On the Influence of Phonetic Content Variation for Acoustic Emotion Recognition

Bogdan Vlasenko[1], Björn Schuller[2], Andreas Wendemuth[1], and Gerhard Rigoll[2]

[1] Cognitive Systems, IESK, Otto-von-Guericke University, Magdeburg, Germany
{Bogdan.Vlasenko,Andreas.Wendemuth}@ovgu.de
[2] Institute for Human-Machine Communication, Technische Universität München, Germany
{Schuller,Rigoll}@tum.de

1 Introduction

Today's approaches to the acoustic recognition of emotion ignore the spoken textual content by using one general model per emotion (see [Batliner, 2006]). Considering that many features highly depend on phonetic structure, such as spectral and cepstral features which have become very popular recently [Batliner, 2006], the question arises if this is the optimal way of acoustic modeling. We therefore aim at answering the question how strongly spoken content variance influences emotion recognition performance, herein. Models trained specifically on the unit at hand could then be considered in future engines to improve on accuracies. This would require a combination with an Automatic Speech Recognition (ASR) engine to pick the right unit-specific emotion models at a time. However, several works already demand for ASR inclusion, e.g. for word-boundary detection (see [Schuller, 2006]). In this context we report results considering specific models vs. general models to demonstrate the amount of dependence of acoustic emotion recognition on phonetic transcription of utterence.

The paper is structured as follows: in sect. 2 we introduce the databases, in sect. 3 and 4 spoken content influence on the turn and on the word level.

2 Acted and Spontaneous Data

To demonstrate the influence of spoken content variation on acted and spontaneous data, we decided first for the popular studio recorded Berlin Emotional Speech

Database (EMODB) [Burkhardt, 2005], which covers the 'big six' emotion set (MPEG-4) besides boredom instead of surprise, and added neutrality. 10 (5f) professional actors speak 10 German emotionally undefined sentences. 494 phrases are marked as min. 60% natural and min. 80% assignable by 20 subjects. 84.3% accuracy are reported for a human perception test.

Secondly, we selected the Speech Under Simulated and Actual Stress (SUSAS) database [Hansen, 1997] as a reference for spontaneous recordings. Here, speech is partly masked by field noise. It consists of five domains, encompassing a wide variety of stresses and emotions. We decided for the 3,663 actual stress speech samples recorded in subject motion fear and stress tasks. 7 speakers, 3 of them female, in roller coaster and free fall actual stress situations are contained in this set. Two different stress conditions have been collected: medium stress, and high stress. Within the further samples also neutral samples, fear during freefall and screaming are contained as classes. SUSAS samples are constrained to a 35 words vocabulary.

3 Text Dependence on the Turn Level

We first investigate the influence of spoken content variation on the turn level. At this level we use frame-level features: speech input is processed using a 25ms Hamming window, with a frame rate of 10ms. Next, we employ a 39 dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. Cepstral Mean Subtraction (CMS) and variance normalization are applied to better cope with channel characteristics. Classification is carried out with GMM as described in [Schuller 2007b]. The priors are chosen as an equal distribution among emotion classes.

Test runs on EMODB and SUSAS for utterance models are carried out speaker independently by Leave-One-Speaker-Out (LOSO) evaluation. Table 1 reports average among all speakers and all utterances accuracies for three cases to address text independent (TI) evaluation. A total of 10 different utterances are found in EMODB and 35 in SUSAS, respectively. We included all utterances from training set for general model training. In other cases we left out all samples with target or non-target utterance from tarining set.

Table 1. Mean Accuracies for turn-level modeling on EMODB and SUSAS. Frame-level features with GMM, LOSO evaluation.

Accuracy [%]	**EMODB**	**SUSAS**
General model	**77.1**	**46.0**
Non-target utterance left out	**75.9**	**45.4**
Target utterance left out	72.7	44.2

From Table 1. it is clear that removal of target utterance from training set fundamentally reduce accuracy of emotion recognition in comparison with removal non-target utterance. Random removal non-target utterances preserves the context, which results in higher accuracy than removing the target utterance, which makes the training data context-independent.

4 Text Dependence on the Word Level

Second, we investigate the influence of spoken content variation on the word level. Therefore we use a different strategy to cover another typical approach to acoustic modeling in emotion recognition from speech: a state-of-the-art brute-force feature generation by projection of a typical prosodic, spectral and voice quality low-level-descriptors (LLD) onto a static feature vector by statistical functionals (see [Schuller, 2006]). The obtained 1,406 dimensional feature vector is classified by SVM with polynomial Kernel and SMO learning [Witten, 2000].

73 different words are found in EMO-DB of which we select only those that have a minimum frequency of occurrence of 3 within each emotion. This comprises a total of 41 words with roughly 200 instances per word. Within an equivalent selection process we picked the according 11 highest frequency terms from SUSAS out of a total of 35.

Table 3. Accuracies for word-level modeling in matched and mismatched condition compared to general models at diverse relative sizes of training corpora on EMODB and SUSAS. tsf *abbreviate training size factor*. Static features with SVM, LOSO.

Accuracy [%]	EMODB	SUSAS
matched	**48.9**	**60.7**
mismatched	37.4	54.2
tsf 1%	43.1	50.6
tsf 2%	44.8	56.1
tsf 5%	49.1	60.7
tsf 10%	51.7	61.5
tsf 100%	**55.5**	**64.7**

Table 3 visualizes the results obtained on these two corpora: first, matched vs. mismatched conditions are analyzed, whereby mismatching is an average of the accuracy of all selected words in a corpus was computed, when the emotion models were taken from all other words. Spoken content clearly does influence accuracy throughout word-model comparison, as can be seen by the mean accuracy in table 1.

We next address the question how a general model trained on any word in the corpus – the common state-of-the-art –performs in relation to the amount of training data available by the relative training size factor (tsf). Random down-sampling preserving class-balance is used. Noting that every word will occur with an average frequency of 2.5% in the corpora, it can be seen that a general model *with that tsf* will perform between matched and mismatched models. The general model will outperform the matched case already at *tsf=10%*.

5 Discussion

The results presented in this work clearly demonstrate dependence of emotion models on the spoken phonetic content for both, acted and spontaneous emotions, and

employing the two typical types of emotion recognition engines (1.4k large-feature-space turn-level SVM and MFCC space frame-level HMM/GMM) (see [Vlasenko, 2007]).. In future works we therefore aim at investigation how this could be exploited by use of unit-specific models.

Acknowledgements

The work has been conducted in the framework of the NIMITEK project (Sachsen-Anhalt Federal State funding) FKZ XN3621H/1005. This project is associated and supported by the Magdeburg Center for Behavioral Brain Sciences (Neuroscience Excellence Cluster). Bogdan Vlasenko acknowledges support by a graduate grant of the Federal State of Sachsen-Anhalt.

References

[Batliner, 2006] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining Efforts for Improving Automatic Classification of Emotional User States. In: Proc. 1st Int. Language Technologies Conference IS-LTC 2006, Ljubljana, Slovenia (2006)

[Burkhardt, 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. INTERSPEECH 2005, pp. 1517–1520 (2005)

[Hansen, 1997] Hansen, J.H.L., Bou-Ghazale, S.: Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In: Proc. EUROSPEECH 1997, Rhodes, Greece, vol. 4, pp. 1743–1746 (1997)

[Schuller, 2006] Schuller, B., Rigoll, G.: Timing Levels in Segment-Based Speech Emotion Recognition. In: Proc. INTERSPEECH 2006, pp. 1818–1821 (2006)

[Schuller, 2007] Schuller, B., Vlasenko, B., Minguez, R., Rigoll, G., Wendemuth, A.: Comparing One and Two-Stage Acoustic Modeling in the Recognition of Emotion in Speech IEEE ASRU 2007, pp. 596–600 (2007)

[Vlasenko, 2007] Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G.: Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech 2007. In: Proc. INTERSPEECH 2007, pp. 2225–2228 (2007)

[Witten, 2000] Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations, p. 133. Morgan Kaufmann, San Francisco (2000)

[Young, 2002] Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK-Book 3.2. Cambridge University Press, Cambridge (2002)