Affect-Robust Speech Recognition by Dynamic Emotional Adaptation

Björn Schuller, Jan Stadermann[1] *& Gerhard Rigoll*

Institute for Human-Machine Communication
Technische Universität München, Germany
{Schuller; Stadermann; Rigoll}@tum.de

Abstract

Automatic Speech Recognition fails to a certain extent when confronted with highly affective speech. In order to cope with this problem we suggest dynamic adaptation to the actual user emotion. The ASR framework is built by a hybrid ANN/HMM mono-phone 5k bi-gram LM recognizer. Based hereon we show adaptation to the affective speaking style. Speech emotion recognition takes place prior to the actual recognition task to choose appropriate models. We therefore focus on fast emotion recognition based on low extra feature extraction effort. As databases for proof-of-concept we use a single digit task and sentences from the well-known WSJ-corpus. These have been re-recorded in acted neutral and angrily speaking style under ideal acoustic conditions to exclude other influences. Effectiveness of acoustic emotion recognition is also proved on the SUSAS corpus. We finally evaluate the need of adaptation and demonstrate significant superiority of our dynamic approach to static adaptation.

1. Introduction

Great efforts have been spent to ensure high performance of Automatic Speech Recognition (ASR) in view of adaptation to noise or speaker characteristics so far. However, among disruptive factors in view of high accuracies also speaking style diversities can be found, as altered speech due to the Lombard effect or speech under stress and emotional influences as anger or sadness [1]. It is a known fact that such affective speech in general downgrades recognition performance of speech or speaker recognition tasks [2, 3]. Throughout this contribution we therefore deal with adaptation to emotional speaker states for affect-robust speech recognition.

As a starting point we analyze emotional speech data dealing with the question whether a performance loss is measurable and significant. Thereby we employ a powerful hybrid ASR engine built up by an Artificial Neural Network (ANN) for phoneme probability estimation profiting from discriminative training abilities in combination with a Hidden-Markov-Model (HMM). Next, it has to be answered whether recognition of the actual underlying affect is necessary aiming at dynamic adaptation, or static adaptation by inclusion of emotional models once suffices.

As data is needed to answer these aspects, we will present two databases collected throughout these works. Firstly, a simple single digit task is considered. Secondly, sentences of the well-known Wall-Street-Journal (WSJ) speech database have been re-recorded in affective speaking styles, and will be released for public access. Likewise, large vocabulary continuous speech recognition is covered as more challenging task.

Since robust emotion recognition is a precondition for dynamic adaptation, effective and robust methods have to be established for speaker independent reliable emotion recognition. Thereby we profit from past expertise in the community [4, 5] and extend our propagated advances so far [6] in view of fast estimation and low extra feature extraction effort, as MFCC and energy are already calculated for speech recognition. These methods shall also be proven effective on a known affective database: the SUSAS corpus. Finally a fully working system that firstly estimates the affect and subsequently adapts to it is evaluated.

The paper is structured as follows: firstly, databases of affective speech are introduced in section 2; afterwards the used ASR engine is described in section 3. Next, we discuss adaptation of the engine to cope with affective speech in section 4. Section 5 deals with emotion recognition based on acoustic features. The contribution ends with results, discussion and conclusion in sections 6 and 7.

2. Databases

For tests and evaluation databases are needed with phonetic transcription of the spoken content and affective and neutral speech under the same recording conditions in view of speaker, noise, microphone, etc. Also, spoken phrases should fit the vocabulary of the ASR engine, and avoid further disruptive influences as unknown background noise.

Under these preconditions we decided to record two databases for this exact purpose: a simple single digit task and continuous speech. The digit task comprises the digits 0-9 in English. For the continuous speech task 26 linguistically affect neutral sentences of the well-known WSJ database [7] were selected out of the set WSJ1-S3-P0. These phrases have an average length of 10.9 words (min. 3 words, max. 18 words). The reason to base the phrases on WSJ sentences is the fact that the acoustic model (AM) of the speech recognizer is trained on WSJ and WSJ is well-known in the speech recognition community. In order to obtain adequate corpus sizes, we decided for acted samples. Though it is disputable whether these are natural, they form a reasonable basis for the aimed at experiments. Recordings have been fulfilled in an anechoic chamber by use of an active condenser microphone AKG 1000S MK II at 48 kHz 16 bit sampling in PCM. Speech data was recorded directly to a hard-disk. The A/D converter had an SNR of 100 dB. For the subsequent test a down-sampling to 16 kHz was performed. All recordings have been performed in random emotion and content order known only to the speaker, to whom the actual text and class label was presented on a screen. After capturing the signal was instantly replayed to two test persons that supervised the recording. Following their phonetic transcription check and emotion labeling, a phrase was kept in case of total

[1] Now with CreaLog GmbH, Munich, Germany

accordance of the speaker and the test subject. In case of disagreement the phrase was saved for later re-recording. In order to avoid anticipation effects the collection was spanned over four weeks. 10 speakers were involved in the dataset creation, 1 of them female. All of the speakers were non-native speakers with excellent English speaking skills with an average age of 26.0 a (min. 23 a, max. 33 a). As emotions anger and neutrality were chosen to keep complexity limited at this time. Each speaker initially passed one hour training.

Likewise, the digits 0-9 have been recorded 50 times per number and emotion, resulting in a total of 1,000 samples named EMO-09 in the ongoing. The WSJ phrases have been collected twice per speaker and emotion, each, resulting in 2x26x10=520 phrases, named EMO-WSJ in the ongoing.

In order to provide results for acoustic emotion recognition on a known public corpus in view of comparability and perform tests on spontaneous emotions we finally selected the Speech Under Simulated and Actual Stress (SUSAS) database [2]. It consists of five domains, encompassing a wide variety of stresses and emotions. We decided for the 3,949 actual stress speech samples recorded in dual-tracking workload or subject motion fear tasks. 4 male speakers in an US Apache helicopter cockpit and 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 within the helicopter situation: *medium stress* during warm-up, where the helicopter is on the ground but running, and *high stress* during flight, where pilots are flying hover, turn and other maneuvers while speaking. Within the further samples also *neutral* samples, *fear during freefall* and *screaming* are contained as classes. Likewise a total of five emotions, respectively speaking styles, are covered. SUSAS samples are constrained to a 35 words vocabulary of short aircraft communication commands. All files are sampled in 8 kHz, 16 bit. The recordings are partly overlaid with heavy noise and background ground controller over-talk. However, this resembles realistic acoustic recording conditions, as also given in many related scenarios of interest as automotive speech interfaces or the mentioned public transport surveillance.

3. Hybrid ANN/HMM ASR Engine

The ASR system consists of a tied-posteriors acoustic model [8] with a multi-layer perceptron (MLP) estimating phoneme posterior probabilities and a set of monophone HMMs. Thereby discriminative training abilities and easily incorporation of seven context frames in an ANN are combined with the warping capabilities of HMM. This concept has been proven highly effective in the past [8].

In a tied-posteriors acoustic model the HMM state likelihood is computed as

$$p\left(\underline{x}|S_i\right) \propto \sum_{j=1}^{J} c_{i,j} \frac{P\left(j|\underline{x}\right)}{P\left(j\right)} \tag{1}$$

having $P(j|\underline{x})$ as a-posteriori phoneme probability of the ANN, $c_{i,j}$ as mixture coefficients of each HMM state and $P(j)$ as prior probability of each phoneme j.

In order to feed the HMM with $P(j|\underline{x})$ phonemes probabilities of the analyzed speech signal are calculated in a lower MLP stage. The acoustic features therefore are 12 MFCCs, the energy and the first and second order derivatives resulting in 39 components per frame, one frame computed every 10 ms. For the non-linearity within the MLP we apply a sigmoid function in the hidden-layer nodes and a softmax function in the output nodes.

The MLP is trained on the WSJ0 [7] speaker independent training set using the back-propagation algorithm. The function to be optimized is the cross entropy function and the topping criterion is an increase of the frame error rate on a cross validation set. Phoneme recognition rate on a word level is 83.24% in average.

Overall, the recognition engine employs a closed vocabulary of 5k terms, mono-phones and a bi-gram language model (LM), trained on WSJ. 8.52% WER is the baseline in this configuration.

4. Emotional Adaptation of Acoustic Models

In order to test whether adaptation of the AM results in an improvement, two methods are viable [9]:

Firstly, adaptation of the ANN by re-training of vital weights in the final layer may be performed. These are tracked by calculation of hidden neurons' variances and search for nodes with such high variance.

Secondly, the mixture-weights $c_{i,j}$ of the HMMs may be adapted by MAP-like strategy [10], whereby β is the learning rate, $\xi_{i,j}$ are the state occupations, and $\gamma_{i,j}$ are the mixture occupations. The latter are estimated from the adaptation data by Baum-Welch iterations:

$$c_{i,j,n+1} = \beta \cdot c_{i,j,n} + \frac{e^{\xi_{i,j}}}{e^{\gamma_{i,j}} + \beta} \tag{2}$$

Both, MLP and HMM adaptation, are fulfilled in a supervised manner.

5. Acoustic Emotion Recognition

As a basis for feature generation we extract low-level contours of a whole phrase. We use the same preprocessing of the audio signal as within the ASR engine: 20 ms Hamming-windowed frames are analyzed every 10 ms. The starting point is a broad feature basis. However, we aim at reduction afterwards, to speed up the process by sparing extraction effort.

For prosodic information we extract contours of elongation, intensity, and intonation and estimate durations of pauses and voiced syllables. Out of the elongation we calculate the zero-crossing-rate. Standard frame energy is used to include intensity information based on physical relations. Intonation is respected by auto-correlation-based pitch estimation. We thereby divide the speech signal correlation function by the normalized correlation function of the window function and search for local maxima besides the origin. Dynamic programming is used to back-track the pitch contour in order to avoid inconsistencies and reduce error from a global point of view. Finally, the named durations are estimated based on intensity considering pause duration, and voiced/unvoiced parts duration for syllable length based on intonation.

In order to include voice quality information we also integrate the location and bandwidth of formants one to seven, harmonics-to-noise-ratio (HNR), MFCC coefficients, and a perception conform dB-corrected FFT spectrum as basis for low-band energies -250 Hz and -650 Hz, spectral roll-off-point, and spectral flux. Formant location and bandwidth estimation is based on resonance frequencies in the LPC-spectrum of the order 18. Back-tracking is used here, as well. The HNR is calculated as logHNR to better model human perception. It also bases on the auto correlation of the input signal. The usage of MFCC for affect recognition is highly discussed, as these tend to depend too strongly on the spoken content. This seems a drawback, as we want to recognize emotion independently of the content. However, they have been proven successful for this task, yet, and are available anyway, as they need to be calculated for the speech recognition itself.

Finally, as articulatory features we use the spectral centroid. Overall, parts of these contours are comprised within the MPEG-7 LLD standard. Likewise, the following methods may be transferred in order to recognize emotion basing on MPEG-7.

In former works we showed the higher performance of derived functionals instead of full-blown contour classification [11]. We therefore use systematic generation of functionals f out of multivariate time-series F by means of descriptive statistics, a common practice in speech emotion recognition [12], on an utterance level:

$$f : F \rightarrow \mathbb{R} \tag{3}$$

First of all the contours are smoothed by symmetrical moving average filtering with a window size of three, to be less prone to noise. Successively, speed (∂) and acceleration (∂^2) coefficients are calculated for each basic contour. Afterwards we compute linear momentums of the first two orders, namely mean, centroid, standard deviation, as well as extrema, turning points and ranges. In order to keep dimensionality within range we decide by expert knowledge which functionals to calculate. Table 1 in section 6 provides a rough overview of calculated functionals.

Besides lower extraction time-effort, reduction of features also often leads to higher classification performance, as the classifier is confronted with less complexity, if only redundant information is spared. In former works [6] we demonstrated the high effectiveness of wrapper-based search of features with the target classifier which aims at optimization of a set as a whole. However, selection of an optimal functional set in general does not spare base contour extraction effort, unless no functional of a base contour is contained within the final set. We therefore focus on evaluation of feature group relevance in the latter sections.

Dealing with classification, the optimal learning method is broadly discussed [4, 5], similar to the optimal features. In [6] we made an extensive comparison including besides Support Vector Machines (SVM) Naïve Bayes, k-Nearest Neighbors, Decision Trees, and Neural Nets. Further more we investigated construction of more powerful classifiers by means of meta-classification as MultiBoosting or Stacking. However, in our experiments SVM prevailed in view of accuracy and effectiveness. We therefore apply these herein.

SVM - kernel machines - are well known in the machine learning community and highly popular at the time due to their remarkable performance and generalization capabilities. The latter result from the applied structural risk minimization oriented training. Generally speaking, SVM base on a linear distance-function classification of a two-class problem. However, multi-class strategies as one-vs.-one, layer-wise decision or one-vs.-all exist. Discriminative training is achieved by optimal placement of a separation hyperplane under the precondition of linear separability. As a consequence, a dual optimization problem has to be solved throughout training process. The precondition of linear separability is approached by a transformation of the original feature space via a kernel function that has to be found empirically. Herein we use a couple-wise one-vs.-one decision for multi-class discrimination and a polynomial kernel found optimal throughout test cycles. For more details on classifiers refer to [13].

6. Results and Discussion

We first show our results for emotion recognition based on acoustic features, as these are crucial for the following results in combination with speech recognition. Table 1, as already mentioned in section 5, shows feature contour and derived functional numbers. The numbers of duration related feature contours are in brackets, as these rely on intonation and intensity. However, the table also shows accuracies obtained by feature groups in a 10-fold stratified cross validation (SCV). Since datasets are sparse in the field of speech emotion recognition, this evaluation method, which allows for training disjunctive test on all samples, is very popular.

Table 1: *Overview of derived acoustic features and group-accuracy within 10-fold SCV with SVM, database EMO-WSJ.*

Group	***F*** **[#]**	***F*+∂+∂²** **[#]**	***f*** **[#]**	**Acc. [%]**
HNR	1	1	3	**57.9**
Duration	(2)	(2)	5	**61.4**
Intonation	1	3	12	**74.6**
Intensity	1	3	11	**77.1**
Formants	14	28	105	**82.9**
FFT based	5	7	17	**86.0**
Elongation	1	1	3	**89.8**
MFCC	15	45	120	**98.5**
Total	**38**	**88**	**276**	**98.1**

As can be seen in the table, MFCC is fortunately the most relevant single feature group. Their accuracy is even higher than use of all features, as the classifier is confronted with too high complexity thereby. However, having the N-best features based on diverse groups outperforms MFCC standalone: As a basis of comparison we also selected the *N* best functionals by SVM Sequential Forward Floating Search (SVM-SFFS), a powerful hill-climbing wrapper-based feature selection method. Thereby the overall maximum of 99.2% for the correct discrimination of angry and neutral sentences could be achieved. However, due to space limitations we cannot name reduced sets functional wisely. Intensity related features also provide a strong basis, and if we combine these only with MFCC related ones setting additionally extracted aside we end up with an accuracy of 98.7% for the database EMO-WSJ. Applying SVM-SFFS only on energy and MFCC related features as calculated for speech recognition leads us to 99.0% accuracy at 40 features. This final set is close to the maximum performance of 99.2% having all features as basis and can be extracted very fast out of the available features.

To manifest this on a well known public speech emotion database, we perform the same experiment on SUSAS. On the spontaneous samples of the SUSAS database we finally achieve 77.8% correct recognition rate within 10-fold SCV and using SVM applying the full feature set for 5 emotions. By feature reduction accuracy is boosted to impressive 84.9% in average. Neutrality is thereby recognized with 76.0%, fear during freefall with 88.6%, medium stress with 82.2%, high stress with 90.6%, and screaming with 97.9% accuracy. Neutral samples are exclusively confused with stress, and mostly with medium stress. However, if we consider only energy and MFCC related features, accuracy drops to 69.9% for five classes. Yet, screaming is not confused with neutral. If we likewise adapt only to screaming, and neutral speech, these features would suffice. Or, if we add high stress as third class, we end up with 82.5%.

Table 2 now shows speech recognition accuracy for the database EMO-09 with no adaptation, adaptation to neutral and angrily speaking style only, and to both emotions. Thereby 2x40 disjunctive digits are used for adaptation in a cross-validation.

Table 2: *Accuracies with diverse static adaptation scenarios, database EMO-09. Adapt. abbreviates adaptation.*

Acc. [%]	**Adapt. -**	**Adapt. N**	**Adapt. A**	**Adapt. A+N**
Neutral (N)	91.4	99.5	71.0	93.0
Anger (A)	47.2	46.0	89.5	90.5

As can be seen, a large gap occurs between neutrally and angrily spoken digits. Adaptation to neutral samples already boosts performance for neutral speaking style. This comes, as the affective adaptation is an adaptation to the acoustic conditions, too, and arguably one to the task, as well. The same is true for adaptation to angrily samples, when testing with such. By anger adaptation the recognition rate is heavily boosted for anger samples. However, diametric adaptation leads to a downgrading in both cases. Static adaptation to both speaking styles results in an overall improvement, but not in the maximum obtainable performance. Table 3 shows this in a more clear way, by analysis over both emotions at a time. It also shows results for dynamic adaptation by integration of the actual emotion recognition prior to the speech recognition with the discussed 40 MFCC and energy related features. Dynamic diametric adaptation shows the worst case risk, if always the exact wrong affect is chosen.

Table 3: *Overall accuracies with diverse adaptation scenarios, database EMO-09. Dyn. abbreviates dynamic, corr. correct, diam. diametric adaptation manner.*

Adaptation	**-**	**static N**	**static A**	**static N+A**	**dyn. corr.**	**dyn. diam.**
Acc. [%]	69.5	72.8	80.3	91.8	**94.5**	58.5

Likewise 2.7% absolute accuracy boost and 32.9% relative error rate reduction can be reported for dynamic adaptation, which is significant at a level of α=0.05.

Tests on continuous speech using the EMO-WSJ set manifest these results: using 2x40 adaptation phrases in cross-manner leads to a significant 1.09% absolute, and 2.55% relative word error rate (WER) improvement for dynamic adaptation compared to static. An absolute WER reduction of 16.59% can thereby be reported for the anger phrases when comparing to no adaptation at all.

7. Conclusions

Within this work we showed that affective speech downgrades recognition performance if training was fulfilled only with emotional neutral phrases. We therefore showed adaptation strategies within a hybrid ANN/HMM ASR framework. A static adaptation to affective speech helps to raise WER, but the maximum performance is only obtained by a dynamic adaptation to the underlying affect. Therefore fast emotion recognition based on available acoustic features could be shown highly effective. A working system that first recognizes emotion, and subsequently adapts to it could be presented resulting in significant WER improvement.

In future works we aim at investigation on larger datasets incorporating more emotions at a time.

8. References

[1] Shriberg, E., 2005. Spontaneous Speech: How People Really Talk And Why Engineers Should Care, *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, 1781-1784.

[2] Hansen, J.H.L.; Bou-Ghazale, S., 1997. Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, *Proc. EUROSPEECH-97*, Rhodes, Greece, vol. 4, 1743-1746.

[3] Klasmeyer, G.; Johnstone, T.; Bänziger, T.; Sappok, C.; Scherer, K. R., 2000. Emotional Voice Variability in Speaker Verification, *Proc. ISCA Workshop on Speech and Emotion: A conceptual framework for research*, Belfast, Ireland.

[4] Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. G., Jan. 2001. Emotion recognition in human-computer interaction, *IEEE Signal Processing magazine*, vol. 18, no. 1, 32–80.

[5] Pantic, M; Rothkrantz, L., Sep. 2003. Toward an Affect-Sensitive Multimodal Human-Computer Interaction, *Proceedings of the IEEE*, vol. 91, 1370-1390.

[6] Schuller, B.; Müller, R.; Lang, M.; Rigoll, G., 2005. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. *Proc. INTERSPEECH 2005*, ISCA, Lisbon, Portugal, 805-809.

[7] Paul, D. B.; Baker, J. M., 1992. The Design for the Wall Street Journal-based CSR Corpus, *Proc. of the DARPA Speech and Natural Language Workshop*, Morgan Kaufmann, Pacific Grove, CA, 357-362.

[8] Stadermann, J.; Rigoll, G., 2003. Comparing NN Paradigms in Hybrid NN/HMM Speech Recognition using Tied Posteriors, *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2003)*, U.S. Virgin Islands.

[9] Stadermann, J.; Rigoll, G., 2005. Two-Stage Speaker Adaptation of Hybrid Tied-Posterior Acoustic Models, *Proc. ICASSP 2005*, Philadelphia, USA.

[10] Gauvain, J.-L. ; Lee, C.-H., 1994. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Transactions on Speech and Audio Processing*, vol. 2, 291-298.

[11] Schuller, B.; Rigoll, G.; Lang, M., 2003. Hidden Markov Model-Based Speech Emotion Recognition, *Proc. ICASSP 2003*, IEEE, Hong Kong, China, vol. II, 1-4.

[12] Amir, N., Ron, S, 1998. Towards an automatic classification of emotions in speech. *Proc. 5th International Conference of Spoken Language Processing*, Sydney, Australia, 555–558.

[13] Witten, I. H.; Frank, E., 2000. *Data Mining, Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 133.