Emotion Recognition from Speech by Combining Databases and Fusion of Classifiers

Iulia Lefter[1,2], Leon J. M. Rothkrantz[1,2], Pascal Wiggers[1], and David. A. van Leeuwen[3]

[1] Delft University of Technology, The Netherlands
[2] The Netherlands Defense Academy
[3] TNO Human Factors, The Netherlands

Abstract. We explore possibilities for enhancing the generality, portability and robustness of emotion recognition systems by combining databases and by fusion of classifiers. In a first experiment, we investigate the performance of an emotion detection system tested on a certain database given that it is trained on speech from either the same database, a different database or a mix of both. We observe that generally there is a drop in performance when the test database does not match the training material, but there are a few exceptions. Furthermore, the performance drops when a mixed corpus of acted databases is used for training and testing is carried out on real-life recordings. In a second experiment we investigate the effect of training multiple emotion detectors, and fusing these into a single detection system. We observe a drop in the Equal Error Rate (EER) from 19.0 % on average for 4 individual detectors to 4.2 % when fused using FoCal [5].

1 Introduction

Emotion recognition from speech is a field that gains more and more attention from researchers. Typically, machine learning techniques are used to train models of features extracted from databases of emotional speech [15]. Even though the general architectures of the systems are similar, no unity can be found within the components. The results are hard to compare due to inconsistencies in data, task and labeling. Details of these problems are outlined in [17] where a call for standardization is being made.

Obtaining data for training is not trivial. A recent trend is to replace acted emotions by real, spontaneous ones. For this purpose, different emotion elicitation methods are used, e.g. children interacting with a remotely controlled pet robot [19].

Recent work aims at finding the most important feature types [2]. The idea is to extract a large number of features and then reduce this figure, keeping most relevant ones. However, the resulting feature set is highly dependent on the database being used. As noted by [23], different features are relevant in the case of acted and spontaneous emotions.

The goal of this paper is to explore the portability of emotion recognition systems and improve the robustness. Usually experiments involve the use of single databases. As a first experiment, we use four databases of emotional speech, three with acted emotions and one with real-life recordings from call centers. Our approach is to choose a fixed database for testing and to use different data combinations for training. This includes training on the same database, on a different database and on a merged database that includes or not the test database, in a speaker independent way. Typically models are database dependent and are not expected to work well on new types of data. A way to remedy this is to provide a larger amount of training data. With this experiment we examine the benefits of using extended corpora as well as the portability of systems trained on acted data to real life scenarios. Research using multi-corpus training and testing is presented in [18] and [21].

Since the performance of emotion recognition systems is still far from 100% accurate, especially when test data is from a different dataset than the training one, we investigate the improvements of fusing the results of more classifiers trained with different feature sets on spontaneous data. We use both utterance and frame level features, whose combination is expected to enhance the recognition as shown in [22]. Late fusion by linear combination of the scores given equal weights and also weights calculated with logistic regression are compared. Both of them yield higher performance than the individual classifiers.

This paper is organized as follows. In Section 2 we introduce the emotional speech databases used in this work and the methodology for training and testing. Details about the setups and the results of the first and the second experiment are provided in Sections 3 and 4 respectively. The last section contains our conclusions.

2 Methods and Materials

We use four databases for training and testing: the German database (BERLIN) [6], the Danish database of emotional speech (DES) [9], the audio part of the eNTERFACE'05 database (ENT) [14] and the South-African Database (SA) [12]. Details about the characteristics of these databases can be found in Table 1.

The idea is to use subsets of the databases that contain the same emotions. Firstly, we use only combinations of the three acted databases and the three emotions they have in common: anger, happiness and sadness (Experiment 1.a). Secondly, we include also the database of spontaneous speech, and consider just two classes: anger and neutral (Experiment 1.b). Even though ENT does not contain a neutral class, we have decided to use its samples from the anger class in this experiment.

Given a fixed test set, three training conditions are implemented for the first experiment: *within corpus* (same database is used for training and testing), *cross corpus* (the databases for training and testing are different), and *mixed corpus* (samples corresponding to the same emotion but belonging to different databases are considered as one class, and speaker independent classification is performed).

Table 1. Characteristics of the databases of emotional speech used

Feature	BERLIN	DES	ENT	SA
# anger	127	50	211	1000
# disgust	38		211	
# fear	55		211	
# happiness	64	52	208	
# sadness	53	52	211	
# surprise		50	211	
# boredom	79			
# neutral	78	52		2000
# speakers	10 (5 male)	4 (2 male)	46	1228
acted/spontaneous	acted	acted	acted	spontaneous
language	German	Danish	English	English/Afrikaans
utterance type	preset	preset	preset	free
mean duration (sec)	2.76	5.46	2.81	4.3
total duration (min)	22.8	30.68	59.06	215.02
recording condition	mic	mic	mic	telephone

Our approach is to consider one emotion as target and the other emotions as non-target. A detector for the target emotion can make two types of errors that can be traded off: misses and false alarms. We asses the performance of our detectors in terms of equal error rates (EER) where false alarm and miss rates are equal.

All experiments are implemented using speaker independent cross-validation with z-normalization of features on the training set in order to achieve $\mu = 0$ and $\sigma = 1$. For BERLIN and DES which have a small number of speakers we use leave-one-speaker-out cross-validation. For ENT and SA we use 10 fold cross-validation.

In the case of the second experiment we fuse detectors whose scores span different ranges (some are log likelihoods, some probabilities). It is therefore important to normalize the scores. For this reason we have used 10-fold speaker independent double-cross validation and an adapted form of t-normalization [1]. The mean and standard deviation of the scores of the non-target development set are used in order to normalize the scores of the evaluation set.

3 Experiment 1 - Multiple Corpus Training and Testing

In this experiment we test the ability of models trained on one database to generalize to another one. We use a prosodic, utterance level feature set inspired from the minimum required set of features proposed by [11] and the approach of [20]. The feature set contains: pitch(mean, standard deviation, range, absolute slope (without octave jumps), jitter), intensity (mean, standard deviation, range, absolute slope, shimmer), means of the first 4 formants, long term averaged spectrum (slope, Hammarberg index, high energy) and center of gravity and skewness of the spectrum. These features were extracted using Praat [3] and we will refer to them as prosodic features. Classification is performed by Support Vector Machines (SVM) with a radial basis function (RBF) kernel by means of LIBSVM [8]. We refer to this method as SVM.

Table 2. Results of Experiment 1.a

Experiment	**Train corpus**	**Test corpus**	**EER**		
			anger	happiness	sadness
within corpus	BERLIN	BERLIN	11.6	18.9	14.8
	DES	DES	31.8	33.0	25.0
	ENT	ENT	26.1	36.7	22.3
cross corpus	DES	BERLIN	31.5	53.2	44.3
	ENT	BERLIN	44.9	45.4	19.9
	DES+ENT	BERLIN	38.4	46.8	24.0
	BERLIN	DES	31.9	44.7	33.0
	ENT	DES	**29.9**	34.0	**13.1**
	BERLIN+ENT	DES	**29.9***	34.5	**17.5**
	BERLIN	ENT	38.8	45.6	30.2
	DES	ENT	33.2	36.9	**16.8**
	BERLIN+DES	ENT	35.7	**36.2***	**16.4***
mixed corpus	BERLIN+DES+ENT	BERLIN	20.5	25.0	**3.5**
	BERLIN+DES+ENT	DES	**26.5**	**32.5**	**15.5**
	BERLIN+DES+ENT	ENT	30.1	**36.2**	**16.3**

The results for Experiment 1.a, which uses the three acted databases and three emotions are presented in Table 2. The results for Experiment 1.b in which all four databases and two classes are used are provided in Table 3. The within corpus results can be considered as reference values. In general the cross corpus tests result in worse EERs than the reference values. Interestingly, there are also some exceptions which are printed in bold. Results marked with a star (*) in the cross corpus experiment highlight that there is an improvement by merging databases for training. The mixed corpus approach gives an improvement to both the within- and cross corpus results for most conditions.

Table 3. Results of Experiment 1.b

Experiment	**Train corpus**	**Test corpus**	**EER**
within corpus	BERLIN	BERLIN	1.4
	DES	DES	28.4
	SA	SA	15.5
cross corpus	BERLIN+DES+ENT	SA	29.9
mixed corpus	BERLIN+DES ENT+SA	BERLIN	3.9
	BERLIN+DES+ENT+SA	DES	**25.5**
	BERLIN+DES+ENT+SA	SA	16.5

When all four databases are used, we are interested, for the cross corpus case, in the performance of classifiers trained on acted and tested on real data. In this case the EER of the cross corpus condition is twice that of the within-corpus condition. The mixed approach shows an improvement only in the case of testing on DES, while for SA the result is slightly lower than the reference value. The ENT database is only used in this experiment for training, since it does not contain a neutral class.

4 Experiment 2 - Fusion of Classifiers

The aim of this experiment is to improve the performance of emotion detection. We use only the SA database, which is more difficult since it contains free natural speech as opposed to preset utterances and the convenient lab conditions are replaced with noisy telephone speech. We are interested in the performance of different classification methods, as well as their fusion.

A first detection approach uses SVM and the prosodic feature set described in Section 3. Further, we use three spectral feature based classifiers popular in speaker recognition. They are based on Relative Spectral Perceptual Linear Predictive (RASTA PLP) coding of speech [10]. In order to extract the features from the sound signal, voice activity detection is performed based on energy levels. Every 16 ms, 26 coefficients are extracted for a frame of 32 ms : 12 PLP coefficients plus log energy and their derivatives.

The Universal Background Model - Gaussian Mixture Model (UBM-GMM) [16] approach models each class by a mixture of Gaussians based on the RASTA PLP features. We use a 512 mixtures precomputed (UBM) trained on NIST SRE 2008 data. This is MAP adapted using either emotion or neutral speech data. We refer to this method as GMM.

The third technique is a UBM-GMM-SVM detector [7]. The feature supervectors are the means of the UBM-GMM model. These feature sets are used for SVM classification.

The final classifier is known as dot-scoring (DS) [4], and it is a linear approximation of UBM-GMM. It uses sufficient fixed size zero and first order statistics of these features. The method includes channel compensation, meaning that the impact of the communication medium is reduced.

Two types of score level fusion are applied on the scores of these four classification methods: a linear combination of the t-normalized scores with equal weights, and fusion by calculating the weights using linear logistic regression using FoCal [5]. For the second fusion type, a constant is added to the formula for calibration. This approach provides simultaneous fusion and calibration in a way that optimizes discrimination and calibration. The fused scores tend to be well-calibrated detection log-likelihood-ratios.

As we expect different classifiers based on different features to complement each other, we fuse in turn the SVM with prosodic features with each of the GMM-like approaches which are based on RASTA PLP features. Finally, we fuse all four classifiers by logistic regression. The results are shown in Table 4 for different weights of each classifier to the final result. The weights different than 1 are calculated with FoCal. SVM gives the highest performance from the individual classifiers and is always assigned high weights for fusion. However, UGS which has a lower performance by itself is assigned slightly higher weights.

Table 4. EERs for individual classifiers and various combinations with different weights

Classifier	**Weight**										
GMM	1				1	3.03					3.35
UGS		1					1	5.62			5.92
DS			1						1	1.81	1.72
SVM				1	1	5.77	1	5.44	1	5.27	5.13
EER(%)	21.2	19.8	19.6	15.5	11.3	9.9	10.5	10.1	11.3	11.0	4.2

The detection error tradeoff (DET) curves [13] of the individual detectors and their fusion are presented in Figure 1. The results show clearly that fusion leads to great improvements.

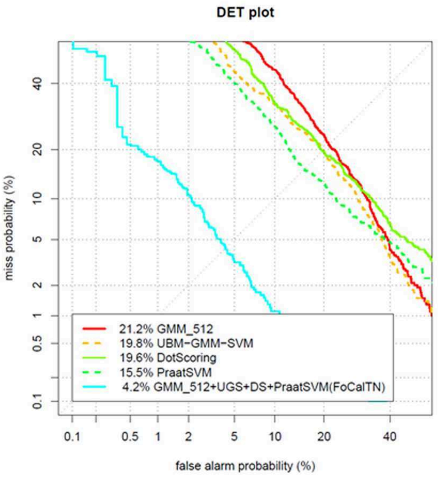

Fig. 1. DET curves for the fusion of Dot Scoring, SVM, GMM and UBM-GMM-SVM by logistic regression

5 Conclusions

In this paper we have investigated several aspects related to emotion recognition in speech. First, we have investigated the ability of an emotion detector to generalize to a different data set. For this we use the common emotions found in three widely used emotion databases: BERLIN, DES, and ENT with emotions anger, happiness and sadness. Surprisingly, we found that for the DES test data we obtained better performance for detectors trained on data including the ENT database, than using training on DES alone. This may be due to the fact that when using DES for training, only 3 actors are available, of which only 1 has the same gender as the test speaker. Here, the classifier obviously can benefit from a wider variability in speakers, even if the recording protocol, way of eliciting the emotions, or even the language of the speech used are different. Another aspect of emotion in speech is whether it results from acted or real emotion. In order to study this, we used data collected from a call center, where two emotions dominate calls from clients: anger and neutral. Here, we observe that mixing in acted emotions does not lead to additional performance with our baseline classifier. We may presume that the emotion cause (real versus acted) is too different between the test and additional training data, although we cannot exclude that other sources of variation (channel, language) also prevent the emotion models from improving.

As a final experiment we have looked into methods for improving a single emotion detector tested on the natural SA data. By using additional features and several frame-based classifiers borrowed from speaker recognition, we could show a very strong improvement in performance from 19 % on average for the 4 individual detectors to 4.2 % for the fused detectors. This is consistent with what has been observed in speaker recognition [5], but it is interesting to note that the fusion still works so well for very short duration utterances (2 seconds, compared to the 2 minutes we are used to in speaker recognition) and for three classifiers that are based on the same spectral features. Although there still is a long way to go before we have robust emotion detectors that are not sensitive to spontaneity, language or recording channel, we do believe that the methods and experiments presented in this paper give some insight into what can be promising approaches in both technology and data collection.

References

1. R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10:42–54, 2000.
2. A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir. Whodunnit - Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech. *Computer Speech and Language*, 2010.
3. P. Boersma. Praat, a System for Doing Phonetics by Computer. *Glot International*, 5(9/10):341–345, 2001.
4. N. Brümmer. Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics. In *Proceedings of Interspeech*. ISCA, 2009.
5. N. Brümmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker

Recognition Evaluation 2006. *IEEE Transactions on Speech, Audio and Language Processing*, 15(7):2072–2084, 2007.

6. F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A Database of German Emotional Speech. *Proceedings of Interspeech*, pages 1517–1520, 2005.
7. W. Campbell, D. Sturim, and D. Reynolds. Support Vector Machines Using GMM Supervectors for Speaker Verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.
8. C. C. Chang and C. J. Lin. *LIBSVM: a Library for Support Vector Machines*, 2001.
9. I. S. Engberg and A. V. Hansen. Documentation of the Danish Emotional Speech Database (DES). Internal AAU report, Center for Person Kommunikation, 1996.
10. H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. RASTA-PLP speech analysis technique. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 121–124, 1992.
11. P.N. Juslin and K.R. Scherer. *In J. Harrigan, R. Rosenthal, and K. Scherer, (Eds.) - The New Handbook of Methods in Nonverbal Behavior Research*, chapter Vocal Expression of Affect, pages 65–135. Oxford University Press, 2005.
12. I. Lefter, L. J. M Rothkrantz, P. Wiggers, and D. A. van Leeuwen. Automatic Stress Detection in Emergency (Telephone) Calls. *Int. J. on Intelligent Defence Support Systems*, 2010. submitted.
13. A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The Det Curve In Assessment Of Detection Task Performance. In *Proceedings Eurospeech '97*, pages 1895–1898, 1997.
14. O. Martin, I. Kotsia, B. Macq, and I. Pitas. The eNTERFACE'05 Audio-Visual Emotion Database. *Data Engineering Workshops, 22nd International Conference on*, 2006.
15. M. Pantic and L.J.M. Rothkrantz. Towards an Affect-sensitive Multimodal Human-Computer Interaction. *Proceedings of the IEEE*, pages 1370–1390, 2003.
16. D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.
17. B. Schuller, S. Steidl, and A. Batliner. The INTERSPEECH 2009 Emotion Challenge. In *Proceedings of Interspeech*, page 312315. ISCA, 2009.
18. M. Shami and W. Verhelst. Automatic Classification of Expressiveness in Speech: A Multi-corpus Study. *Speaker Classification II: Selected Projects*, pages 43–56, 2007.
19. S. Steidl. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech.* Logos Verlag, 1 edition, 2009.
20. K. P. Truong and S. Raaijmakers. Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features. In *Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction (MLMI)*, pages 161–172. Springer-Verlag, 2008.
21. L. Vidrascu and L. Devillers. Anger Detection Performances Based on Prosodic and Acoustic Cues in Several Corpora. In *LREC 2008*, 2008.
22. B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. In *Proceedings of Interspeech*, 2007.
23. T. Vogt and E. Andre. Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. In *IEEE International Conference on Multimedia and Expo*, pages 474–477, July 2005.