Comparison of Gender- and Speaker-adaptive Emotion Recognition

Maxim Sidorov, Stefan Ultes, and Alexander Schmitt
Institute for Communications Engineering, University of Ulm, Germany
{maxim.sidorov, stefan.ultes, alexander.schmitt}@uni-ulm.de

Abstract

Deriving the emotion of a human speaker is a hard task, especially if only the audio stream is taken into account. While state-of-the-art approaches already provide good results, adaptive methods have been proposed in order to further improve the recognition accuracy. A recent approach is to add characteristics of the speaker, e.g., the gender of the speaker. In this contribution, we argue that adding information unique for each speaker, i.e., by using speaker identification techniques, improves emotion recognition simply by adding this additional information to the feature vector of the statistical classification algorithm. Moreover, we compare this approach to emotion recognition adding only the speaker gender being a non-unique speaker attribute. We justify this by performing adaptive emotion recognition using both gender and speaker information on four different corpora of different languages containing acted and non-acted speech. The final results show that adding speaker information significantly outperforms both adding gender information and solely using a generic speaker-independent approach.

Keywords: adaptive emotion recognition, speaker identification, gender recognition

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 monitoring and analyzing human-human calls 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 (Scherer, 2002), humans have a fine-tuned emotional model of people they know allowing for recognizing their emotions more accurately. Furthermore, speaker-specific models have shown to improve speech recognition as well (e.g., (Leggetter and Woodland, 1995)). Hence, we presented work on speaker-adaptive emotion recognition (Sidorov et al., 2014) proving the general benefit of this approach. However, recent work on combined gender and emotion recognition by Vogt and André (Vogt and André, 2006) has shown improved recognition accuracy without having a speaker-specific model. Thus, we present investigate the approach on adding speaker-specific information to the emotion recognition process and compare it with a similar approach only regarding the gender of the speaker. Moreover, our particular interested lies on the question if adding speaker information may result in increased performance compared to solely adding gender information. The ground truth about the speaker (or gender) is used for modeling adaptive emotion recognition. In a second step, a real speaker (gender) identification system is applied and evaluated. In order to generate more general results, all approaches are applied to four different databases with different characteristics. Finally, all approaches are investigated using different feature sets.

The rest of the paper is organized as follows: Significant related work on emotion recognition including the gender-adaptive approach by Vogt and André is presented in the 2. Section. The 3. Section presents the applied corpora and renders their characteristics along with statistical details. Our approach on speaker-specific emotion recognition as well as gender-specific emotion recogniton is proposed in the 4. Section having its results of numerical evaluations using the presented corpora in the 5. Section. Conclusion and future work are described in the 6. Section.

2. Significant Related Work

One of the pilot experiments which deals with speech based emotion recognition has been presented by Kwon et al. (2003). 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 (Hansen et al., 1997) and the AIBO (Batliner et al., 2004) 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 (2006) 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 (Burkhardt et al., 2005) and the SmartKom (Steininger et al., 2002) corpora. They concluded that the combined gender and emotion recognition system improved the recognition rate of a gender-independent emotion recognition system by 2–4% relatively by applying the Naive Bayes classifier for building the emotion models.

Another approach for improving emotion recognition has been proposed by Polzehl et al. (2011) 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.

Table 1: Databases description

Database	Language	Full length (min.)	Number of emotions	File level duration		Emotion level duration		Notes
				Mean(sec.)	Std. (sec.)	Mean (sec.)	Std. (sec.)	
Berlin	German	24.7	7	2.7	1.02	212.4	64.8	Acted, single utterances
Let's Go	English	118.2	5	1.6	1.4	1419.5	2124.6	Non-acted, human-machine
UUDB	Japanese	113.4	4	1.4	1.7	1702.3	3219.7	Non-acted, human-human
VAM	German	47.8	4	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 and gender-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 (Burkhardt et al., 2005) 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 (Schmitt et al., 2012) 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).

UUDB The UUDB (The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies) database (Mori et al., 2011) 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 (Schuller et al., 2009b).

VAM Based on the popular German TV talk-show "Vera am Mittag" (Vera in the afternoon), the VAM-Audio database (Grimm et al., 2008) 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 axes.

While the Berlin corpus consists of acted emotions, the other three databases comprise real emotions. Furthermore, for German, acted and non-acted emotions have been considered, while only non-acted emotions were available for Japanese and English. 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).

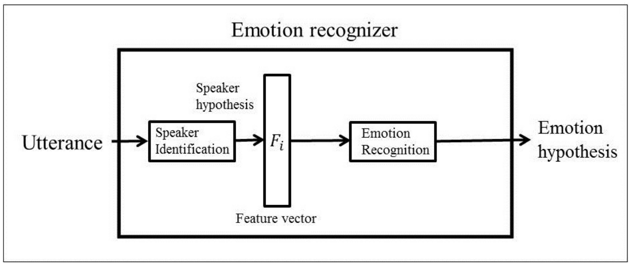
4. Statistical Approach

Incorporating information about the speaker into the emotion recognition process may be done in many ways. A straightforward way, which will be investigated in this contribution, is to add this information to the set of features used for creating the emotion recognition model. This results in a two-stage recognition approach (see Figure 1 for an example using speaker-identification): In the first stage, the speaker information is identified. This step is independent of the actual utilization of this information for emotion recognition. After adding this information to the feature set (or feature vector), the emotion may be recognized using a speaker-specific emotion recognition model directly. The described emotion recognition speaker identification (ER-SI) hybrid systems has been investigated in this study and compared to a emotion recognition gender identification (ER-GI) hybrid system. For the latter, the speaker identification module in Figure 1 is simply replaced by a gender recognition module.

The choice of the appropriate speech signal features is still an open question. As the focus of this study lies on comparing emotion recognition with added speaker dependency with gender-dependent emotion recognition, no feature selection has been applied. Two feature sets have been investigated for both approaches: the first set consists of the most popular features for emotion recognition (cf. (Schmitt et al., 2009)): 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 second feature set is equal to the feature set of the Interspeech Emotion Challenge of 2009 (Schuller et al., 2009a) consisting of a total of 384 features also including a part of the first feature set. Additionally, so called functionals have been added. We investigate the performances of the proposed approaches with two different features sets, as the effect of adding one feature to a small feature set might be higher compared to adding one feature to a larger feature sets.

Figure 1: Hybrid Emotion Recognition System: Addition of Speaker information to the feature set.



The speech signal features of the first set have been extracted from wave files using the Praat system (Boersma, 2002). For the second feature set, the OpenSMILE framework (Eyben et al., 2010) has been used for feature extraction.

For both gender recognition and speaker identification, the models have been created independently of the emotion recognition module. All models have been created in a static mode. That means that one feature vector consisting of all corresponding average values of 37 features (or 384 features, respectively) has been derived from each speech signal, i.e., wave audio file, and then used for training of the models. For training the speaker- and gender-dependent emotion recognition modules, of course, the number of feature used was 38 features (or 385 features, respectively).

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 emotion recognition (ER), the true information about the speaker and their gender has been used. Then, in order to provide pilot experiments, a real speaker identification (SI) component has been applied as well as a real gender identification (GI) component. For both tasks (ER and SI/GI), a multi-layer perceptron, which is a baseline type of artificial neural networks, has been chosen as a modelling algorithm.

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

In the first experiment (E1), the focus was on investigating the theoretical improvement of adaptive emotion recognition and thus comparing the performances of adding speaker information compared to solely adding gender information to the emotion recognition module. The theoretical improvement is investigated using the true labels for training and evaluating the emotion recognizer. Thus, E1 is conducted without actually identifying the speaker or recognizing their gender, respectively as the ground truth was used instead. As described in the 4. Section, the speaker or gender information has simply been added to the feature vector of the emotion recognition module. For both, all utterances along with the corresponding speaker or gender information have been used to train and evaluate the ANN-based emotion model.

In the second experiment (E2), the speaker and the gender of the speaker were actually estimated using a preceding recognition module. This results in the described two-stage recognition process. Correspondingly, the training process is also divided into two phases. First, to create the speaker identification model and the gender recognition model, the corresponding ANN-based recognizers identifying the speaker or the gender out of the speech signal are built in the training phase. For training of the adaptive emotion classifier, again, the ground truths about the speaker and the gender are used and added to the feature vector. For speaker-adaptive emotion recognition, 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. Thus, in contrast to E1, E2 is not free of speaker identification errors. For testing gender-adaptive emotion recognition, the general procedure was the same having a gender recognition module instead of speaker identification.

In order to generate statistical significant results, each complete classification process was run 25 times for each database and experiment for both gender- and speaker-adaptive emotion recognition. For each run, the databases were randomly divided into training and testing sets (70–30% correspondingly).

The final results for speaker-adaptive emotion recognition for both feature sets are shown in Table 2. These results are calculated taking the average of all runs. For each feature set, the first columns correspond to ANN-based emotion recognition accuracy which was achieved without speaker specific information (baseline). In the respective second column, the accuracy of the emotion recognition system using known speaker information is shown. The next column 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 is clearly visible, adding speaker information improves the accuracy for both feature sets for all databases.

The results of gender-adaptive emotion recognition are depicted in Table 3. Again, the results are separated according to the feature set. As can be seen, gender-adaptive emotion recognition also increases the recognition performance compared to the baseline of estimating emotions without additional information for both feature sets.

For comparing speaker- and gender-adaptive emotion recognition, Figure 2 depicting results of E1 shows clearly that speaker-specific emotion recognition outperforms all other approaches for all databases. This also applies when using actual estimated information which is shown in Figure 3 for E2: speaker-adaptive emotion recognition outperforms all other approaches.

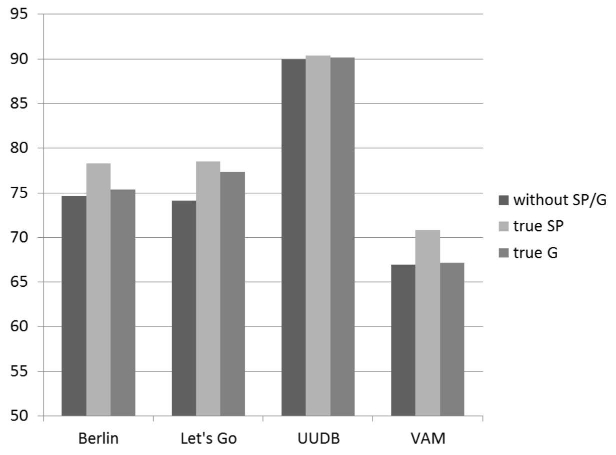
Table 2: Evaluation result (mean / standard deviation) of speaker-adaptive emotion recognition (37- and 384- dimensional feature vectors) in percent: Accuracy of baseline (Without SP), Experiment 1 (True SP) and Experiment 2 (ANN SP, having SP accuracy in parentheses.)

Database	Without SP	True SP	ANN SP	Without SP	True SP	ANN SP
	37-dimensional feature vector			384-dimensional feature vector		
Berlin	74.63/3.78	78.29/3.28	75.98/3.36 (74.93/2.53)	80.99/2.53	83.98/2.50	83.08/2.52 (88.65/2.13)
Let's Go	74.11/1.33	78.53/1.38	78.22/1.36 (44.27/1.14)	78.62/1.03	79.01/3.15	78.30/4.76 (57.55/1.27)
UUDB	89.91/0.71	90.37/0.84	89.78/0.93 (72.63/1.15)	89.45/0.30	89.26/2.50	89.16/2.52 (73.63/4.19)
VAM	66.97/2.51	70.85/2.03	68.46/2.06 (68.07/2.45)	68.99/1.95	72.08/1.71	70.17/1.67 (77.97/1.86)

Table 3: Evaluation result (mean / standard deviation) of gender-adaptive emotion recognition (37- and 384- dimensional feature vectors) in percent: Experiment 1 (True G) and Experiment 2 (ANN G, having G accuracy in parentheses.)

Database	True G	ANN G	True G	ANN G
	37-dimensional feature vector		384-dimensional feature vector	
Berlin	75.38/4.32	74.29/4.11 (95.75/1.08)	82.16/2.83	81.74/2.90 (97.42/1.09)
Let's Go	77.32/0.96	76.89/1.01 (85.67/0.95)	80.32/1.13	79.67/1.14 (86.42/1.89)
UUDB	90.13/0.65	90.05/0.62 (97.01/0.38)	88.80/3.06	88.78/3.09 (98.65/0.29)
VAM	67.20/2.62	67.18/2.59 (96.44/0.78)	69.37/1.38	69.17/1.30 (96.06/0.76)

Figure 2: Accuracy comparison: ER without additional information, ER with true SP-specific information and ER with true G-specific information. All differences are significant with $\alpha < 0.05$ when applying the t-test (Student, 1908) except for UUDB.

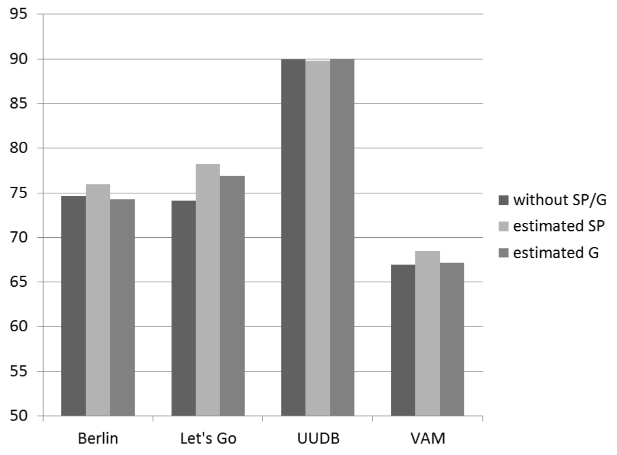


The results of speaker- and gender-adaptive emotion recognition as well as the baseline results have been tested for significance using the t-test (Student, 1908) for comparing the results of each of the 25 runs of the experiments. All differences are significant with at least $\alpha < 0.1$ except for UUDB.

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) even for a large feature set. It was figured out that speaker-adaptive ER outperforms state-of-the-art gender-adaptive ER for almost all corpora.

Figure 3: Accuracy comparison: ER without additional information, ER with estimated SP-specific information and ER with estimated G-specific information. All differences are significant with $\alpha < 0.1$ when applying the t-test (Student, 1908) except for UUDB.



This improvement is significant when using both true SP/G information and estimated SP/G 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 methods known from speaker-dependent speech recognition.

However, such a kind of a problem decomposition favors the accumulation of errors. Hence, there is still a gap between emotion recognition accuracy using the known speaker information and an actual SI module.

While an ANN already provides reasonable results for speaker identification, we still examine its general appropriateness. The usage of other—possibly more accurate—

identifiers may improve the performance of this hybrid system. Furthermore, dialogues do not only consist of speech, but also of a visual representation. Hence, an analysis of picture or even video recordings may also improve SI and ER performance.

7. References

Batliner, A., Hacker, C., Steidl, S., Nöth, E., D'Arcy, S., Russell, M. J., and Wong, M. (2004). " you stupid tin box"-children interacting with the aibo robot: A cross-linguistic emotional speech corpus. In *LREC*.

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Interspeech*, pages 1517–1520.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of ACM Multimedia (MM)*, pages 1459–1462. ACM, October.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The vera am mittag german audio-visual emotional speech database. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 865–868. IEEE.

Hansen, J. H., Bou-Ghazale, S. E., Sarikaya, R., and Pellom, B. (1997). Getting started with susas: a speech under simulated and actual stress database. In *EUROSPEECH*, volume 97, pages 1743–46.

Kwon, O.-W., Chan, K., Hao, J., and Lee, T.-W. (2003). Emotion recognition by speech signals. In *INTERSPEECH*.

Leggetter, C. J. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50.

Polzehl, T., Schmitt, A., Metze, F., and Wagner, M. (2011). Anger recognition in speech using acoustic and linguistic cues. *Speech Communication*, Special Issue: Sensing Emotion and Affect - Facing Realism in Speech Processing.

Scherer, K. (2002). Emotion. In *Sozialpsychologie*, pages 165–213. Springer.

Schmitt, A., Heinroth, T., and Liscombe, J. (2009). 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), September. Association for Computational Linguistics.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*.

Schuller, B., Steidl, S., and Batliner, A. (2009a). The interspeech 2009 emotion challenge. In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, September.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009b). Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*, pages 552–557. IEEE.

Sidorov, M., Ultes, S., and Schmitt, A. (2014). Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May.

Steininger, S., Schiel, F., Dioubina, O., and Raubold, S. (2002). Development of user-state conventions for the multimodal corpus in smartkom. In *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*.

Student. (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*. Citeseer.