VOICE-BASED SADNESS AND ANGER RECOGNITION WITH CROSS-CORPORA EVALUATION

Orith Toledo-Ronen and Alexander Sorin

IBM Research – Haifa, Haifa University Mount Carmel, Haifa 31905, Israel
Email: {oritht,sorin}@il.ibm.com

ABSTRACT

Real-life scenarios often require detection of few target emotional categories under a high mismatch between training and operation conditions. We present results of a study on sadness and anger detection with cross-corpora evaluations using two publically available databases. We demonstrate the influence of the mismatch on the detection accuracy comparing cross-corpora results to a single test corpus cross-validation results. We introduce the methodology of representing the broad complementary category by a number of hidden classes. We show performance improvements in sadness and anger detection by using the hidden-classes approach in both cross-corpora and single-corpus evaluations. We explore feature subset selection achieving further improvement in the cross-corpora settings.

Index Terms— emotion recognition, cross-corpora, mismatch, feature selection, eNTERFACE

1. INTRODUCTION

Extracting vocal biomarkers for medical diagnosis and monitoring applications is an emerging and active research area with big potential for medical conditions such as Alzheimer disease, depression, Attention Deficit Hyperactivity Disorder, Parkinson disease, Schizophrenia, and pathologies related to the speech production organs (e.g. pathologies of the vocal folds) [1-6].

In a recent European project, called Dem@Care [7], we are dealing with diagnosis and assessment of people with Dementia using multiple sensing devices including audio, video, and physiological sensors. The audio analysis part aims mainly to detect the presence and state of dementia based on its manifestation in the human voice along three axes: 1) impact of dementia-specific cognitive deficit on the voice; 2) certain mood states typical for dementia patients; 3) certain impairments of the neuromuscular mechanism of speech production. In this work, we focus on the second aspect of the audio analysis. We believe that voice-based emotion detection can play an important role in complementing other vocal biomarkers for diagnosis and monitoring these medical conditions. In the context of the Dem@Care project, and similar foreseen applications, we try to detect certain states of mood of a subject. Apathy is the most relevant mood state, while irritability or aggressiveness is also of interest. We translate this task to the detection of two basic emotional states: sadness and anger, and we try to distinguish between sadness vs. other emotions and sadness vs. anger vs. other emotions. The goal of this work is to explore the applicability of state-of-the-art emotion recognition techniques to the problem outlined above using publically available data corpora.

The rest of the paper is organized as follows. In Section 2 we describe our experimental evaluation methodology and the relation of our work to prior art. In Section 3 we describe the experimental setup including the data sets, the feature set, the classifier, and performance measures. Then, in Section 4, we present the experimental results and we conclude with a discussion in Section 5.

2. EVALUATION METHODOLOGY

Lack of annotated corpora containing real-life emotional speech and the difficulties in collecting such data are widely acknowledged. Typically, a system used in a real-life application would be developed and trained mainly using existing corpora of artificial emotional speech. Hence, we expect high level of mismatch between training and operation conditions. One way to increase the amount of training data is to gather data from a number of corpora. As a result, the training data will be heterogeneous in terms of its emotion categories and their interpretation by speakers, and the training data sources may differ in their recording conditions, language, and the type of emotions produced (acted, elicited, or spontaneous).

To simulate the above situation, we focus on cross-corpora (CC) evaluation where two databases from two independent sources are used, one for training and the other for testing. This is a challenging setup due to the mismatch between the training and testing conditions and has recently become an active research area.

In addition to the CC evaluation, we also perform cross validation (CV) experiments on the test corpus itself, as done in previous studies. The CV is expected to yield more optimistic results because the mismatch between the training and test data is significantly reduced when training and

testing are performed on data from the same corpus. We do the CV experiments for assessing the gap between the in-lab performance and the predicted real-life performance and for testing some ideas.

In relation to prior art, our work continues other recent studies [8-13] in the emerging area of cross-corpora emotion detection. Some of these studies used the same test database that we use along with other corpora, but focused on arousal and valence recognition. Others are focused on feature selection, feature normalization, and data fusion techniques. The focus in our study is on performance optimization for the practical problem of sadness/anger detection and comparison to the single-corpus cross-validation results.

3. EXPERIMENTAL SETUP

The training data set used in the CC evaluations is derived from the Emotional Prosody Speech and Transcripts (EPST) database [14] available from the Linguistic Data Consortium. This corpus contains acted emotional speech. It consists of 2207 utterances with neutral content in English, produced by 8 actors with 15 emotional categories including the neutral class. We use only the lip microphone channel in our data set. The audio part of the eNTERFACE 2005 audio-visual database [15] is used as the test set in the CC evaluations and for the CV evaluations. This database contains a set of pre-defined text sentences in English. The emotional speech is elicited as a reaction to a short story expected to induce a particular emotion. The corpus consists of 1277 utterance from 43 subjects with 6 basic emotions. No neutral data is included in this data set. The same version of this database was used in the study [16].

For feature extraction, we used the openSMILE toolkit [17]. Our setup is based on the emobase2010 feature set containing 1582 features attributed to prosody, spectral envelope and voice quality. We tested some subsets of the low-level descriptors (LLD) and their sentence-level statistics (*functionals*) with the goal of optimizing the emotion detection accuracy in the CC evaluation.

For classification, we used a Support Vector Machine (SVM) classifier with a linear kernel in all our experiments. We used the libSVM software (version 2.82) [18] and normalized each feature to the range [0,1]. In the CC evaluations, the classifier was trained on the EPST data and tested on the entire eNTERFACE database. For the CV evaluation, we followed the leave-one-subject-out (LOSO) scheme in which each fold contains the data from one subject, achieving the desired speaker independence property.

For assessing the classification performance, we used the unweighted average (UA) of the per-class recall or precision rates. The UA performance measure compensates the imbalance between the sizes of the class populations which is significant in the "sadness vs. other" and "sadness vs. anger vs. other" tasks.

4. EXPERIMENTAL RESULTS

As a pre-study step of performance verification, we performed the 6-class LOSO CV evaluation using the emobase2010 feature set. The result is shown in the last line of Table 1. This result is compared to two other relevant evaluations done on the eNTERFACE 2005 database. The first result, as reported by Schuller in [16] and therefore shown in parenthesis, was obtained using a different classifier (SVM with polynomial kernel using SMO) and a slightly different definition of the test set in terms of the audio samples included. The second result in Table 1 was obtained by our system using the Emo_IS09 feature set. The comparison reveals significant advantage of the emobase2010 feature set over the earlier versions and supports our decision to use this feature set and the classifier configuration for our study. The per-class recall rates obtained with the emobase2010 feature set experiment are shown in Table 2 for reference. The abbreviations in the table are used for the six emotional classes of anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), and surprise (SU).

Table 1: Unweighted average recall rate in percent for 6-class LOSO cross validation on the eNTERFACE 2005 database.

Feature Set	# Features	UA
Schuller 2008	1406	(54.2)
Emo_IS09	384	59.6
Emobase2010	**1582**	**65.1**

Table 2: Per-class and UA recall rates obtained with the Emobase2010 feature set for 6-class LOSO cross validation on the eNTERFACE 2005 database.

AN	DI	FE	HA	SA	SU	UA
79.5	61.9	61.4	59.4	71.9	56.3	65.1

4.1. Hidden classes vs. direct classification

Our ultimate goal is to distinguish between one or two target classes (SA or SA/AN) and their complement class referred to as "other" (OT). To establish an optimistic reference we performed two corresponding CV LOSO evaluations. In the SA vs. OT evaluation, the classifier has two classes, one is SA and the other is OT. Class OT is formed by pooling together all the data with labels other than SA. Similarly, in the 3-class case of SA vs. AN vs. OT evaluation, the OT class is composed from all the data labeled differently than SA and AN. Hereafter, this straight forward approach is referred to the direct classification approach. The results of the direct classification evaluations with either two or three classes are presented in the first two lines of Table 3. Note that in the direct classification evaluation, we applied a simple class weighting based on the ratio of class instances to account for the imbalance in the training data between classes due to the large amount of data in class OT.

Comparing the SA and AN recall rates obtained in the direct classification (Table 3) to the ones reported in the 6-class evaluation of Table 2, we hypothesize that the direct classification results can be improved by using the approach hereafter referred to as the hidden-classes classification. In this approach, we perform the classification with the 6 basic classes and then collapse the results of all the classes of no interest into the OT category. The hidden classes in the OT category are transparent to the observer of the final classification results, and confusion errors among the hidden classes within category OT are ignored. The results of hidden classes approach for the SA vs. OT and SA vs. AN. vs. OT tasks are presented at the bottom of Table 3. Note that the multi-class evaluations (3 and 6 classes) are based on the one-versus-one approach implemented in libSVM.

Table 3: Emobase2010 baseline recall rates obtained by LOSO cross validation on the eNTERFACE 2005 database using the direct classification and the hidden classes approach.

#Classes	SA	AN	OT	UA
2 cls	70.0	---	92.0	81.0
3 cls	68.6	74.0	85.4	76.0
6 cls/2 categ	71.9	---	94.5	83.2
6 cls/3 categ	71.9	79.5	86.5	79.3

Comparing to the direct classification results, we observe that the hidden classes approach improved the unweighted average recall rate in the two and three categories evaluations by 3% and 4% respectively. This observation can be explained by that the collection of hidden classes enables better modeling of the wide and diverse data samples associated with the OT category. Also, in the direct classification approach the category representations in the training set is significantly unbalanced because the OT category contains much more samples than the target classes SA and AN. The hidden classes approach removes this effect. Based on the results reported above, we selected the hidden-classes approach as the default technique used in the cross-corpora evaluations below.

4.2. Cross-corpora baseline results

For the CC evaluation, we have to establish a mapping between the 15 emotional classes of the EPST training database and the 6 emotions of the eNTERFACE test corpus. This is not an easy task. We would like to use as much data as possible for training but to have a good match between the emotions in training and testing. The mapping is summarized in Table 4 showing the EPST emotions grouped into six classes in correspondence to the eNTERFACE emotions, and the amount of data per emotion in each corpus. Although the EPST database contains some neutral data, we excluded it because its limited amount of data. We acknowledge that mapping 'shame' to 'fear' or 'interest' to 'surprise' may not be ideal, but since we are focusing on detecting anger and sadness it is not critical.

The cross-corpora results obtained with emobase2010 feature set and the hidden-classes approach are shown in Table 5. As one can see, the performance is very poor not only comparing to the CV results of Table 3 but also in absolute terms. In particular, the OT recall rate is low. The AN recall rate is especially low with most of the instances not classified correctly. Hypothesizing on possible reasons for the poor performance, we realized that the training set is limited in terms of the number of speakers and amount of the data while the emobase2010 feature set contains vast amount of features. This may lead to overfitting of the classifier to the training data, which explains the dramatic performance gap between the CV and CC baseline results. This hypothesis raises an idea to look for a feature subset which will reduce the overtraining effect and improve the generalization ability of the system.

Table 4: Emotional classes mapping between the EPST and the eNTERFACE corpora and the number of instances of each emotion class in each corpus.

EPST		eNTERFACE		Cls
cold anger, hot anger, contempt	454	anger	215	AN
disgust	179	disgust	215	DI
anxiety, shame, panic	435	fear	215	FE
happy, pride, elation	472	happiness	207	HA
sadness, despair, boredom	490	sadness	210	SA
interest	177	surprise	215	SU
Total	**2207**		**1277**	

Table 5: Emobase2010 baseline recall rates for the 2 and 3 category hidden-classes approach in the cross-corpora evaluation.

#Classes	SA	AN	OT	UA
6 cls/2 categ	79.0	---	37.1	58.1
6 cls/3 categ	79.0	2.8	33.7	38.5

4.3. Feature subset selection

Feature selection is important for within-domain classification and even more important for cross-domain or cross-corpora evaluations. For example, in [8], Bone et. al. present a robust set of prosodic features for arousal detection, which are coherent across corpora. The small feature set consists of intensity and pitch attributes. The method used in that study is based on scoring a Gaussian model built on neutral data and a speaker followed by scores fusion using rank-correlation weighting. The results in [8] show gain compared to another approach of using a large feature vector without normalization as in [10].

In order to improve the generalization ability of our classifier across the different corpora, we tested various subsets derived from the emobase2010 feature set, as outlined below. First, a dedicated analysis revealed that C0

and C1 components of the MFCC features degrade the cross-corpora performance. C0 and C1 reflect the average spectral level and tilt respectively. Therefore, their values are influenced significantly by the recording channel characteristics. Apart from that, in the EPST database many anger and panic samples are overemphasized and contain shouting, which especially affects the spectral tilt and thus influences C1 values. Hence we excluded C0 and C1 components from the MFCC features. Secondly, we changed the set of the statistical functionals defined on the LLDs of the emobase2010. Our new and reduced set of selected features consists of the following statistical functionals: extreme relative distances (e.g. range), standard deviation, skewness, kurtosis, linear regression errors, quartiles' differences, percentiles, rising and up-level timings. The rationale behind this modification was to exclude mean values that tend to reflect speaker identity and utterance contents and may appear not representative on a small training set. The results of the 2-category hidden-classes CC evaluations obtained with different LLD subsets of the modified feature set are shown in Table 6. Among all the subsets tested, the best accuracy is achieved using the LLD set containing F0 and the reduced MFCC features without temporal derivatives, with a total of 287 features.

Table 6: Cross-corpora 2-category hidden-classes evaluation recall rates for several feaure subsets of the modified emobase2010 feature set.

Feature Type	# Feat	SA	OT	UA
Loudness,shimmer	64	79.5	37.3	58.4
F0	66	65.2	71.0	68.1
F0,voicing	100	67.1	63.2	65.2
F0,jitter	126	73.3	60.3	66.8
F0,jitter,loudness, shimmer	190	71.4	52.9	62.1
F0,Mfcc-no deriv	287	76.7	60.5	68.6
F0,loudness, Mfcc-no deriv	321	72.9	60.9	66.9
F0,loudness, Mfcc	542	87.1	37.3	62.2
All	1244	81.9	35.7	58.8

4.4. Hidden vs. direct classification for selected features

To complete the study and assess the effect of the hidden-classes approach, we now compare the performance of 2 and 3 categories CC evaluations using the best features subset in both hidden-classes and direct classification. The results are summarized in Table 7, where per-category and unweighted average recall and precision rates are presented.

From Table 7 one can see that the recall for the category OT is consistently and significantly higher for the hidden-classes method than in direct classification. As a result, the precision of the SA and AN categories detection were improved with the cost of certain decrease in the recall rates for these classes. It is explained by that the hidden-classes approach provides a sharper model for the underlying classes within the OT category relatively to the direct classification, which uses data pooling from different classes to build a single and broader model. Thus, the hidden-classes approach would potentially reduce the false positive rate in a sadness and anger detection application. Overall, the hidden-classes method yielded 7% and 3% improvement of the unweighted average recall and precision respectively in the 2-category evaluation while it did not affect the results of the 3-category evaluation.

Comparing the last unweighted average rate figures to the baseline results shown in Table 5, we observe a significant accuracy improvement of 18% and 27% for the 2 and 3 category evaluations respectively. This improvement is attributed to the overtraining reduction due to the proper narrowing of the feature set. Still the cross corpora results of Table 7 exhibit a huge performance gap relatively to the cross validation results presented in Table 3. In our view the cross validation results are not representative of the true accuracy level achievable in a real-life application.

Table 7: Cross-corpora recall and precision rates for hidden-classes and direct classification with 2 classes (SA,OT) and 3 classes (SA,AN,OT) using the selected feature set with F0 and MFCC-no derivatives (287 features).

	Recall				Precision			
#Cls	**SA**	**AN**	**OT**	**UA**	**SA**	**AN**	**OT**	**UA**
2 cls	88.1	---	40.5	64.3	22.6	---	94.5	58.5
3 cls	81.0	37.2	28.8	49.0	25.0	29.7	74.9	43.2
6 cls/ 2 categ	76.7	---	60.5	68.6	27.7	---	92.9	60.3
6 cls/ 3 categ	76.7	28.8	41.2	48.9	27.7	30.4	71.5	43.2

5. CONCLUSIONS

Real-life applications often require detection of few target emotional categories under a high mismatch between training and operation conditions. To simulate this setup, we studied the detection of sadness and anger in cross-corpora evaluations using two publically available databases. We demonstrated the influence of the training-test mismatch on the detection accuracy comparing the cross-corpora results to the single test corpus results. We introduced the approach of representing the broad complementary category by multiple hidden-classes. Our experiments revealed the accuracy gain achieved by the hidden-class approach and a proper feature subset selection. Referring to the absolute accuracy level achievable with the state-of-the-art open source tools, we acknowledge that more work is required for matching the real-life application performance requirements.

6. ACKNOWLEDGEMENTS

This work is supported by the Dem@Care FP7 project, partially funded by the EC under contract number 288199.

8. REFERENCES

[1] Juan J. G. Meilán, Francisco Martínez-Sánchez, Juan Carro1, José A. Sánchez, and Enrique Pérez, "Acoustic Markers Associated with Impairment in Language Processing in Alzheimer's Disease", *The Spanish Journal of Psychology*, Vol. 15, No. 2, 487-494, 2012.

[2] Douglas E. Sturim, Pedro A. Torres-Carrasquillo, Thomas F. Quatieri, Nicolas Malyska, and Alan McCree, "Automatic Detection of Depression in Speech Using Gaussian Mixture Modelling with Factor Analysis", in Proc. *INTERSPEECH* 2011, p. 2981-2984.

[3] Hamdan AL, Deeb R, Sibai A, Rameh C, Rifai H, Fayyad J., "Vocal characteristics in children with attention deficit hyperactivity disorder", *Journal of Voice*. 2009 Mar;23(2):190-4.

[4] T. Bocklet, E. Nöth, G. Stemmer, H. Ruzickova, and J. Rusz, "Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis", in Proc. *ASRU* 2011, pp.478-483.

[5] Thakore J, Rapcan V, D'Arcy S, Yeap S, Afzal N, Reilly RB., "Acoustic and temporal analysis of speech: a potential marker for Schizophrenia", *International Clinical Psychopharmacology*, vol. 26, p. e131, 2011.

[6] Z. Kons, A. Satt, R. Hoory, V. Uloza, E. Vaiciukynas, A. Gelzinis and M. Bacauskiene, "On Feature Extraction for Voice Pathology Detection from Speech Signals", in Proc. *Afeka-AVIOS Speech Processing Conference,* Tel Aviv, Israel, 2011.

[7] The Dem@Care FP7 EU project. Dementia Ambient Care: Multi Sensing Monitoring for Intelligent Remote Management and Decision Support. http://www.demcare.eu/.

[8] D. Bone, C.-C. Lee, S. S. Narayanan, "A Robust Unsupervised Arousal Rating Framework using Prosody with Cross-Corpora Evaluation", in Proc. *INTERSPEECH* 2012.

[9] R. Sun, E. Moore II, "A Preliminary Study on Cross-Databases Emotion Recognition using the Glottal Features in Speech", in Proc. *INTERSPEECH* 2012.

[10] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? ", in Proc. *INTERSPEECH* 2011.

[11] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting Training Data for Cross-Corpus Speech Emotion Recognition: Prototypicality vs. Generalization", in Proc. *Afeka-AVIOS Speech Processing Conference,* Tel Aviv, Israel, 2011.

[12] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition", in *Proc. Automatic Speech Recognition and Understanding Workshop,* Big Island, HY, USA, pp. 523–528, 2011.

[13] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments", in *Proc. 7th Intern. Conf. on Language Resources and Evaluation,* Valletta, Malta, 2010.

[14] M. Liberman, et al. *"Emotional Prosody Speech and Transcripts",* Linguistic Data Consortium, Philadelphia, 2002. (http://www.ldc.upenn.edu/Catalog/LDC2002S28.html)

[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database", In Proc. *IEEE Workshop on Multimedia Database Management*, Atlanta, 2006.

[16] B. W. Schuller, "Speaker, Noise, and Acoustic Space Adaptation for Emotion Recognition in the Automotive Environment", in Proc. *8th ITG Conference on Speech Communication*, 2008.

[17] F. Eyben, M. Wöllmer, B. Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", in Proc. *ACM Multimedia (MM), ACM*, Firenze, Italy, 2010.

[18] Chang, C.-C. and C.-J. Lin. *"LIBSVM: a library for support vector machines"*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.