Towards Robust Spontaneous Speech Recognition with Emotional Speech Adapted Acoustic Models

Bogdan Vlasenko, Dmytro Prylipko, and Andreas Wendemuth

Cognitive Systems, IESK & Center for Behavioral Brain Sciences,
Otto von Guericke University, D-39016 Magdeburg, Germany
bogdan.vlasenko@ovgu.de

Abstract. Speech signal in addition to the linguistic information contains additional information about the speaker: age, gender, social status, accent (foreign accent, dialects, etc.), emotional state, health etc. Some of these informational channels induce changes of the speech acoustic characteristics. This article presents evaluation of the ASR acoustic models (first trained on neutral, read speech) on acted and spontaneous emotional speech. In our research we used adaptation approaches to compensate the mismatch of acoustic characteristics between neutral speech samples and affective speech material. During experiments we observed that the affective-speech-adapted ASR acoustic models provide better emotional-speech-recognition performance. The improvements of affective speech recognition performance were 6.24% absolute (7.1% relative) for speaker-independent evaluations on the EMO-DB database and 7.08% absolute (25.43% relative) for cross-corpora evaluation on the VAM database.

Keywords: Emotional Speech, Adaptation, ASR

1 Introduction

The speech signal comprises not only linguistic content but also various additional information about the speaker: *age, gender, social status, accent, emotional state, health* etc. Characterization of the influence of some of these speech signal variations, together with related methods to improve automatic speech recognition (ASR) performance, is an important research field. In order to deal with spontaneous speech we should not cut the above mentioned information channels from the input signal, but use them as an additional knowledge source and thus boost the performance.

In real-life applications training and evaluation conditions (speaking rate, acoustic environment, vocal tracts variability, affected state etc.) usually do not match, which cause a severe degradation of the recognition performance. In our previous research [7] we characterized acoustical difference between emotional

S. Wölfl (Ed.): Poster and Demo Track of the 35th German Conference on
Artificial Intelligence (KI-2012), pp. 103-107, 2012.

and neutral speech. We have shown a significant difference between vowel triangles form and their position in F1-/F2-dimensional space for emotionally colored and neutral speech samples. This difference illustrates why ASR models trained on neutral speech are not able to provide a reliable performance for affective speech recognition.

To compensate such a mismatch, acoustic models' adaptation techniques are usually applied. However, these techniques are usually employed to compensate the mismatch of acoustic characteristics between various speaker, acoustic channels and noisy environments. Acoustic models' adaptation towards affective speech is a less popular adaptation concept.

In our research, we used adaptation approaches to compensate the mismatch of acoustic characteristics between neutral speech samples and affective speech material. We used acted affective speech samples from the popular public available database EMO-DB to adapt acoustic models trained on emotionally neutral speech samples from The Kiel Corpus of Read Speech.

2 Corpora

For initial acoustic models training we used a part of The Kiel Corpus of Read Speech [5] which contains emotionally neutral German read speech samples. For our evaluation we used speech samples from 1041 utterances produced by 6 female and 1033 utterances spoken by 6 male speakers.

For affective speech we decided to use the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [2] and The Vera am Mittag (VAM) corpus [4]. The EMO-DB contains acted emotional speech samples. 10 professional actors (5 male and 5 female) spoke 10 German sentences with emotionally neutral linguistic meaning. For our evaluations we used 494 sentences classified as more than 60% natural and at least 80% clearly assignable throughout perception tests.

The VAM database [4] consists of 12 hours of audio-visual recordings taken from a German TV talk show. The corpus contains 947 utterances with spontaneous emotions from 47 guests of the talk show which were recorded from unscripted, authentic discussions. Since VAM corpus does not provide such a lexicon, we created it by ourselves using two ways. The major part of the word transcriptions (1216 words) has been taken from other German corpora, namely Verbmobil and SmartKom. For the rest (688 words) we created transcriptions using a grapheme-to-phoneme conversion with *Sequitur* G2P converter [1]. The converter was trained on a joined lexicon based on SmartKom and Verbmobil lexicons (12460 German words at all).

3 Emotional adaptation of acoustic models

For our evaluations we used the HTK toolkit to create and test continuous density hidden Markov models (HMMs) based on a multivariate Gaussian mixture model (GMM) with 32 mixture components. We created *left-to-right* mono-

phone models with three emitting states for acoustic modeling. Speech input is processed using a 25 ms Hamming window, with a frame rate of 10 ms. We employed 39-dimensional MFCC feature vectors (12 cepstral coefficients + log frame energy plus speed and acceleration coefficients).

3.1 Adaptation configuration

Two adaptation schemes have been tested: *Maximum Likelihood Linear Regression (MLLR)* and *Maximum a Posteriori (MAP)* [8]. During the adaptation only the mean values of Gaussian mixture were updated because variance compensation provides only minor imrovement and requires additional computational overhead of non-diagonal Gaussian likelihood calculations [3]. Prior to adaptation and recognition on VAM, optimal parameters for each scheme should be determined. For MLLR a number of regression classes is important. MAP depends on the τ parameter (weight of the prior knowledge).

For the MLLR, regression trees with $2, 4, 8, 16$ and 32 terminal nodes have been tested. Prior knowledge weight for MAP has been evaluated in range of $\tau = 2, ..., 20$. For MLLR adaptation the best emotion recognition performance on EMO-DB samples has beed obtained with 32 regression class trees ($rc = 32$). These configurations have been used further for adaptation and test on the VAM corpus. For MAP adaptation the best emotion recognition performance has been obtained with $\tau = 2$ (see Table 1).

Table 1. Optimal adaptation parameters selection. Basic models trained on Kiel, adapted and evaluated with LOSO on EMO-DB

Acoustic model	**Parameters**	**Word accuracy** [%]
Non-adapted basic		88.06
MLLR-adapted basic	rc=32	93.67
MAP-adapted basic	τ=2	94.30
EMO-DB trained		96.70

As one can see from Table 1, HMM/GMM models trained on neutral speech samples from the Kiel dataset are not able to provide acceptable emotional-speech-recognition performance without adaptation on affective speech samples. For this configuration (72 words in lexicon, only 10 possible sentences) state-of-the-art recognition accuracy is higher than 95% [6]. However, MLLR-adapted basic models provide better recognition performance, which is close to the value achieved during the evaluation of EMO-DB-trained (*native*) acoustic models.

3.2 Experiments and results

Prior to adaptation we tested the baseline performance of the acoustic models trained on the Kiel corpus. Except acoustic models, other components such as

lexicon or language models have been taken from the test database. Training and testing on the VAM (single corpora mode) has been done in a speaker independent fashion using Leave-One-Speaker-Group-Out (LOSGO with 5 speaker groups at all) strategy.

Table 2. Word accuracy rates for cross-corpora evaluation of acoustic models with and without pre-adaptation on EMO-DB samples, evaluated on the VAM database

Training set	Adaptation scheme	Evaluation set	Word accuracy [%]
Kiel	–	VAM	27.84
Kiel	MLLR on EMO-DB	VAM	34.92
Kiel	MAP on EMO-DB	VAM	33.54
VAM	–	VAM	42.75

The results presented in Table 2 show that training ASR models on neutral speech, and subsequent adaptation on affective speech samples, does have an impact on the recognition performance within emotional speech recognition. These results have been obtained after evaluations in a cross-corpora way. We used speech samples from the Kiel and EMO-DB databases for training and adaptation purposes, respectively. Finally, these acoustic models have been evaluated on the VAM database speech samples.

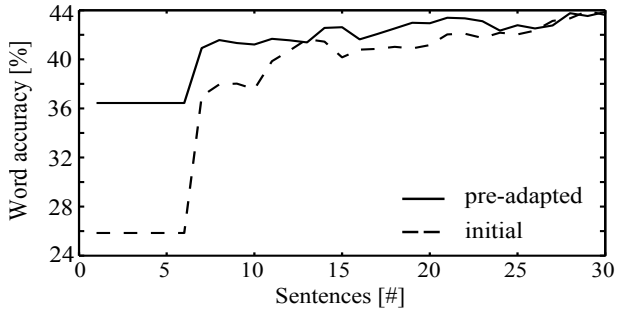

Fig. 1. Performance evolution during the incremental unsupervised adaptation on VAM database. Initial models trained on only Kiel samples, pre-adapted - trained on Kiel and adapted on EMO-DB samples.

Also, we compared initial acoustic models with pre-adapted ones with unsupervised incremental MLLR adaptation. 30 adaptation sentences were selected randomly from the whole VAM corpus. During the adaptation process they were fed to HVite sequentially. Other 917 sentences formed the test set. This difference in procedure is the reason why initial values of both curves depicted in Fig. 1 slightly differ from the values provided in Table 2. The transformations were applied after some number of frame occurrences (namely 800) which in our case corresponds to 6 sentences or 6.52 seconds of speech. One can see from Fig. 1, if

we do not have at least 25 sentences for unsupervised adaptation, pre-adapted acoustic models provide much better speech recognition performance.

4 Discussion and conclusions

The main issue of this research is to show that training ASR models on neutral speech, and subsequent adaptation on affective speech samples, does have an impact on the recognition performance within emotional speech recognition. It has been found that the adaptation on acted emotional speech samples favor a significant gain (about 25.43% relative improvement for word-accuracy rate) in spontaneous emotional speech recognition performance (34.92% with adapted models) over the basic ASR models trained on neutral speech samples. In comparison to results presented for the EMO-DB database, speech recognition performance for the VAM database obtained with adapted models is relatively low. This result can be compared with a low word-accuracy rate of 42.75% obtained during speaker-independent LOSGO evaluation on the VAM database.

Comparison of these values to the state-of-the-art speech recognition performance is unfortunately hardly possible, due to the nature of the corpora. Both EMO-DB and VAM were designed for research in emotion recognition from speech, rather than for speech recognition. That is why most publications report on accuracies in emotion classication. For our best knowledge, there is no paper reporting on the accuracies of speech recognition on EMO-DB or VAM.

As a conclusion, we showed that acoustic models trained on read speech samples and adapted to acted emotional speech could provide better performance of spontaneous emotional speech recognition.

References

1. M. Bisani and H. Ney. Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Communication*, 50(5):434–451, May 2008.
2. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. A database of German emotional speech. In *Proc. of EUROSPEECH*, pages 1517–1520, 2005.
3. M. Gales, D. Pye, and P. Woodland. Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation. In *Proc. of ICSLP*, pages 1832–1835. IEEE, 1996.
4. M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German audio-visual emotional speech database. In *Proc. of ICME*, pages 865–868, 2008.
5. K. J. Kohler. Labelled data bank of spoken standard German - the Kiel Corpus of read and spontaneous speech. In *Proc. of ICSLP*, pages 1938–1941, 1996.
6. D. Pallett. A Look at NIST's Benchmark ASR tests: Past, Present, and Future, 2003.
7. B. Vlasenko, D. Prylipko, D. Philippou-Hübner, and A. Wendemuth. Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In *Proc. of Interspeech*, Florence, Italy, 2011.
8. P. Woodland. Speaker adaptation for continuous density HMMs: A review. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, pages 11–19, Antipolis, France, 2001.