Improving Automatic Emotion Recognition from Speech via Gender Differentiation

Thurid Vogt[*][†], Elisabeth André[*]

[*]Multimedia Concepts and Applications Group
Augsburg University, Germany
{andre,vogt}@informatik.uni-augsburg.de
[†] Applied Computer Science Group
Bielefeld University, Germany

Abstract

Feature extraction is still a disputed issue for the recognition of emotions from speech. Differences in features for male and female speakers are a well-known problem and it is established that gender-dependent emotion recognizers perform better than gender-independent ones. We propose a way to improve the discriminative quality of gender-dependent features: The emotion recognition system is preceded by an automatic gender detection that decides upon which of two gender-dependent emotion classifiers is used to classify an utterance. This framework was tested on two different databases, one with emotional speech produced by actors and one with spontaneous emotional speech from a Wizard-of-Oz setting. Gender detection achieved an accuracy of about 90 % and the combined gender and emotion recognition system improved the overall recognition rate of a gender-independent emotion recognition system by 2–4 %.

1. Introduction

The automatic recognition of emotions has recently received much attention for building more intuitive human-computer interfaces. Speech usually comes to mind first when thinking about possible sources to exploit for emotion recognition. It provides two types of information that are relevant for emotions: its acoustic properties and its linguistic content. Our focus here lies on emotion recognition from acoustic features.

So far, a lot of work has been done on finding a set of acoustic features describing optimally the properties of the speech signal that are relevant for emotions. Because, currently, there exists no such set, most researchers compute a multitude of features trying to be as exhaustive as possible, and from this set of many correlated features the best are chosen by an automatic selection algorithm. This procedure has proved to be successful in our earlier work (Vogt and André, 2005) and has also been applied e. g. by (Oudeyer, 2003), (Batliner et al., 2003a) and (Küstner et al., 2004).

The most commonly used features in the literature are related to pitch, energy and speaking rate, as well as spectral features such as mel-frequency cepstral coefficients (MFCCs). Especially pitch is gender-dependent, with ca. 160 Hz forming an upper resp. lower bound for male resp. female pitch (this holds for pitch of speakers in a neutral emotional state). This threshold is used in (Abdulla and Kasabov, 2001) for separating genders to improve automatic speech recognition. But also most of the other features are gender-dependent to varying degrees. If the recognition system ignores this issue a misclassification of utterances might be the consequence. For instance highly aroused men utterances might be confused with neutral women utterances. This is a great disadvantage for an automatic system, particulary compared to human listeners who would easily distinguish these two cases, even if they cannot see their conversational partner.

One possibility to deal with this phenomenon is to normalize gender-dependent features, e. g. by relating the mean pitch of a speech segment to the minimum and maximum pitch values in that segment. Still, gender-specific emotion recognizers perform better than those with both genders mixed (Lee and Narayanan, 2005). This is demonstrated by (Ververidis and Kotropoulos, 2004) who show that the combined performance of a male and a female emotion recognizer is better than that of a gender-independent recognizer. However, one must know the gender of a speaker to separate classifications. (Ververidis and Kotropoulos, 2004) assumed gender information to be a priori given, but this is only the case in offline or purely academic systems. This is where we apply our scheme: we precede the emotion recognition with an automatic gender detection module that decides whether a male or a female emotion recognizer is given the respective utterance. Since automatic gender detection cannot be 100 % correct, this method could have a negative effect on the overall classification accuracy. On the other hand, combining gender and emotion recognition could also lead to an even higher improvement compared to a gender-independent emotion recognition system than an emotion recognition system based on correct gender information. We achieve classification accuracies that at least exceed gender-independent systems.

We will describe our framework in detail in Section 2. Section 3 presents the databases on which tests were performed. Section 4 and 5 discuss our results in terms of relevant features and classification results.

2. Combined gender and emotion recognition

Our basic emotion recognition system, which is in detail presented in (Vogt and André, 2005), works as follows: From the series of energy, MFCCs, the center of gravity of the spectrum, duration and pause values of a time unit of emotional analysis, further series such as the se-

This work was partially funded by a grant from the DFG in the graduate program 256 and the European Network of Excellence Humaine.

ries of maxima or minima, distances, differences or slope between adjacent extrame are derived. For every series, mean, minimum, maximum, range, variance, median, first quartile, third quartile and interquartile range are computed. From the resulting 1289 features, the most relevant ones for the given task are chosen. The optimized feature subsets are chosen by a best-first search through the feature space according to the classification performance achieved by a Naïve Bayes classifier, which is also used for the final classification of utterances. We decided on the Naïve Bayes classifier because it is fast to train which is advantageous for our feature selection. Though other classifiers such as support vector machines generally perform better, Naïve Bayes is still satisfactorily and we are interested here in the relative improvements different feature constellations can lead to and not in the absolute classification results. Feature selection as well as classification are carried out with the data mining software Weka (Witten and Frank, 2000).

We observed that the features we used for emotion recognition were also relevant for gender detection. It is often suggested to do gender detection based on mean pitch, but, as described above, this can be misleading in very emotional speech. Therefore, we argue for a greater number of features which actually leads to better recognition results, as we will show in Section 5. Generally, automatic gender detection achieves a high accuracy with little effort, so it is straight forward to integrate gender detection and emotion recognition into a two-stage recognizer, which first predicts the gender of the speaker and then, depending on the outcome, uses a gender-specific emotion recognizer (see figure 1).

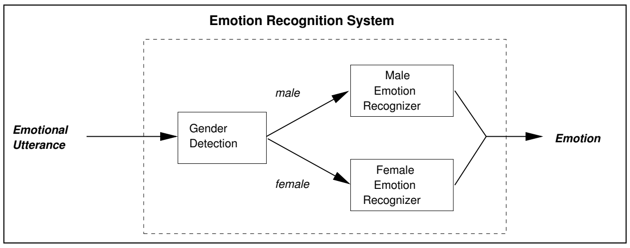

Figure 1: Combining a gender detection system and two gender-specific emotion classifiers into a single emotion recognition system.

All these recognizers are trained in the same manner as the basic emotion recognition system, only the input data or class definition (emotions vs. gender) changes. For the training of the gender-specific emotion systems, only those utterances of the training set that were classified to the respective gender by the gender detection system were used.

In the following, the combined gender and emotion detection system will be compared to an emotion recognition system without gender information and to one with information about the correct gender.

3. Databases

The framework was tested on 2 different databases, the Berlin database of emotional speech (Burkhardt et al., 2005) and the SmartKom mobile database (Steininger et al., 2002). Both databases are in German. The Berlin database contains speech with acted emotions while the SmartKom database was recorded in a Wizard-of-Oz setting and thus contains more natural emotions. The data was evaluated differently than in our work published in (Vogt and André, 2005) in having predefined test and training sets, because this time, speakers should be kept separate and not mixed, so that the test was completely speaker-independent.

3.1. Berlin database of emotional speech

The Berlin database of emotional speech was recorded at the Technical University, Berlin. It contains 493 utterances of 10 professional actors who spoke 10 sentences with emotionally neutral content in 7 different emotions. The emotions were anger, joy, sadness, fear, disgust, boredom and a neutral emotional state.

This database was evaluated in 5 folds with always 1 male and 1 female speaker in the test set and the remaining 4 male and 4 female speakers in the training set. Always whole utterances were analysed.

3.2. SmartKom mobile database

This database was recorded within the SmartKom project at the University of Munich, from persons interacting with a multi-modal dialogue system. We evaluated only those utterances that were recorded in the mobile setting with a head-set microphone. This subset was splitted into a training set with 56 speakers (24 male and 32 female) and a test set with 14 speakers (7 male and 7 female). The original annotation comprises 12 emotional states. Since 12 emotional classes in spontaneous speech are – for the current state-of-the-art in speech emotion recognition – a too complex task, we merged them into a 4-class problem applying a scheme suggested in (Batliner et al., 2003b): neutral and unidentifiable utterances to *neutral*, strong and weak joy and strong and weak surprise into *joy*, strong and weak pondering/reflecting and strong and weak helplessness into *helplessness* and strong and weak anger into *anger*.

The distribution of emotions is 81 % neutral, 4 % joy, 10.5 % helplessness and 4.5 % anger, which is obviously very unbalanced and makes a good classification especially challenging. The units of emotional analysis were speech segments with no silent parts of more than 0.2 seconds.

4. Relevant features

As described in Section 2, relevant features were selected by a best-first search through the feature space using the classification accuracy of a Naïve Bayes classifier on the test set as selection criterion. This always resulted in only a very small subset of the original 1289 features, but altogether, a wide range of features was selected in all tasks.

In the following, we discuss the selected features. However, we consider here only classes of features, because a more in depth examination of single features is not practical. Since the number of features is very large and many features are highly correlated, it is not reasonable to make an assertion about a particular feature.

4.1. Relevant features for gender detection

The features used for gender detection used in the two databases are listed in Table 1. Apparently, not only pitch-related features, but also MFCC and energy-related features are meaningful for gender detection. In combination with other features, energy and MFCCs even play a more important role than pitch. Pitch, however, *is* the most discriminating feature when used alone.

Features	Berlin	SmartKom
Pitch	1	2
Energy	2	3
MFCC	17	7
$\sum$	20	12

Table 1: Relevant features for gender detection

4.2. Differences in relevant features for male and female emotions

The features selected for the classification of gender-independent, male and female emotions in the Berlin database are displayed in Table 2.

Features	gender-independent	male emotions	female emotions
Pitch	1	2	3
Energy	2	1	2
MFCC	17	7	15
$\sum$	20	10	20

Table 2: Relevant features for gender-independent, male and female emotion classification on the Berlin database

It is notable that the energy features were all derived from the energy derivation series and that, compared to the features chosen for emotion recognition without gender information, few energy features are found. About half the features in both male and female emotions were first quartile or interquartile range of a feature series. The number of features was constrained by the number of instances in the least frequent class as represented in one of the five training sets and is therefore not meaningful.

The respective features for the SmartKom database are shown in Table 3. The number of features is striking as for male emotions, almost 3 times as many features were selected. The number of features for the gender-independent classification lies in between. Again, a lot of features were the first quartile of a feature series.

Features	gender-independent	male emotions	female emotions
Pitch	1	3	1
Energy	3	1	1
MFCC	6	13	4
$\sum$	10	17	6

Table 3: Relevant features for gender-independent, male and female emotion classification on the SmartKom database

Altogether, slightly more pitch-related features can be found in the gender-specific selected feature sets which confirms that pitch is more meaningful when gender effects are eliminated.

5. Classification results

5.1. Gender detection

Table 4 shows the results for the gender detection on the two databases, 1) using only mean pitch as suggested in (Lee and Narayanan, 2005) and 2) after selection of the most relevant features from the original emotion recognition feature set of 1289 features. The optimized feature sets are in both cases significantly better compared to mean-pitch-only separation. The difference is higher for the Berlin database which contains a much higher proportion of emotional speech. It is therefore critical to use mean pitch only for gender detection in this kind of speech, so we are confirmed in opting for more features than only mean pitch when processing gender detection from emotional speech, as in our system.

While emotion recognition is much harder for spontaneous than for acted speech, no extreme difference could be observed for gender detection. This is not surprising, as gender differences are much more obvious and do not depend on speech quality. That the gender detection is even slightly better for spontaneous speech is presumably due to the larger number of training instances in the SmartKom database compared to the Berlin database and to the less emotional speech in it.

	mean pitch	optimized feature set
Berlin	69.37 %	90.26 %
SmartKom	87.56 %	91.85 %

Table 4: Accuracy of gender detection from mean pitch only and from the optimized feature set on two databases

In the Berlin database, misclassified utterances were observed only for certain emotional categories: Wrongly classified female utterances were found in disgusted, bored, sad and neutral emotional states, while wrongly classified male utterances were found in joy, fear and anger. For the SmartKom database, there could not be made such an distinction for males and females.

5.2. Emotion recognition

Table 5 shows the recognition results for emotion recognition without gender information, with gender information and with automatically detected gender information. Overall recognition accuracy and averaged class recognition accuracy is given, since for the SmartKom database with the vast majority of data belonging to only one class (neutral), the overall recognition accuracy does not reveal the discrimination of emotions adequately (while this holds for the Berlin database with a very balanced distribution of emotions). Combining gender and emotion recognition leads to a relative improvement of 2 % (Berlin database) resp. 4 % (SmartKom database) of the classification rate

		Berlin database		SmartKom database	
		RR	CL	RR	CL
Without gender information		81.14 %	79.18 %	75.11 %	34.43 %
With correct gender information	female	84.62 %	83.35 %	78.99 %	37.38 %
	male	87.92 %	79.95 %	75.36 %	38.85 %
	combined	86.00 %	83.48 %	76.74 %	38.02 %
With recognized gender information	female	84.93 %	83.87 %	81.38 %	41.21 %
	male	80.09 %	72.28 %	75.84 %	41.58 %
	combined	82.76 %	79.79 %	78.22 %	41.09 %

Table 5: Overall recognition rate (RR) and class-wise recognition rate (CL) of emotion detection using a gender-independent emotion recognizer, a gender-dependent emotion recognizer, and a combined gender and emotion recognizer

compared to the gender-independent classification of emotions. The class-wise recognition rate on the SmartKom database is even improved by 16 %. Obviously, this is a significant improvement. The recognition rate on the SmartKom database is different to the rate in our previously published work. This is due to the different evaluation strategy which is completely gender-independent. For the SmartKom database, the new system even outperformed the system with correct gender information. We suppose that gender detection should be interpreted here as a detection of male or female sounding voices rather than of male or female persons leading to a better partition of the data.

These results disprove a possible objection against two consecutive classifiers, that they might perform worse than one classifier when the second classifier gets false input from the first classifier, and so, errors might sum up. The reasons for this are probably that the gender detection is very accurate, and that even if it fails, this is in cases which are “untypical” for the true gender and so can as well or even better be recognized by the emotion recognizer of the detected gender.

Though splitting the classification task into two is more time-consuming, time constraints can be neglected as classification can still be done in real-time when using a simple but fast algorithm like the Naïve Bayes classifier.

6. Conclusion

We presented a new framework to improve emotion recognition from speech by making use of automatic gender detection. Starting from the fact that gender-specific emotion recognizers work more accurately than gender-independent ones, we extended our basic emotion recognition system by a preceding stage of gender detection that determines which gender-specific emotion recognition system should be used. We showed that gender detection is more reliable when not only based on mean pitch, as usually proposed in the literature, but also on energy, spectral and MFCC features.

Tests were carried out on two different databases, one with acted and one with spontaneous emotions. We pointed out some differences in the relevant features for the classification of emotions in male and female voices. With two-stage emotion recognition we could achieve an improvement of the recognition accuracy, which was even higher for the database with spontaneous emotions. For this database, gender separation does still not solve the problem of finding good discriminative features for emotion recognition, but it leads to a considerable improvement.

7. References

W. H. Abdulla and N. K. Kasabov. 2001. Improving speech recognition performance through gender separation. In *Proceedings of ANNES*, pages 218–222.

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. 2003a. How to find trouble in communication. *Speech Communication*, 40:117–143.

A. Batliner, V. Zeißler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. 2003b. We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proceedings of EUROSPEECH*, pages 733–736, Geneva.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech. In *Proceedings Interspeech*, Lissabon, Portugal.

D. Küstner, R. Tato, T. Kemp, and B. Meffert. 2004. Towards real life applications in emotion recognition. In *Workshop on Affective Dialogue Systems*, pages 25–35, Kloster Irsee, Germany.

C. M. Lee and S. S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, March.

P.-Y. Oudeyer. 2003. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, 59(1–2):157–183.

S. Steininger, F. Schiel, O. Dioubina, and S. Raubold. 2002. Development of user-state conventions for the multimodal corpus in SmartKom. In *Proceedings Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas.

D. Ververidis and C. Kotropoulos. 2004. Automatic speech classification to five emotional states based on gender information. In *Proc. 12th European Signal Processing Conference*, pages pp. 341–344, Vienna, September.

T. Vogt and E. André. 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of ICME*, Amsterdam.

I. H. Witten and E. Frank. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco.