RANSAC-based Training Data Selection for Emotion Recognition from Spontaneous Speech

Çiğdem Eroğlu Erdem
Dept. Electrical and Electronics Engineering
Bahçeşehir University
Beşiktaş, İstanbul, Turkey
(90)212 381 0895
cigdem.eroglu@bahcesehir.edu.tr

Elif Bozkurt, Engin Erzin
Dept. Electrical and Computer Engineering
Koç University
Sarıyer, İstanbul, Turkey
(90)212 338 1533
{ebozkurt, eerzin}@ku.edu.tr

A. Tanju Erdem
Dept. Electrical and Computer Engineering
Özyeğin University
Altunizade, İstanbul, Turkey
(90)216 559 2337
tanju.erdem@ozyegin.edu.tr

ABSTRACT
Training datasets containing spontaneous emotional expressions 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 various number of states and Gaussian mixtures per state indicate that utilization of RANSAC in the training phase provides an improvement of up to 2.84% in the unweighted recall rates on the test set. . This improvement in the accuracy of the classifier is shown to be statistically significant using McNemar's test.

Categories and Subject Descriptors
I.5.4. [**Computing Methodologies**]: Pattern Recognition – Applications: *signal processing, speech processing.*

General Terms
Algorithms

Keywords
Affect recognition, emotional speech classification, RANSAC, data cleaning, data pruning

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". 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]. 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.

AFFINE'10, October 29, 2010, Firenze, Italy.

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. [10] 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. [16], 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][12].

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 concentrates on different aspects and the second component is a probabilistic reasoning 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 [13] or averaging decisions of several functions such as bagging [6]. 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 [16]. 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 and the RANSAC-based HMM fitting approach. The details of the statistical method used for assessing the significance of the improvement in the accuracy of the HMM classifier is given in Section 4. In Section 5, our experimental results are provided, which is followed by conclusions and future work in Section 6.

2. BACKGROUND

2.1 The Spontaneous Speech corpus

The FAU AIBO corpus [1] is used in this study. The corpus consists of spontaneous, German and emotionally colored recordings of children interacting with Sony's 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, which may adversely affect the recognition 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 [1], [15]. 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 [8]. 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 [16][8]:

1. Suppose we have n training data samples $X = \{x_1, x_2, \ldots, x_n\}$ to which we hope to fit a model determined by (at least) m samples $(m \le n)$.
2. Set an iteration counter $k = 1$.
3. Choose at random m items from X and compute a model.
4. For some tolerance ε, determine how many elements of X are within ε 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 [16][8]. 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:

- ε, 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 ε 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 [8].
- *K,* which is the maximum number to run the algorithm while searching a satisfactory fit. Values of $K = 2\omega^{-m}$ or $K = 3\omega^{-m}$ have been argued to be reasonable choices [8], where ω is the probability of a randomly selected sample to be within ε of the model.

3. RANSAC-BASED DATA CLEANING METHOD

3.1 Speech Features and the Classifier

The feature set that we employ is the basic and widely use spectral features [14] and its derivatives. We also selected a simple Hidden Markov Model as the classifier, which will be described below.

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 f_C.

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. The MFCC feature vector is extended to include the first and second order derivative features, and the resulting feature vector with dynamic components is represented as: $\left[f_C^T \quad \Delta f_C^T \quad \Delta\Delta f_C^T\right]$, where ^{T} is the vector transpose operator.

As for the classifier, we chose to use a simple left-to-right Hidden Markov Model for each emotion class with single or two emitting states [18]. We considered HMMs with various number of Gaussian mixtures per state [4][5].

3.2 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 $X = \{x_1, x_2, \ldots, x_n\}$ to which we hope to fit a model determined by (at least) *m* samples $(m \le n)$. Initially, we randomly select $m = 320$ utterances considering use of 20 utterances per Gaussian mixture is sufficient for the training 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 $\varepsilon = \mu - 1.5\sigma$, where mean (μ) and standard deviation (σ) 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 [6]:

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

Here we set $\omega = \frac{m_i}{n}$, 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.

4. COMPARISON OF 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 [7][11]. McNemar's 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. McNemar's test has been shown to have low probability of incorrectly detecting a difference when no difference exists (type I error) [7].

In order to apply McNemar's test, the sample set is divided into a training set *R* and a test set *T*. In our case, the training set consists of the recordings of one school and the test set consists of the recordings of another school. Then, the two classification algorithms to be compared are trained on the training set yielding classifiers D_1 and D_2. In our case, let D_1 denote HMM classifiers trained using all training data and let D_2 denote HMM classifiers trained using RANSAC. Note that D_1 and D_2 consist of five HMMs, corresponding to the five emotions to be recognized. These classifiers are then tested on the same test set. Next, the following contingency table is constructed based on how each sample in the test set was classified by each of the two classifiers:

	D_2 correct	D_2 wrong
D_1 correct	N_{11}	N_{10}
D_1 wrong	N_{01}	N_{00}

If the null hypothesis, H_0, that is "the accuracies of the two classifiers are not different" is correct, then the two algorithms should have the same error rate, which implies $N_{01} = N_{10}$. This implies that the expected counts for both off-diagonal entries are $(N_{01} + N_{10})/2$. In order to measure the discrepancy between the expected and observed counts, the following statistic is used, which is approximately distributed as χ^2 [11]:

$$y^2 = \frac{\left(|N_{01} - N_{10}| - 1\right)^2}{N_{01} + N_{10}} \quad (1)$$

The McNemar's test is carried out by calculating y^2 and and by comparing it with the tabulated χ^2 value for a significance level (i.e. type I error) of 0.05. That means if $y^2 > 3.841$, the null hypothesis is rejected and we accept that the accuracies of the two classifiers are significantly different.

5. 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 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. The highest improvement (2.84%) is seen for the 1-state HMM with 160 Gaussian mixtures.

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. If we compare the curves denoted by circles and squares, we can say that for the 1-state HMM the RANSAC based data cleaning method brings significant improvements to the emotion recognition rate.

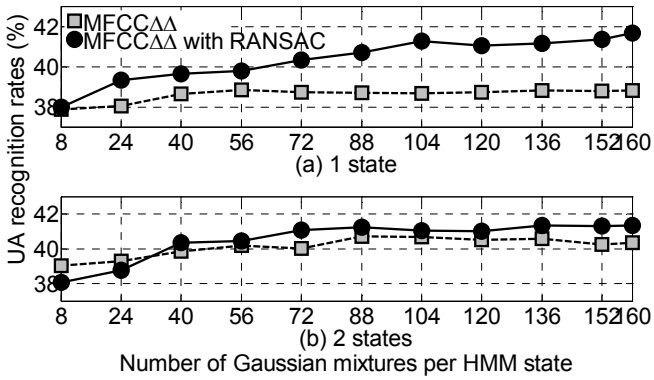

Figure 1. Unweighted recall rate versus number of Gaussian mixtures per state for (a) 1-state and (b) 2-state HMMs with and without RANSAC.

Comparison of the Classifiers: We performed the McNemar's test, which was described in Section 4, 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 contingency table for HMM classifiers with a single state and 160 Gaussian mixtures are given in Table 3. The McNemar's value is computed from (1) as $y^2 = 231.246$. Since this value is larger than the statistical significance threshold $\chi^2_{(1,0.95)} = 3.8414$, we can conclude that the improvement provided by RANSAC-based cleaning is statistically significant.

The contingency table for HMM classifiers with a two states and 160 Gaussian mixtures are given in Table 4. The McNemar's value is computed from (1) as $y^2 = 8.917$. Again, since this value is larger than the statistical significance threshold $\chi^2_{(1,0.95)} = 3.8414$, we can reject the null hypothesis and 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.

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.

Table 1. Unweighted Recall Rates (UA) for 1-state HMM

Number of Gaussian Mixtures per state	HMM Training Using		Improvement in UA
	All data	RANSAC cleaning	
16	38.39	39.51	1.12
56	38.84	39.79	0.95
80	38.63	40.62	1.99
160	38.82	41.66	2.84

Table 2. Unweighted Recall Rates (UA) for 2-state HMM

Number of Gaussian Mixtures per state	HMM Training Using		Improvement in UA
	All data	RANSAC cleaning	
16	38.46	38.63	0.17
56	40.17	40.45	0.28
80	40.18	40.95	0.77
160	40.36	41.32	0.96

Table 3. Contingency table for 1-state HMMs with 160 Gaussian mixtures.

	RANSAC Cleaned, (D_2) correct	RANSAC Cleaned, (D_2) wrong
All training data, (D_1) correct	2771	417
All training data, (D_1) wrong	988	4081

Table 4. Contingency table for 2-state HMMs with 160 Gaussian mixtures.

	RANSAC Cleaned, (D_2) correct	RANSAC Cleaned, (D_2) wrong
All training data, (D_1) correct	3097	623
All training data, (D_1) wrong	521	4016

6. 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 McNemar's test.

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.

7. ACKNOWLEDGMENTS

Ç. E. Erdem's work has been supported by Turkish Scientific and Technical Research Council (TUBITAK) under project EEAG-110E056 and Bahçeşehir University Research Fund. This work was also supported in part by TUBITAK under project 106E201 and COST2102 action.

8. REFERENCES

[1] Angelova, A., Abu-Mostafa, Y., and Perona, P. 2005. Pruning Training Sets for Learning of Object Categories, Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR).

[2] Barandela, R., and Gasca, E. 2000, Decontamination of Training Samples for Supervised Pattern Recognition Methods. Lecture Notes in Computer Science, vol. 1876, pp. 621-630.

[3] Ben-Gal I., Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, 2005.

[4] Bozkurt, E., Erzin, E., Erdem, C. E., Erdem, A. T. 2009. Improving Automatic Emotion Recognition from Speech Signals. Interspeech 2009, ISCA.

[5] Bozkurt, E., Erdem, C. E., Erdem, A. T., Erzin, E. 2010. Use of Line Spectral Frequencies for Emotion Recognition from Speech. Int. Conf. on Pattern Recognition, August 2010, İstanbul, Turkey.

[6] Breiman, L., 1996. Bagging Predictors. Machine Learning, 24(2), 123-140.

[7] Dietterich, T. G., Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 7(10):1895-1924, 1998.

[8] Fischler, M. A., and Bolles, R. C. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Graphics and Image Processing, Vol. 24, No. 6.

[9] Gu, B., Hu, F., Liu, H. 2000. Sampling and its Applications in Data Mining: A Survey. Tech. Rep. School of Computing, National University of Singapore, Singapore.

[10] Guyon, I., Matin, N., Vapnik, V. 1994. Discovering informative Patterns and Data Cleaning. Workshop on Knowledge Discovery in Databases.

[11] Kuncheva, L. I. Combining Pattern Classifiers. John Wiley & Sons, 2004.

[12] Olken, F. 1993. Random Sampling From Databases. Ph. D. Thesis, Department of Computer Science, University of California, Berkeley.

[13] Ratsch, G., Onoda, T., and Muller, K. 2000. Regularizing Adaboost, Advances in Neural Information Processing Systems, vol. 11, 564-570.

[14] Schuller, B., Steidl, S., and Batliner, A. 2009. The INTERSPEECH 2009 Emotion Challenge. Interspeech (2009), ISCA.

[15] Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Aharonson, V.2008. Patterns, Prototypes, Performance: Classifying Emotional User States. Interspeech (2008), ISCA.

[16] Sonka, M., Hlavac, V., and Boyle, R. 2008. Image Processing, Analysis and Machine Vision, Thomson.

[17] Wang, S., Dash, M., Chia, L. and Xu, M. 2007. Efficient sampling of training set in large and noisy multimedia data. ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 3, No.3.

[18] The Hidden Markov Toolkit, http://htk.eng.cam.ac.uk/