# EMOTIONS ARE A PERSONAL THING: TOWARDS SPEAKER-ADAPTIVE EMOTION RECOGNITION

*Maxim Sidorov, Stefan Ultes, and Alexander Schmitt*

Institute of Communications Engineering
Ulm University
Ulm, Germany

## ABSTRACT

In this paper, we present novel work on speech-based adaptive emotion recognition through addition of speaker-specific information. We propose a two-stage approach of first determining the speaker and then using this information during the emotion recognition process. The proposed technique has been evaluated using five emotional speech databases of different languages using both artificial neural network-based speaker identifier and the ground truth. The addition of speaker-specific information improves the emotion recognition accuracy by up to +10.2%. Moreover, emotion recognition performance scores for all applied databases are improved.

***Index Terms***— speech-based adaptive emotion recognition, speech-based speaker identification

## 1. INTRODUCTION

In human-human communication, people are usually quite capable of determining the emotions of the other person, while machines still do have a hard time recognizing people's emotions with the same accuracy. This information, however, is very beneficial. It may be applied in Interactive Voice Response (IVR) systems for adapting the course of the dialogue to the emotional state of the caller. Furthermore, it may also be used for analyzing human-human dialogues in call centers for identifying problematic calls. These calls may then be used as the basis for internal training of the agents.

State-of-the-art approaches for automatic emotion recognition regard the problem independently of the speaker. However, while the basic emotions are shared between all people and cultures [1], humans have a fine-tuned emotional model of people they know allowing for recognizing their emotions more accurate. Furthermore, speaker-specific models have shown to improve speech recognition as well (e.g., [2]). Hence, we present a novel approach on adding speaker-specific information to the emotion recognition process. Moreover, our particular interest lies on the question if adding speaker information may result in increased performance in general. Hence, the ground truth about the speaker is used for evaluating speaker-adaptive emotion recognition. In a second step, a real speaker identification system is applied and evaluated. In order to generate more general results, both approaches are applied to five different databases with different characteristics.

The rest of the paper is organized as follows: Significant related work is presented in Section 2. Section 3 describes the applied corpora and renders their differences. Our approach on speaker-specific emotion recognition is proposed in Section 4 having its results of numerical evaluations in Section 5. Conclusion and future work are described in Section 6.

## 2. SIGNIFICANT RELATED WORK

One of the pilot experiments which deals with speech-based emotion recognition has been presented by Kwon et al. [3]. The authors compared emotion recognition performance of various classifiers: support vector machine, linear discriminant analysis, quadratic discriminant analysis and hidden Markov model. For evaluation, the classifiers have been applied on the SUSAS [4] and the AIBO [5] databases of emotional speech. The authors achieved the highest value of accuracy by applying a Gaussian support vector machine (70.1% and 42.3% on the databases, correspondingly).

Vogt and André [6] improved the performance of emotion classification by automatic gender detection. The authors have used two different classifiers in order to classify male and female voices from the Berlin [7] and the SmartKom [8] corpus. They concluded that a combined gender and emotion recognition system improved the recognition rate of a gender-independent emotion recognition system by 2-4% relatively applying the Naive Bayes classifier.

Another approach for improving emotion recognition has been proposed by Polzehl et al. [9] by adding linguistic information, e.g., Bag-of-Words or Self-Referential Information. Evaluation with three different databases showed that fusion at the decision level adding confidence scores slightly improves the overall scores. However, evaluating acoustic and linguistic models on separate levels showed the dominance of acoustic models.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

**Table 1**. Databases description

| Database | Language | Full length (min.) | Num. of em. (sp.) | File level duration | | Emotion level duration | | Notes |
|---|---|---|---|---|---|---|---|---|
| | | | | Mean(sec.) | Std. (sec.) | Mean (sec.) | Std. (sec.) | |
| Berlin | German | 24.7 | 7 (10) | 2.7 | 1.02 | 212.4 | 64.8 | Acted, single utterances |
| Let's Go | English | 118.2 | 5 (291) | 1.6 | 1.4 | 1419.5 | 2124.6 | Non-acted, human-machine |
| SAVEE | English | 30.7 | 7 (4) | 3.8 | 1.07 | 263.2 | 76.3 | Acted, single utterances |
| UUDB | Japanese | 113.4 | 4 (14) | 1.4 | 1.7 | 1702.3 | 3219.7 | Non-acted, human-human |
| VAM | German | 47.8 | 4 (47) | 3.02 | 2.1 | 717.1 | 726.3 | Non-acted, human-human |

## 3. CORPORA

For the study, a number of speech databases has been applied for speaker-adaptive emotion recognition. In this Section, a brief description of each corpus is provided. Furthermore, their main differences are outlined including database language, acted vs. non-acted speech, and number of emotions.

**Berlin** The Berlin emotional database [7] was recorded at the Technical University of Berlin and consists of labeled emotional German utterances which were spoken by 10 actors (5 female). Each utterance has one of the following emotional labels: neutral, anger, fear, joy, sadness, boredom, and disgust.

**Let's Go** The Let's Go emotion database [10] comprises non-acted American English utterances extracted from an automated bus information system of the Carnegie Mellon University in Pittsburgh, USA. The utterances are requests to the Interactive Voice Response system spoken by real users with real concerns. Each utterance is annotated with one of the following emotional labels: angry, slightly angry, very angry, neutral, friendly, and non-speech (critical noisy recordings or just silence).

**SAVEE** Haq and Jackson [11] recorded the SAVEE (Surrey Audio-Visual Expressed Emotion) corpus for research on audio-visual emotion classification from four native English male speakers. The emotional label for each acted utterance is one out of the standard set of emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral).

**UUDB** The UUDB (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database [12] consists of spontaneous Japanese human-human speech. Task-oriented dialogue produced by seven pairs of speakers (12 female) resulted in 4,737 utterances in total. Emotional labels for each utterance were created by three annotators on a five-dimensional emotional basis (interest, credibility, dominance, arousal, and pleasantness). For this work, only pleasantness (or evaluation) and the arousal axis are used. The corresponding quadrant (counterclockwise, starting in positive quadrant, assuming arousal as abscissa) are then assigned to emotional labels: happy-exciting, angry-anxious, sad-bored and relaxed-serene [13].

**VAM** Based on the popular German TV talk-show "Vera am Mittag" (Vera in the afternoon), the VAM-Audio database [14] has been created at Karlsruhe Institute of Technology. The emotional labels of the first part of the corpus (speakers 1–19) were given by 17 human evaluators and the rest of the utterances (speakers 20–47) were labeled by six annotators, both on a three-dimensional emotional basis (valence, activation, and dominance). The emotional labeling was performed in a similar way to the UUDB corpora, using valence (or evaluation) and arousal axis.

While the Berlin and SAVEE corpora consist of acted emotions, the other three databases comprise real emotions. Furthermore, for both, English and German, acted and non-acted emotions have been considered, while only non-acted emotions were available for Japanese. A statistical description of the used corpora may be found in Table 1. Please also note that Let's Go, UUDB, and VAM are highly unbalanced (see Emotion level duration columns in Table 1).

Emotions itself and their evaluations have subjective nature. That is why it is important to have at least several evaluators of emotional labels.

## 4. STATISTICAL APPROACH

Incorporating speaker-specific information into the emotion recognition process may be done in many ways. A very straight forward way is to add this information to the set of features (System A). Another way is to create speaker-dependent models: While, for conventional emotion recognition, one statistical model is created independently of the speaker, one may create a separate emotion model for each speaker (System B). Both approaches result in a two-stage recognition procedure (see Figure 1 and Figure 2): First, the speaker is identified and then this information is included into feature set directly (System A), or the corresponding emotion model is used for estimating the emotions (System B). Both emotion recognition speaker identification (ER-SI) hybrid systems have been investigated and evaluated in this study.

The choice of the appropriate speech signal features for both problems is still an open question. As the focus of

**Fig. 1**. Hybrid Emotion Recognition System A: Addition of Speaker information to the feature set.

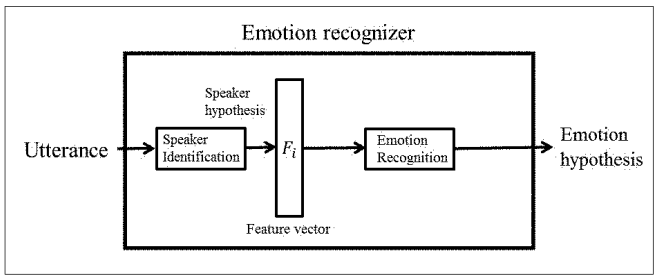
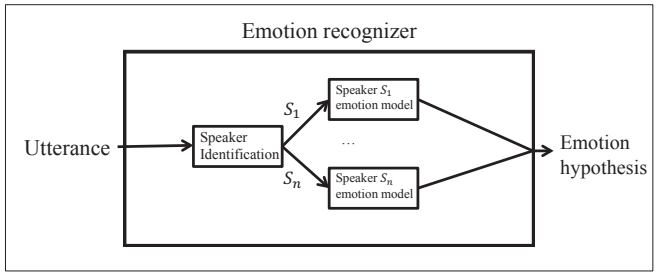



**Fig. 2**. Hybrid Emotion Recognition System B: A separate model for each speaker.




this study lies on improving emotion recognition by adding speaker dependency, no feature set optimization has been applied and the most popular features have been chosen (cf. [15]). Hence, the features vector includes average values of the following speech signal features: power, mean, root mean square, jitter, shimmer, 12 MFCCs, and five formants. Mean, minimum, maximum, range, and deviation of the following features have also been used: pitch, intensity and harmonicity. This results in a 37-dimensional feature vector for one speech signal file. The Praat system [16] has been used in order to extract speech signal features from wave files.

Each algorithm has been applied in a static mode, i.e., each speech signal was parameterized by one single 37-dimensional feature vector consisting of corresponding average values. As this study concentrates on the theoretical improvement of emotion recognition using speaker-specific information, usage of other speech signal features or modelling algorithms may improve the recognition performance.

## 5. EVALUATION AND RESULTS

To investigate the theoretical improvement of using speaker-specific information for ER, the true information about the speaker has been used. Then, in order to provide pilot experiments, a real SI component has been applied. For both tasks (ER and SI), a multi-layer perceptron, which is a baseline type of artificial neural networks, has been chosen as a modelling algorithm for both approaches.

As a baseline, an emotion recognition process without speaker-specific information has been conducted. The training set was used to create and train an artificial neural network (ANN) based emotion model. The test set was used to evaluate the model. Hence, one single neural network has been created addressing the emotions of every speaker in the database.

In the first experiment, the focus was on investigating the theoretical improvement, which may be achieved using speaker-based adaptivity. For this, known speaker information (true labels) was used for both approaches. In System A, the speaker information was simply added to the feature vector. Hence, all utterances with the corresponding speaker information were used to create and evaluate an ANN-based emotion model. For the System B, individual emotion models were built for each speaker. During the training phase, for each speaker, all speaker utterances were used for creating the emotion models. During testing, all speaker utterances were evaluated with the corresponding emotion model.

Additionally, a second experiment was conducted including a real speaker identification module instead of using known speaker information. First, an ANN-based speaker identifier was created during training phase. Furthermore, for System A, the known speaker information was included into the feature vector for the training of the emotion classifier. The testing phase starts with the speaker identification procedure. Then, the speaker hypothesis was included into the feature set which was in turn fed into the emotion recognizer. For System B, an ANN-based emotion recognizer was created for each speaker separately. For testing, the speaker hypothesis of the speaker recognition is used to select the emotion model which corresponds to the recognized speaker to create an emotion hypothesis. In contrast to the first experiment, these experiments are not free of speaker identification errors.

In order to generate more statistical significant results, the complete classification process was run 25 times for each database and experiment. For each run, the databases were randomly divided into training and test sets (70–30% correspondingly). While each database was stratified into training and testing sets by the *emotion* class, the Let's Go database was stratified into subsets by the *speaker* class, due to highly unbalanced distribution of the speaker class. For all experiments, z-score normalization has been applied for all features. The final results are shown in Table 2 for System A, where speaker-specific information is included into the feature set, and in Table 3, where separate emotion models are created for each speaker. The results are calculated taking the mean and standard deviation of all runs. The first column in Table 2 (Without SI) corresponds to ANN-based emotion recognition accuracy, which was achieved without speaker-specific infor-

**Table 2**. Evaluation Result of System A in percent: Accuracy of baseline (Without SI), Experiment 1 (True SI) and Experiment 2 (ANN SI, having SI accuracy in parentheses.)

| Database | Without SI | True SI | ANN SI |
|---|---|---|---|
| SAVEE | 59.31/3.27 | 63.78/3.02 | 63.78/2.94 (99.19/0.67) |
| Berlin | 73.76/2.54 | 77.74/3.18 | 74.61/3.60 (74.61/3.60) |
| VAM | 66.11/2.63 | 70.62/2.54 | 68.30/2.97 (67.44/2.59) |
| UUDB | 89.99/0.57 | 90.42/0.60 | 90.10/0.61 (73.47/1.30) |
| Let's Go | 76.99/1.09 | 78.53/1.38 | 78.22/1.36 (44.27/1.14) |

**Table 3**. Evaluation Result of System B in percent: Experiment 1 (True SI) and Experiment 2 (ANN SI, having SI accuracy in parentheses.)

| Database | True SI | ANN SI |
|---|---|---|
| SAVEE | 65.36/3.05 | 65.36/3.04 |
| Berlin | 74.01/2.66 | 68.84/3.29 |
| VAM | 68.72/2.59 | 65.35/2.86 |
| UUDB | 89.91/0.60 | 89.46/0.67 |
| Let's Go | 80.84/1.01 | 75.62/0.93 |

mation (baseline). In the second column (True SI), the accuracy of the emotion recognition system using known speaker information. The next column (ANN SI) contains the emotion recognition accuracy which used an ANN-based speaker identification module. Values within the parentheses depict the performance of the speaker identification module.

As a result, we can conclude that addition of speaker-specific information in emotion recognition procedure significantly improves the recognition performance. For all corpora, recognition accuracy has been improved by adding speaker information to the feature vector (see Table 4). This improvement is even significant for almost all databases.

## 6. CONCLUSION AND FUTURE WORK

It is evident that already a very simple method as extending the feature vector with additional speaker-specific information could improve the ER accuracy for all databases (even using a real SI module). This improvement is significant when using true SI information for most of the used corpora (see Table 1). These results are very encouraging leading to further more sophisticated approaches on speaker-dependent emotion recognition, e.g., applying speaker adaptation methods known from speaker-independent speech recognition.

However, for building accurate individual emotion models, balanced databases are required. In order to build emotion model for each speaker, a high number of emotional samples are needed for each speaker. Hence, for some of the corpora (VAM and UUDB), the addition of speaker information (as a building of separate emotion models) could not improve recognition accuracy (see emotion level duration columns for corresponding corpora in Table 1). The high standard deviation values show that, for some of the speakers, only a few

**Table 4**. Improvement in ER performance using the true speaker information (True SI) and using SI information of an ANN (ANN SI). Significant differences are marked with ** ($\alpha < .0001$) and * ($\alpha < .01$) using the T-test.

| Database | SI in FS (Sys. A) | | Separate Models (Sys. B) | |
|---|---|---|---|---|
| | True SI | ANN SI | True SI | ANN SI |
| SAVEE | +7.54% ** | +7.54% ** | +10.20% ** | +10.20% ** |
| Berlin | +5.40% ** | +1.15% | +0.34% | -6.67% ** |
| VAM | +6.82% ** | +3.31% * | +3.95% * | -1.15% |
| UUDB | +0.48% | +0.12% | -0.09% | -0.59% * |
| Let's Go | +2.00% ** | +1.60% * | +5.00% ** | -1.78% ** |

emotional samples are presented in the database. However, this is not enough data for building emotion models.

Moreover, such kind of problem decomposition favor accumulation of errors as can easily be seen by the complete probability formula

$$P(em) = P(sp) \cdot P(em|sp) + P(\overline{sp}) \cdot P(em|\overline{sp}) \ , \quad (1)$$

where $\mathtt{P}(em)$ and $\mathtt{P}(sp)$ denote the probabilities of correct emotion classification and speaker identification correspondingly and $\mathtt{P}(\overline{sp})$ is the probability of wrong speaker identification. $\mathtt{P}(em|sp)$ is a conditional probability of correct emotion recognition given the correct speaker label and $\mathtt{P}(em|\overline{sp})$ is the probability of correct emotion recognition given the false speaker label. In other words, estimating the emotion correctly may also happen when estimating the speaker wrongly but still estimating the emotion correctly. This corresponds to the similarity of emotions between speakers.

While an ANN already provides reasonable results for speaker identification, we are still examining its general appropriateness. The usage of other—possibly more accurate—identifiers may improve the performance of this hybrid system. Furthermore, dialogues may not only consist of speech, but also of a visual representation. Hence, an analysis of captured speaker images or even video recordings may also improve SI and ER performance. In addition, by applying feature selection techniques (principal component analysis, genetic algorithm-based feature selection, etc.) for this problem, further performance improvement could be achieved.

## 7. ACKNOWLEDGMENTS

This work was supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" which is funded by the German Research Foundation (DFG).

## 8. REFERENCES

[1] Klaus Scherer, "Emotion," in *Sozialpsychologie*, pp. 165–213. Springer, 2002.

[2] Christopher J Leggetter and PC Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[3] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee, "Emotion recognition by speech signals.," in *INTERSPEECH*, 2003.

[4] John HL Hansen, Sahar E Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom, "Getting started with susas: a speech under simulated and actual stress database.," in *EUROSPEECH*, 1997, vol. 97, pp. 1743–46.

[5] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong, ""' you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus.," in *LREC*, 2004.

[6] Thurid Vogt and Elisabeth André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. Citeseer, 2006.

[7] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of german emotional speech.," in *Interspeech*, 2005, pp. 1517–1520.

[8] Silke Steininger, Florian Schiel, Olga Dioubina, and S Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*, 2002.

[9] Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Communication*, vol. Special Issue: Sensing Emotion and Affect - Facing Realism in Speech Processing, 2011.

[10] Alexander Schmitt, Stefan Ultes, and Wolfgang Minker, "A parameterized and annotated corpus of the cmu let's go bus information system," in *International Conference on Language Resources and Evaluation (LREC)*, 2012.

[11] S. Haq and P.J.B. Jackson, *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pp. 398–423, IGI Global, Hershey PA, Aug. 2010.

[12] Hiroki Mori, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.

[13] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, and Andreas Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 552–557.

[14] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 865–868.

[15] Alexander Schmitt, Tobias Heinroth, and Jackson Liscombe, "On nomatchs, noinputs and bargeins: Do non-acoustic features support anger detection?," in *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, SigDial Conference 2009*, London (UK), Sept. 2009, Association for Computational Linguistics.

[16] Paul Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.