RANSAC-Based Training Data Selection on Spectral Features for Emotion Recognition from Spontaneous Speech

Elif Bozkurt[1], Engin Erzin[1], Çiğdem Eroğlu Erdem[2], and A. Tanju Erdem[3]

[1] Multimedia, Vision and Graphics Laboratory,
College of Engineering, Koç University, 34450, Sariyer, Istanbul, Turkey
{ebozkurt,eerzin}@ku.edu.tr
[2] Department of Electrical and Electronics Engineering,
Bahçeşehir University, 34349 Beşiktaş, Istanbul, Turkey
cigdem.eroglu@bahcesehir.edu.tr
[3] Department of Electrical and Electronics Engineering,
Özyeğin University, 34662 Üsküdar, Istanbul, Turkey
tanju.erdem@ozyegin.edu.tr

Abstract. Training datasets containing spontaneous emotional speech are often imperfect due the ambiguities and difficulties of labeling such data by human observers. In this paper, we present a Random Sampling Consensus (RANSAC) based training approach for the problem of emotion recognition from spontaneous speech recordings. Our motivation is to insert a data cleaning process to the training phase of the Hidden Markov Models (HMMs) for the purpose of removing some suspicious instances of labels that may exist in the training dataset. Our experiments using HMMs with Mel Frequency Cepstral Coefficients (MFCC) and Line Spectral Frequency (LSF) features indicate that utilization of RANSAC in the training phase provides an improvement in the unweighted recall rates on the test set. Experimental studies performed over the FAU Aibo Emotion Corpus demonstrate that decision fusion configurations with LSF and MFCC based classifiers provide further significant performance improvements.

Keywords: Affect recognition, emotional speech classification, RANSAC, data cleaning, decision fusion.

1 Introduction

For supervised pattern recognition problems such as emotion recognition from spontaneous speech, large training sets need to be recorded and labeled to be used for the training of the classifier. The labeling of large training datasets is a tedious job, carried out by humans and hence prone to human mistakes. The mislabeled (or noisy) examples of the training data may result in a decrease in the classifier performance. It is not easy to identify these contaminations or imperfections of the training data since they may also be hard to learn examples.

A. Esposito et al. (Eds.): Communication and Enactment 2010, LNCS 6800, pp. 36–47, 2011.
Springer-Verlag Berlin Heidelberg 2011

In that respect, pointing out troublesome examples is a chicken-and-egg problem, since good classifiers are needed to tell which examples are noisy [1].

Spectral features play an important role in emotion recognition. The dynamics of the vocal tract can potentially change under different emotional states. Hence spectral characteristics of speech differ for various emotions [14]. The utterance level statistics of the spectral features have been widely used in speech emotion recognition and demonstrated a considerable success [13] [12].

In this work, we assume that outliers in the training set of emotional speech recordings mainly result from mislabeled or ambiguous data. Our goal is to remove such noisy samples from the training set to increase the performance of Hidden Markov Model based classifiers modeling spectral features.

1.1 Previous Work

Previous research on data cleaning, which is also called as data pruning or decontamination of training data shows that removing noisy samples is worthwhile [1] [2] [3]. Guyon et al. [9] have studied data cleaning in the context of discovering informative patterns in large databases. They mention that informative patterns are often intermixed with unwanted outliers, which are errors introduced non-intentionally to the database. Informative patterns correspond to atypical or ambiguous data and are pointed out as the most "surprising" ones. On the other hand, garbage patterns are also surprising, which correspond to meaningless or mislabeled patterns. The authors point out that automatically cleaning the data by eliminating patterns with suspiciously large information gain may result in loss of valuable informative patterns. Therefore they propose a user-interactive method for cleaning a database of hand-written images, where a human operator checks those patterns that have the largest information gain and therefore the most suspicious.

Batandela and Gasca [2] report a cleaning process to remove suspicious instances of the training set or correcting the class labels and keep them in the training set. Their method is based on the Nearest Neighbor classifier. Wang et al. [22], present a method to sample a large and noisy multimedia data. Their method is based on a simple distance measure that compares the histograms of the sample set and the whole set in order to assess the representativeness of the sample set. The proposed method deals with noise in an elegant way, and has been shown to be superior to the simple random sample (SRS) method [8][16].

Angelova et al. [1] present a fully automatic algorithm for data pruning, and demonstrate its success for the problem of face recognition. They show that data pruning can improve the generalization performance of classifiers. Their algorithm has two components: the first component consists of multiple semi-independent classifiers learned on the input data, where each classifier concen-

machine for identifying examples which are in contradiction with most learners and therefore noisy.

There are also other approaches for learning with noisy data based on regularization [17] or averaging decisions of several functions such as bagging [4]. However, these methods are not successful in high-noise cases.

1.2 Contribution and Outline of the Paper

In this paper, we propose an algorithm for automatic noise elimination from training data using Random Sample Consensus. RANSAC is a paradigm for fitting a model to noisy data and utilized in many computer vision problems [21]. RANSAC performs multiple trials of selecting small subsets of the data to estimate the model. The final solution is the model with maximal support from the training data. The method is robust to considerable noise. In this paper, we adopt RANSAC for training HMMs for the purpose of emotion recognition from spontaneous emotional speech. To the best of our knowledge, RANSAC has not been used before for cleaning an emotional speech database.

The outline of the paper is as follows. In Section 2, background information is provided describing the spontaneous speech corpus and the well known RANSAC algorithm. In Section 3, the proposed method is described including the speech features, the Hidden Markov Model, the RANSAC-based HMM fitting approach and the decision fusion method. In Section 4, our experimental results are provided, which is followed by conclusions and future work given in Section 5.

2 Background

2.1 The Spontaneous Speech Corpus

The FAU AIBO corpus is used in this study [19]. The corpus consists of spontaneous, German and emotionally colored recordings of children interacting with Sony's pet robot Aibo. The data was collected from 51 children and consisted of 48,401 words. Each word was annotated independently from each other as neutral or as belonging to one of the ten other classes, which are named as: joyful (101 words), surprised (0), emphatic (2,528), helpless (3), touchy (i.e., irritated) (225), angry (84), motherese (1,260), bored (11), reprimanding (310), rest (i.e., non-neutral but not belonging to the other categories) (3), neutral (39,169), and there were also 4,707 words not annotated since they did not satisfy the majority vote rule used in the labeling procedure. Five labelers were involved in the annotation process, and a majority vote approach was used to decide on the final label of a word, i.e., if at least three labelers agreed on a label, the label was attributed to the word. As we can see from the above numbers, in 4,707 of the words, the five listeners could not agree on a label. Therefore, we can say that labeling spontaneous speech data into emotion classes is not an easy task, since the emotions are not classified easily and may even contain a mixture of more than one emotion. This implies that the labels of the training may be imperfect,

ion performance of the trained pattern classifiers.

In the INTERSPEECH 2009 emotion challenge, the FAU AIBO dataset was segmented into manually defined chucks consisting of one or more words, since that was found to be the best unit of analysis [19], [20]. A total of 18,216 chunks was used for the challenge and the emotions were grouped into five classes, namely: Anger (including angry, touchy, and reprimanding classes) (1,492), Emphatic (3,601), Neutral (10,967), Positive (including motherese and joyful) (889), and Rest (1,267). The data is highly unbalanced. Since the data was collected at two different schools, speaker independence is guaranteed by using the data of one school for training and the data of the other school for testing. This dataset is used in the experiments of this study.

2.2 The RANSAC Algorithm

Random Sample Consensus is a method for fitting a model to noisy data [7]. RANSAC is capable of being robust to error levels of significant percentages. The main idea is to identify the outliers as data samples with greatest residuals with respect to the fitted model. These can be excluded and the model is re-computed. The steps of the general RANSAC algorithm are as follows [21] [7]:

1. Suppose we have n training data samples $X = x_1, x_2, ..., x_n$ to which we hope to fit a model determined by (at least) m samples ($m \leq n$).
2. Set an iteration counter $k = 1$.
3. Choose at random m items from X and compute a model.
4. For some tolerance , determine how many elements of X are within of the derived model . If this number exceeds a threshold t, re-compute the model over this consensus set and stop.
5. Set $k = k + 1$ If , $k < K$ for some predetermined K, go to 3. Otherwise, accept the model with the biggest consensus set so far, or fail.

There are possible improvements to this algorithm [21] [7]. The random subset selection may be improved if we have prior knowledge of data and its properties, that is some samples may be more likely to fit a correct model than others.

There are three parameters that need to be chosen:

- , which is the acceptable deviation from a good model. It might be empirically determined by fitting a model to m points, measuring the deviations and setting to some number of standard deviations above the mean error.
- t, which is the size of the consensus set. There are two purposes for this parameter: to represent enough sample points for a sufficient model and to represent the enough number of samples to refine the model to the final best estimate. For the first point a value of t satisfying $t - m > 5$ has been suggested [7].
- K, which is the maximum number to run the algorithm while searching a satisfactory fit. Values of $K = 2^{-m}$ or $K = 3\omega^{-m}$ have been argued to be reasonable choices [7], where ω is the probability of a randomly selected sample to be within of the model.

3 RANSAC-Based Data Cleaning Method

3.1 Extraction of the Speech Features

We represent spectral features of speech using mel-frequency cepstral coefficients (MFCC) and line spectral frequencies (LSF) with their first and second order derivatives.

MFCC features. Spectral features, such as mel-frequency cepstral coefficients (MFCC), are expected to model the varying nature of speech spectra under different emotions. We represent the spectral features of each analysis window of the speech data with a 13-dimensional MFCC vector consisting of energy and 12 cepstral coefficients, which will be denoted as $\mathbf{f}_C$.

LSF features. Line spectral frequency (LSF) decomposition has been first developed by Itakura [10] for robust representation of the coefficients of linear predictive (LP) speech models. LP analysis of speech assumes that a short stationary segment of speech can be represented by a linear time invariant all pole filter of the form $H(z) = \frac{1}{A(z)}$, which is a p^{th} order model for the vocal tract.

LSF decomposition refers to expressing the p-th order inverse filter $A(z)$ in terms of two polynomials $P(z) = A(z) - z^{p+1}A(z^{-1})$ and $Q(z) = A(z) + z^{p+1}A(z^{-1})$, which are used to represent the LP filter as,

$$H(z) = \frac{1}{A(z)} = \frac{2}{P(z) + Q(z)}. \tag{1}$$

The polynomials $P(z)$ and $Q(z)$ each have $p/2$ zeros on the unit circle, where phases of the zeros are interleaved in the interval $[0,\ \pi]$. Phases of p zeros from the $P(z)$ and $Q(z)$ polynomials form the LSF feature representation for the LP model. Extraction of LSF features, which is finding p zeros of $P(z)$ and $Q(z)$

Note that the formant frequencies correspond to the zeros of $A(z)$. Hence, $P(z)$ and $Q(z)$ will be close to zero at each formant frequency, which implies that the neighboring LSF features will be close to each other around formant frequencies. This property relates the LSF features to the formant frequencies [15], and makes them good candidates to model emotion related prosodic information in the speech spectra. We represent the LSF feature vector of each analysis window of speech as a p dimensional vector $\mathbf{f}_L$.

Dynamic features. Temporal changes in the spectra play an important role in human perception of speech. One way to capture this information is to use dynamic features, which measure the change in the short-term spectra over time. We compute the first and second time derivatives of the thirteen dimensional MFCC features using the following regression formula:

$$\Delta \mathbf{f}_C[n] = \frac{\sum_{k=-2}^{2} k\mathbf{f}_C[n+k]}{\sum_{k=-2}^{2} k^2}, \tag{2}$$

where $\mathbf{f}_C[n]$ is the MFCC feature vector at time frame n. Then, the extended MFCC feature vector, including the first and second order derivative features, is represented as $\mathbf{f}_{C\Delta} = \left[\mathbf{f}_C^T\ \Delta\mathbf{f}_C^T\ \Delta\Delta\mathbf{f}_C^T\right]^T$, where ^{T} is the vector transpose operator. Likewise, the extended LSF feature vector including dynamic components is denoted as $\mathbf{f}_{L\Delta}$.

3.2 Emotion Classification Using Hidden Markov Models

Hidden Markov model has been deployed with great success in automatic speech recognition to model temporal spectral information, and they were also used similarly for emotion recognition as well [18]. We model the temporal patterns of the emotional speech utterances using HMM. We target to make a decision for syntactically meaningful chunks of speech segments, where in each segment typically a single emotional evidence is expected. Furthermore, in each speech segment emotional evidence may exhibit temporal patterns. Hence, we employ N states left-to-right HMM to model each emotion class. Feature observation probability distributions are modeled by M mixture Gaussian density functions with diagonal covariance matrix. Structural parameters N and M are determined through a model selection method and discussed under experimental studies.

In the emotion recognition phase, the likelihood of a given speech segment is computed over HMM with the Viterbi decoding for each emotion class. Then, the utterance is classified as expressing the emotion, which yields the highest likelihood score.

3.3 RANSAC-Based Training of HMM Classifiers

Our goal is to train an HMM for each of the five emotion classes in the training set (Anger, Emphatic, Positive, Neutral and Rest). For each emotion class, we want to select a training set such that the fraction of the number of inliers (consensus set) over the total number of utterances in the dataset is maximized. In order to apply the RANSAC algorithm for fitting an HMM model, we need to estimate suitable values for the parameters m, ε , t, K and ω , which were defined in Section 2.2.

For determining the biggest consensus set (inliers) for each of the five emotions, we use a simple HMM structure with single state and 16 Gaussian mixtures per state. The steps of the RANSAC-based HMM training method are as follows:

1. For each of the five emotions suppose we have n training data samples $X = x_1, x_2, ..., x_n$ to which we hope to fit a model determined by (at least) m samples ($m \leq n$). Initially, we randomly select $m = 320$ utterances consid-

process.

2. Set an iteration counter $k = 1$.
3. Choose at random m items from X and compute an HMM with a given number of states and Gaussian mixtures per state. Estimate the normalized likelihood values for the rest of the training set, using the trained HMM.
4. Set tolerance level to ε = (μ - 1.5 *), where mean (μ) and standard deviation ( ) values are calculated using the normalized likelihood values of the initial randomly selected m utterances. Determine how many elements of X are within ε of the derived model. If this number exceeds a threshold t, recompute the model over this consensus set and stop.
5. Increase the iteration counter $k = k + 1$, If $k < K$, and $k < 200$, for some predetermined K, go to step 3. Otherwise, accept the model with the biggest consensus set so far, or fail. Here, we estimate K, the number of loops required for the RANSAC algorithm to converge, using the number of inliers [4]:

$$K = \frac{\ln(1-p)}{\ln(1-\omega^m)} \tag{3}$$

Here we set $\omega = \frac{m_i}{m}$, where m_i is the number of inliers for iteration i and $p = 0.9$ is the probability that at least one of the sets of random samples does not include an outlier.

3.4 Decision Fusion for Classification of Emotions

Decision fusion is used to compensate for possible misclassification errors resulting from a given modality classifier with other available modalities, where scores resulting from each unimodal classification are combined to arrive at a conclusion. Decision fusion is especially effective when contributing modalities are not correlated and resulting partial decisions are statistically independent.

We consider a weighted summation based decision fusion technique to combine different classifiers [6] for emotion recognition. The HMM classifiers with MFCC and LSF features output likelihood scores for each emotion and utterance, which need to be normalized prior to the decision fusion process. First, for each utterance, likelihood scores of both classifiers are mean-removed over emotions. Then, sigmoid normalization is used to map likelihood values to the $[0, 1]$ interval for all utterances [6]. After normalization, we have two likelihood score sets for the HMM classifiers for each emotion and utterance.

Let us denote normalized log-likelihoods of MFCC and LSF based HMM classifiers as $\bar{\rho}_{\gamma_e(C)}$ and $\bar{\rho}_{\gamma_e(L)}$ respectively, for the emotion class e. The decision fusion then reduces to computing a single set of joint log-likelihood ratios, ρ_e, for each emotion class e. Assuming the two classifiers are statistically independent, we fuse the two classifiers, which will be denoted by $\gamma_e(C) \oplus \gamma_e(L)$, by computing the weighted average of the normalized likelihood scores

$$\rho_e = \alpha \bar{\rho}_{\gamma_e(C)} + (1-\alpha)\bar{\rho}_{\gamma_e(L)}, \tag{4}$$

where the parameter is selected in the interval $[0, 1]$ to maximize the recognition rate on the training set.

4 Experimental Results

In this section, we present our experimental results for the 5-class emotion recognition problem using FAU-Aibo speech database provided by the INTERSPEECH 2009 emotion challenge. The distribution of emotional classes in the database is highly unbalanced that the performance is measured as unweighted recall (UA) rate which is the average recall of all classes. In Table 1 and Table 2, we list the UA rates for classifiers modeling MFCC and LSF features with 1-state and 2-state HMMs with number of Gaussian mixtures in the range [8, 160] per state. In the experiments further increasing number of states did not improve our results. We can see that incorporation of a RANSAC based data cleaning procedure yields an increase in the unweighted recall rates in all cases. For the MFCC feature set, the highest improvement (2.84%) is seen for the 1-state HMM with 160 Gaussian mixtures, whereas for the LSF feature set the highest improvement is obtained as 2.73 % for 1-state HMM with 80 Gaussian mixtures.

Table 1. Unweighted recall rates (UA) for 1- and 2- state HMMs modeling MFCC features with and without RANSAC

Number of	1 state		2 states	
mixtures	All-data	RANSAC	All-data	RANSAC
16	38.39	39.51	38.46	38.63
56	38.84	39.79	40.17	40.45
80	38.63	40.62	40.18	40.95
160	38.82	**41.66**	40.36	**41.32**

Table 2. Unweighted recall rates (UA) for 1- and 2- state HMMs modeling LSF features with and without RANSAC

Number of	1 state		2 states	
mixtures	All-data	RANSAC	All-data	RANSAC
16	34.53	34.24	36.59	36.71
56	36.69	**38.39**	35.38	37.54
80	36.67	**39.40**	35.65	36.95
160	36.82	39.30	35.98	37.50

We also provide a plot of unweighted recall rate versus number of Gaussian mixtures per state for 1-state and 2-state HMMs with and without RANSAC cleaning in Figure 1 and 2 for feature sets MFCCs and LSFs, respectively. If we compare the curves denoted by circles and squares for the feature sets, we can say that the RANSAC based data cleaning method brings significant improvements to the emotion recognition rate.

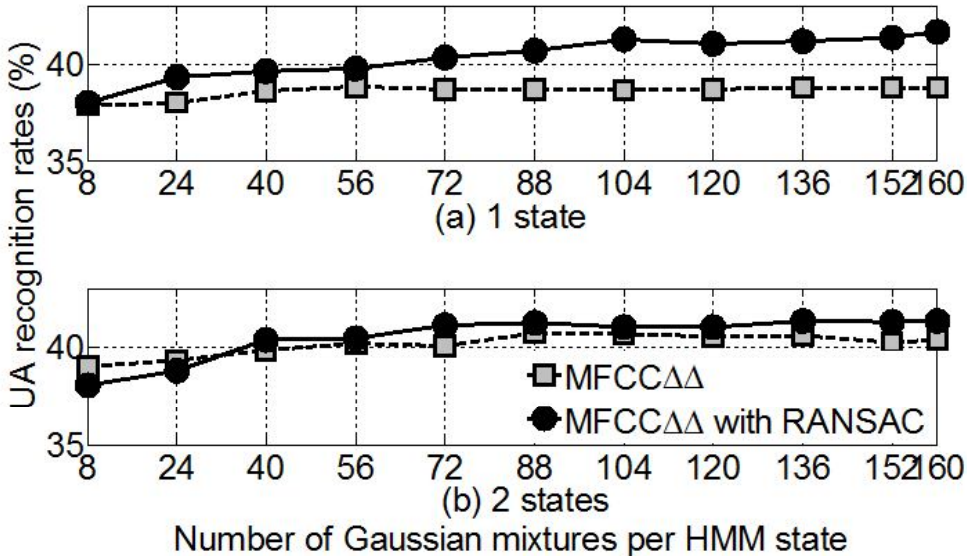

Fig. 1. Unweighted recall rate versus number of Gaussian mixtures per state for (a) 1-state and (b) 2-state HMMs modeling $MFCC\Delta\Delta$ features with and without RANSAC

Comparison of the Classifiers. We would like to compare the accuracies of the HMM classifiers with and without using RANSAC-based training data selection. There are various statistical tests for comparing the performances of supervised classification learning algorithms [5] [11]. The McNemars test tries to assess the significance of the differences in the performances of two classification algorithms that have been tested on the same testing data. The McNemars test has been shown to have low probability of incorrectly detecting a difference when no difference exists (type I error) [5].

We performed the McNemars test to show that the improvement achieved with the proposed RANSAC-based data cleaning method, as compared to employing all the available training data is significant. The McNemar's values for the MFCC feature set modeled by 1- and 2- state HMM classifiers with 160 Gaussian mixtures per state are computed as 231.246 and 8.917, respectively. Since these values are larger than the statistical significance threshold $^{2}_{(1,.95)} = 3.8414$, we can conclude that the improvement provided by RANSAC-based cleaning is statistically significant. The McNemar's values for the LSF feature set modeled by 1- and 2-state HMMs with 160 Gaussian mixtures per state are calculated as 196.564 and 22.448, respectively. Again, since these values are greater than the statistical significance threshold we can claim that the RANSAC based classifier has a better accuracy, which is statistically significant.

Note that the data we fed to the RANSAC-based training data selection algorithm consisted of chunks of one or more words for which three of the five labelers agreed on the emotional content. Using five labelers may not always be possible and if only one labeler is present, the training data is expected to be more noisy. In such cases, the proposed RANSAC based training data selection algorithm has the potential to bring even higher improvements to the performance of the classifier.

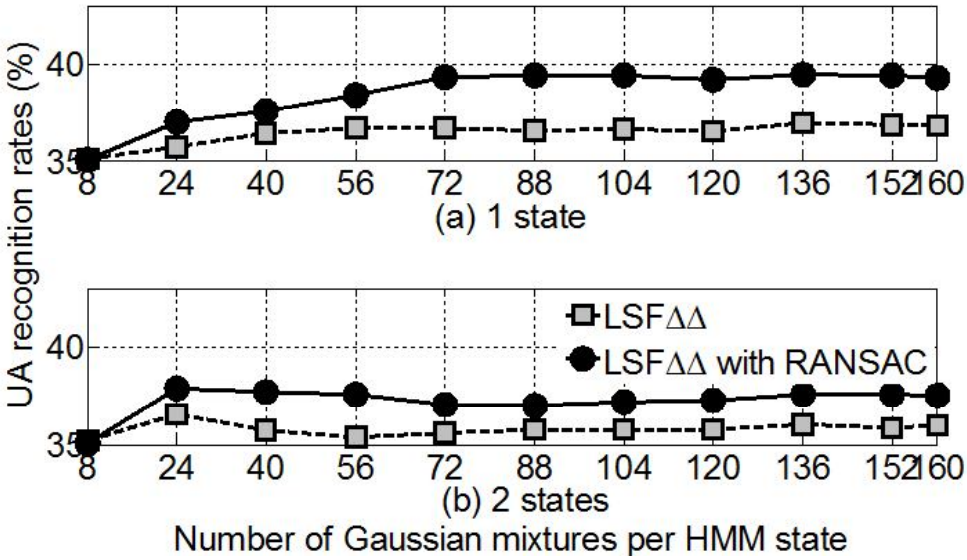

Fig. 2. Unweighted recall rate versus number of Gaussian mixtures per state for (a) 1-state and (b) 2-state HMMs modeling $LSF\Delta\Delta$ features with and without RANSAC.

One drawback of the RANSAC algorithm that was observed during the experiments is that it is time consuming, since many random subset selections need to be tested.

Decision Fusion of the RANSAC-based Trained Classifiers. Decision fusion of the RANSAC-based trained HMM classifiers is performed for various combinations of MFCC and LSF features. The fusion weight, α, is optimized over a subset of the training database prior to be used on the test data. The highest recall rate observed with the classifier fusion is 42.22 % for = 0.84 when 1-state HMMs with 80 mixtures modeling RANSAC-cleaned MFCCs are fused with 2-state HMMs with 104 mixtures modeling RANSAC-cleaned LSF features.

5 Conclusions and Future Work

In this paper, we presented a random sampling consensus based training data selection method for the problem of emotion recognition from a spontaneous emotional speech database. The experimental results show that the proposed method is promising for HMM based emotion recognition from spontaneous speech data. In particular, we observed an improvement of up to 2.84 % in the unweighted recall rates on the test set of the spontaneous FAU AIBO test set, significance of which have been shown by the McNemar's test. Moreover, the decision fusion of the LSF features with the MFCC features resulted in improved classification rates over the state-of-the-art MFCC-only decision for the FAU Aibo database.

In order to increase the benefits of the data cleaning approach, and to decrease the training effort, the algorithm may be improved by using semi-deterministic subset selection methods. Further experimental studies are planned to include more speech features (e.g., prosodic features), more complicated HMM structures and other spontaneous datasets.

Acknowledgments. This work was supported in part by the Turkish Scientific and Technical Research Council (TUBITAK) under projects 106E201, 110E056 and COST2102 action.

References

1. Angelova, A., Abu-Mostafa, Y., Perona, P.: Pruning training sets for learning of object categories. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2005)
2. Barandela, R., Gasca, E.: Decontamination of training samples for supervised pattern recognition methods. In: Amin, A., Pudil, P., Ferri, F., Iñesta, J.M. (eds.) SPR 2000 and SSPR 2000. LNCS, vol. 1876, pp. 621–630. Springer, Heidelberg (2000)
3. Ben-Gal, I.: Outlier Detection, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, Dordrecht (2005)
4. Breiman, L.: Bagging predictors. Machine Learning 24, 123–140 (1996)
5. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 7, 1895–1924 (1998)
6. Erzin, E., Yemez, Y., Tekalp, A.M.: Multimodal speaker identification using an adaptive classifier cascade based on modality realiability. IEEE Transactions on Multimedia 7(5), 840–852 (2005)
7. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Graphics and Image Processing 24 (1981)
8. Gu, B., Hu, F., Liu, H.: Sampling and its applications in data mining: A survey. Tech. Rep. School of Computing, National University of Singapore (2000)
9. Guyon, I., Matin, N., Vapnik, V.: Discovering informative patterns and data cleaning. In: Workshop on Knowledge Discovery in Databases (1994)

signals. Journal of the Acoustical Society of America 57(1), S35 (1975)
11. Kuncheva, L.I.: Combining Pattern Classifiers. John Wiley and Sons, Chichester (2004)
12. Kwon, O., Chan, K., Hao, J., Lee, T.: Emotion recognition by speech signals. In: Proc. of Eurospeech 2003, Geneva (September 2003)
13. Lee, C.M., Narayanan, S.S.: Toward detecting emotions in spoken dialogs. Journal 13, 293–303 (2005)
14. Lee, C.M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S.: Emotion recognition based on phoneme classes. In: Proc. ICSLP 2004, pp. 889–892 (2004)
15. Morris, R.W., Clements, M.A.: Modification of formants in the line spectrum domain. IEEE Signal Processing Letters 9(1), 19–21 (2002)

16. Olken, F.: Random Sampling from Databases. Ph. D. Thesis, Department of Computer Science, University of California, Berkeley (1993)
17. Ratsch, G., Onada, T., Muller, K.: Regularizing adaboost. Advances in Neural Information Processing Systems 11, 564–570 (2000)
18. Schuller, B., Rigoll, G., Lang, M.: Hidden markov model based speech emotion recognition. In: Proc. Int. Conf. Acoustics, Speech and Signal Processing, ICASSP (2003)
19. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: Interspeech (2009), ISCA. Brighton, UK (2009)
20. Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Aharonson, V.: Patterns, prototypes, performance: Classifying emotional user states. In: Interspeech (2008) ISCA (2008)
21. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis and Machine Vision. Thomson (2008)

and noisy multimedia data. ACM Transactions on Multimedia Computing, Communications and Applications 3 (2007)