination has by far the best generalization capability and outperforms the other presented methods with an improvement in the recognition rate of about 10%. It achieved the best recognition rates on all speakers and all emotions. Concluding we can say, the diversity of the subclassifier, that is accomplished by different feature sets, lead to a very robust ensemble classifier.

REFERENCES

[1] Burkhardt et. al., "A database of German emotional speech," *Proceedings of Interspeech*, 2005.

[2] M. Lugger and B. Yang, "The relevance of voice quality features in speaker independent emotion recognition," in *International Conference on Audio, Speech, and Signal processing, Honolulu, USA*, 2007.

[3] M. Lugger and B. Yang, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *International Conference on Acoustics, Speech, and Signal Processing, Las Vegas*, 2008.

[4] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotional recognition in the noise applying large acoustic feature sets," in *Speech Prosody, Dresden*, 2006.

[5] R. Kohavi and G. John, "Wrappers for feature selection," *Artificial Intelligence Journal, Special Issue on Relevance*, vol. 97, pp. 273–324, 1997.

[6] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *Journal of Machine Learning Research*, 2003.

[7] Orsak et. al., "Collaborative SP education using the internet and matlab," *IEEE Signal processing magazine*, vol. 12, no. 6, pp. 23–32, 1995.

[8] P. Pudil, F. Ferri, J. Novovicova, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," in *12th IAPR International Conference on Pattern Recognition*, 1994.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Vision and Pattern Recognition Conference*, 2001.

[10] Y. Saatci and C. Town, "Cascaded classification of gender and facial expression using active appearance models," in *7th International Conference on Automatic Face and Gesture Recognition*, 2006.

[11] L.I Kuncheva, C.J. Whitaker, C.A Shipp, and R.P.W Duin, "Is independance good for combining classifiers?," in *15th International Conference on Pattern Recognition*, 2000.

[12] A. Tsymbal, M. Pechenizky, and P. Cunningham, "Sequential genetic search for ensemble feature selection," Tech. Rep., Dep. of Computer Science, Trinity College Dublin, 2005.

[13] M. Skurichina and R.P.W Duin, "Combining feature subsets in feature selection," in *International Workshop on Multiple Classifier Systems*, 2005.

[14] J.J Rodriguez, L.I Kuncheva, and C.J Alonso, "Rotation forest: a new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[15] L. Lam and C.Y. Suen, "Application of majority voting to pattern recognition: an analysis of its behaviour and performance," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 27, pp. 553–568, 1997.

[16] L. Breiman, "Bias, variance, and arcing classifiers," Tech. Rep., Statistics Department, University of California, 1997.

[17] J. Loughrey and P. Cunningham, "Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets," Tech. Rep., Department of Computer Science, Trinity College Dublin, 2005.

[18] P. Cunningham, "Overfitting and diversity in classification ensembles based on feature selection," Tech. Rep., Department of Computer Science, Trinity College Dublin, 2000.