Multi-class and hierarchical SVMs for emotion recognition

Ali Hassan and Robert I. Damper

School of Electronics and Computer Science,
University of Southampton, SO17 1BJ, UK
{ah07r|rid}@ecs.soton.ac.uk

Abstract

This paper extends binary support vector machines to multi-class classification for recognising emotions from speech. We apply two standard schemes (one-versus-one and one-versus-rest) and two schemes that form a hierarchy of classifiers each making a distinct binary decision about class membership, on three publicly-available databases. Using the OpenEAR toolkit to extract more than 6000 features per speech sample, we have been able to outperform the state-of-the-art classification methods on all three databases.

Index Terms: emotion recognition, support vector machines, multi-class classification, hierarchical classification.

1. Introduction

It is important that practical and robust speech systems are able to cope with the realities of everyday conversational speech. In recent years, appreciation has been growing of the fact that a speaker's emotional state can affect profoundly the nature of their utterances, so that emotion recognition from speech has been gaining increasing attention, culminating in the Emotion Challenge at the last Interspeech conference [1]. In this paper, we aim to extend the state-of-the-art in emotion recognition beyond that reported last year by the use of multi-class support vector machines (SVMs).

2. Databases

From the current literature [2, 3], three different kinds of emotional speech database are observed: simulated or acted, natural or spontaneous and elicited speech databases. As we move from acted to spontaneous emotions, the complexity of the task increases [4].

The most popular acted databases are the the German emotional speech database [5], also known as EMO-DB, and Danish emotional speech database (DES) [6], and the Serbian emotional speech database [7]. These are all publicly available for non-commercial research use. Other acted emotional speech databases exist (e.g., BabyEars [8]), but we have used only EMO-DB, DES and the Serbian database in this work. Most natural or spontaneous databases are proprietary as they are typically recorded in customer interaction services like call centres. Hence, results on these can not be easily reproduced or compared. In elicited speech, emotions are induced by putting the subjects into certain controlled conditions. Examples of elicited databases are SmartKom [9] and German AIBO [10].

Most research in emotion recognition is on acted speech rather than spontaneous or elicited speech. Whereas spontaneous speech is clearly more realistic [1], and there is a shift towards using spontaneous data, we continue to use acted speech here for comparability with our earlier work [11].

3. Approaches to Classification of Emotions

3.1. Features

Murray and Arnott [12] studied the effect of different emotions on acoustic features of speech from the point of view of synthesis. The acoustic features of pitch, energy, speech rate and voice quality were extracted from the whole utterance on the assumption that the emotion will remain constant over the utterance, which can therefore be treated as a single unit. Most research in this area uses the same assumption, which works well. However, it has been argued that this is not appropriate for spontaneous speech. As the emotions in spontaneous speech are instantaneous, features should be extracted over smaller segments of speech. So far, attempts to use short-term features alone for emotion classification have not been as successful as those using utterance level features [13, 14].

Batliner et al. [15] pooled together a number of features from sets independently developed at different sites. This resulted in improved classification accuracies. Subsequently, Eyben et al. [16] have described an open-source toolkit called OpenEAR for extracting these pooled features (more than 6000 in number).

3.2. Classifiers

A variety of methods from the simplest like k-nearest neighbour (k-NN) through decision trees to more complex classifiers like support vector machines (SVMs), artificial neural networks (ANN), hidden Markov models (HMMs) have been used to solve this problem. Static classifiers like k-NN, SVM, ANN and decision trees with long-term features have been the methods of choice.

Several researchers have tried to create classifiers that combine features from different sources of information to achieve good accuracy. Schuller et al. [17] used an ensemble of classifiers for recognising different sets of emotions from film sound tracks. Similarly, Sidorova [18] proposed the so-called TGI+ classifier that uses a combination of many classifiers. However, the gain in accuracy compared to the increased complexity is not very high.

3.3. Hierarchical Classification

Many of the papers presented in the 2009 Interspeech Emotion Challenge have used some form of hierarchical multi-class classification methods. Lee et al. [19] used a hierarchical binary decision tree; at each node of the tree, a binary Bayes logistic regression or SVM classifier is trained. The authors placed Angry/Emphatic versus Positive emotions at the root node based upon prior empirical testing. Planet et al. [20] used multiple classifiers with different hierarchical structures. In one of these, they used a binary classifier to distinguish between

Neutral and other classes at the first level and then used a multi-class classifier at the next level to separate the rest of the classes. In another scheme, a cascade of binary classifiers was used in which at each level they separate the most populated class from the rest. So for m classes, they have to train $m-1$ classifiers. Luengo et al. [21] used a one-versus-rest scheme for training two different types of classifiers for each emotion and fused the results using a scoring scheme. Shaukat and Chen [22] used the valence arousal model to determine a hierarchical structure for classification of the Serbian emotional speech database. At the root node, they divide the emotions into active and nonactive emotions using a binary SVM classifier. They further divided the emotions depending upon the valence to give the final classification result.

4. Methodology

Support vector machines were originally developed for binary classification; much subsequent work has been done to extend SVMs to multi-class classification. Currently there are two main approaches: one is by combining several binary classifiers, the other is by considering all classes in one single optimisation function. One-versus-one (1v1) [23], one-versus-rest (1vR) [24], directed acyclic graph (DAG) [25], and unbalanced decision tree (UDT) [26] are methods based on binary classification. An example of a method which considers all the classes at once can be found in [27]. In this paper, we have only considered binary classification methods.

4.1. 1vR, 1v1, DAG and UDT

The one-versus-rest (1vR) scheme applies a ‘winner takes all’ strategy as in Figure 1(a) for the case of four classes. To classify m classes, this method has to construct m binary SVM classifiers. The jth classifier is trained with all the training data in the jth class given positive labels, and all the rest given negative labels. Thus, given l training samples $(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)$ where $x_i \in \mathbb{R}^n$, $i = 1, 2, \ldots, l$ and $y_i \in \{1, 2, \ldots, m\}$, we have m binary classifiers each with a classification function f_j. For any input sample x_i, the scheme assigns it to the class with the highest classification function value:

$$\text{class of } x_i = \underset{j=1,\ldots,m}{\arg\max} \quad f_j(x_i) \tag{1}$$

The benefit of the 1vR scheme is that we only have to train m classifiers. However, we have to deal with highly unbalanced training data for each binary classifier.

The one-versus-one (1v1) classifier uses a ‘max-wins’ voting strategy as illustrated in Figure 1(b) for the case of four classes. It constructs $m(m-1)/2$ binary classifiers, one for every pair of distinct classes. Each binary classifier C_{ij} is trained on the data from the ith and jth classes only. For a given test sample, if classifier C_{ij} predicts it is in class i, then the vote for class i is increased by one; otherwise the vote for class j is increased by one. Then the ‘max-wins’ voting strategy assigns the test sample to the highest scoring class. In this method, we don’t have to handle highly unbalanced classes as in the case of 1vR. However, we have to train more classifiers than for 1vR.

The DAG method also has to train $m(m-1)/2$ binary classifiers. The training phase is same as for 1v1; however, in the testing phase, it uses a rooted binary directed acyclic graph with $m(m-1)/2$ internal nodes and k leaves. Each node is a binary SVM classifier, C_{ij} for ith and jth classes. For every test sample, starting at the root node, the sequence of binary decisions at each node determines a path to a leaf node that indicates the predicted class. Figure 2(a) shows the architecture of DAG for the classification of four classes.

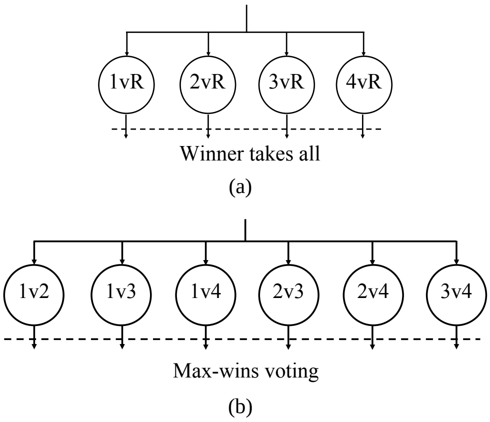

Figure 1: *Non-hierarchical architectures for classification of four classes: (a) One-versus-rest (1vR); (b) One-versus-one (1v1).*

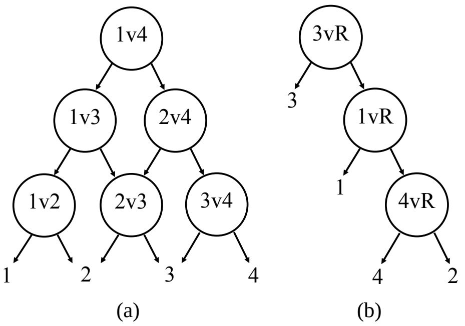

Figure 2: *Hierarchical architectures for classification of four classes: (a) Directed acyclic graph (DAG); (b) Unbalanced decision tree (UDT).*

The unbalanced decision tree (UDT) converts the 1vR scheme into a right-branching tree structure. In this method, we have to train $(m-1)$ binary classifiers as compared to the m classifiers of 1vR. These are arranged in a cascade structure similar to [20]. Starting at the root node, one selected class is evaluated against the rest. The decision about which classifier to place at the root is made by taking the most separable class from the rest using the results of the 1vR scheme. Then, the UDT method proceeds to the next level by eliminating the class from the previous level from the training samples. That is, UDT uses a ‘knock-out’ strategy that, in the worst-case scenario, requires $(m-1)$ classifiers, and in the best case needs only one classifier. Figure 2(b) shows the architecture of UDT for classification of four classes. Class 3 has been chosen as the root node on the assumption that this class is maximally separable from the rest.

4.2. Experimental setup

As stated earlier, we have used three databases of acted speech: German EMO-DB Danish DES and the Serbian emotional speech database. Table 1 shows the number of emotion classes in each database, and the number of sentences for each of

Table 1: *Number of emotion classes in the EMO-DB, DES and Serbian databases and number of sentences per class. Key – A : Angry, N : Neutral, S : Sad, H : Happy, F = Fear, B : Bored, D : Disgust and U : Surprised.*

EMO-DB	# sent.	DES	# sent.	Serbian	# sent.
A	127	A	52	A	558
N	79	N	52	N	558
S	62	S	52	S	558
H	71	H	52	H	558
F	69	U	52	F	558
B	81				
D	46				

Table 2: *Unweighted percentage average results for 10-FCV with four classifier schemes on three databases of acted speech. The 4 classes identified for EMO-DB and DES are those common the the two (i.e., A, N, S, H as in Table 1).*

Database	# of classes	1vR	UDT	1v1	DAG	Human acc.
EMO-DB	4	87.9	92.0	95.6	**95.8**	87.4
	7	80.4	88.8	**92.3**	91.5	86.1
DES	4	67.3	78.8	**85.6**	84.8	69.4
	5	50.4	70.4	74.2	**75.0**	67.3
Serbian	5	93.0	**94.6**	94.1	93.3	94.7

these classes. The open-source toolkit OpenEAR has been used to extract 6553 features from each test sample. Each feature is normalised and then discretised using the method described in [28]. For the individual binary classifiers, we use the LibSVM [29] implementation of support vector machines. We could have chosen any classifier for this task but, as SVMs have been the method of choice for most researchers in this area, we have also used them. We use a linear kernel instead of, say, a radial basis function (RBF) kernel since, according to Hsu et al. [30], when the number of features is very large compared to the number of instances, there is no significant benefit to using an RBF kernel over a linear SVM. Usually, the value of the cost parameter C for an SVM is determined by a search over some appropriate space, but for fair comparison of the different schemes we have fixed it to $C = 0.1$. Training/testing is by 10-fold cross-validation (10-FCV).

5. Results and Discussion

Figure 3 shows the average accuracy for 10-FCV for the four classes that are common between all three databases, using the 1vR scheme. Individually, the Sad emotion is most easily discriminated from the other classes. We have used this in the UDT scheme and place Sad versus Rest at the root note of the classification tree. Using the four classification schemes on the three databases, the average accuracy for 10-FCV evaluation is given in Table 2. We also tabulate human accuracy at recognising the emotions, which is documented with each database. For 1vR and UDT, we have very unbalanced datasets. We have tried simply up-sampling the minority class, but this did not give a significant increase of accuracy and so it was not pursued further.

For many cases, the classification results in Table 2 appear much better than any previously published results, although training/testing conditions may not be entirely comparable. The best known previous result on EMO-DB for 7 classes is by Schuller et al. [17] in which the authors have used ensemble classifiers to achieve 87.5% accuracy while we report 92.3% using a 1v1 linear SVM. For the EMO-DB 4 class problem, Shami and Verhelst [13] report 75.5% accuracy, whereas we achieve 95.8%. For DES, the latter authors achieve 64.9% average accuracy for 4 classes using SVMs and 10-FCV but with different folds from those used here. The best unweighted average accuracy that we have achieved is 85.6% and 75.0% for 4 and 5 classes, respectively. For the Serbian database, Shaukat and Chen [22] report 89.7% accuracy using hierarchical SVMs; our best results on this database are 94.6% using UDT with linear SVMs.

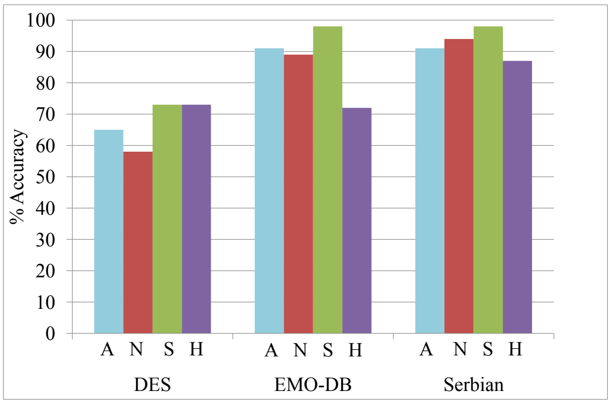

Figure 3: *Percentage accuracy for four emotional classes for each database using 1vR scheme with linear SVM. Key – A : Angry, N : Neutral, S : Sad and H : Happy.*

Several observations can be made about the results in Table 2. DAG and 1v1 seem to be giving very similar results. This is almost certainly because both use the same training methods; they only differ in their testing methods. Out of the other two schemes (i.e., 1vR and UDT), 1vR performs relatively poorly whereas UDT performs closer to 1v1 and DAG. Ramanan et al. [26] also found UDT performance very close to DAG and 1v1 on a variety of benchmark datasets. In UDT, after every level of classification, we drop the test samples for the tested class. In other words, we are actually making the classification task more balanced by removing some of the data from the majority class after every level of classification, which is the reason that UDT outperforms 1vR. Another observation from the results is that we are approaching or in some cases exceeding human accuracy on these databases. There are a few possible reasons for this. First, we can challenge the quality of the human recognisers. They might not be very good at the job or they might not have been given enough time to adjust to the speaker's delivery style. In some cases, the listeners were allowed to listen to the test sample only once. And, of course, if it is difficult to compare machine recognisers (e.g., because the folds used in 10-FCV are different), it is even more difficult to compare machine and human recognition on a fair basis.

6. Conclusions

We have explored different methods for applying multi-class (hierarchical and non-hierarchical) classification schemes using linear binary SVMs to emotion recognition from speech using three publicly-available databases, with results that exceed the best previously-report performance on these datasets. The four classification schemes explored are not restricted to SVMs;

in principle, we can use any binary classifier. Of the four schemes tested, DAG and 1v1 perform equally well with UDT not far behind.

Our future work includes comparing segment-based and utterance-based feature extraction with respect to recognition performance. We also intend to try different classifiers (other than linear SVMs) within the same hierarchical structures to assess the effect on classification. Lastly, we would like to see the effect of training on one database and testing on the others, to assess the extent to which emotion in speech is language-independent.

7. References

[1] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Interspeech'09, Proceedings of 10th Annual of the International Speech Communication Association*, Brighton, UK, 2009, pp. 312–315.

[2] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 2, pp. 33–60, 2003.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[4] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech," in *Interspeech 2006, Proceedings of International Conference on Spoken Language Processing*, Pittsburgh, PA, 2006, pp. 1093–1097.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech'05, Proceedings of 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, pp. 1517–1520.

[6] I. Engberg and A. Hansen, *Documentation of the Danish Emotional Speech Database DES*, Center for PersonKommunikation, Institute of Electronic Systems, Alborg University, Aalborg, Denmark, 1996.

[7] S. T. Jovicic, Z. Kacic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: design, processing and evaluation," in *Proceedings of 9th Conference on Speech and Computer, SPECOM'04*, St. Petersburg, Russia, 2004, pp. 77–81.

[8] M. Slaney and G. McRoberts, "BabyEars: A recognition system for affective vocalizations," *Speech Communication*, vol. 39, no. 3–4, pp. 367–384, 2003.

[9] F. Schiel, S. Steininger, and U. Türk, "The SmartKom multimodal corpus at BAS," in *Proceedings of the 3rd Language Resources and Evaluation Conference, LREC'02*, Canary Islands, Spain, 2002, pp. 200–206.

[10] A. Batliner, C. Hacker, S. Steidl, E. Noth, S. D'Arcy, M. Russell, and M. Wong, "'You stupid tin box' – children interacting with the AIBO robot: a cross-linguistic emotional speech corpus," in *Proceedings of 4th Language Resources and Evaluation Conference LREC, 2004*, Lisbon, Portugal, 2004, pp. 171–174.

[11] A. Hassan and R. I. Damper, "Emotion recognition from speech using extended feature selection and a simple classifier," in *Interspeech'09, Proceedings of 10th Annual of the International Speech Communication Association*, Brighton, UK, 2009, pp. 2403–2406.

[12] I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[13] M. Shami and W. Verhelst, "An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech," *Speech Communication*, vol. 49, no. 3, pp. 201–212, 2007.

[14] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *IEEE International Conference on Semantic Computing, '08*, Santa Clara, CA, 2008, pp. 158–165.

[15] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Proceedings of Information Society-Language Technologies Conference IS-LTC*, Ljubljana, Slovenia, 2006, pp. 240–245.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR - Introducing the Munich Open-source Emotion and Affect Recognition Toolkit," in *Proceedings of International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, 2009, pp. 1–6.

[17] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *IEEE International Conference on Multimedia and Expo, ICME 05*, Amsterdam, The Netherlands, 2005, pp. 864–867.

[18] J. Sidorova, "Speech emotion recognition with TGI+.2 classifier," in *Proceedings of 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL'09*, Athens, Greece, 2009, pp. 54–60.

[19] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," in *Interspeech'09, Proceedings of 10th Annual of the International Speech Communication Association*, Brighton, UK, 2009, pp. 320–323.

[20] S. Planet, I. Iriondo, J. Socoró, C. Monzo, and J. Adell, "GTM-URL contribution to the Interspeech 2009 Emotion Challenge," in *Interspeech'09, Proceedings of 10th Annual of the International Speech Communication Association*, Brighton, UK, 2009, pp. 316–319.

[21] I. Luengo, E. Navas, and I. Hernáez, "Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge," in *Interspeech'09, Proceedings of 10th Annual of the International Speech Communication Association*, Brighton, UK, 2009, pp. 332–335.

[22] A. Shaukat and K. Chen, "Towards automatic emotional state categorisation from speech signals," in *Interspeech'08, Proceedings of 9th Annual of the International Speech Communication Association*, Brisbane, Australia, 2008, pp. 2771–2774.

[23] C. Hsu and C. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2001.

[24] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge, UK: Cambridge University Press, 2000.

[25] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Proceedings of Neural Information Processing Systems, NIPS'99*, Denver, CO, 2000, pp. 547–553.

[26] A. Ramanan, S. Suppharangsan, and M. Niranjan, "Unbalanced decision trees for multi-class classification," in *International Conference on Industrial and Information Systems, ICIIS'07*, Sri Lanka, 2007, pp. 291–294.

[27] J. Weston and C. Watkins, "Multi-class support vector machines," in *Proceedings of European Symposium on Artificial Neural Networks, ESANN'09*, Bruges, Belgium, 1999, pp. 219–224.

[28] U. Fayyad and K. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry, France, 1993, pp. 1022–1027.

[29] O. Chapelle, P. Haffner, and V. Vapnik, "Support vector machines for histogram-based image classification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1055–1064, 1999.

[30] C. Hsu, C. Chang, and C. Lin, *A Practical Guide to Support Vector Classification*, Department of Computer Science, National Taiwan University, Taiwan, 2003.