Speaker and Listener Variations in Emotion Assessment

Michael Grimm[1], Kristian Kroschel[1], and Shrikanth Narayanan[2]
[1] *Universität Karlsruhe (TH), Institut für Nachrichtentechnik, 76128 Karlsruhe, Germany*
[2] *University of Southern California, Speech Analysis and Interpretation Lab, Los Angeles CA 90089, USA*
Email: grimm@int.uni-karlsruhe.de *

Introduction

In this paper we discuss both the speaker dependent and the listener dependent aspects in the assessment of emotions in speech. These dependencies form a basis to improve current emotion recognition systems as they can be applied in man-machine interaction, for instance.
Emotion recognition in speech has gained much attention in recent years [1, 2, 3]. However, human evaluation of emotions has not been exploited to its full extent so far. Each person perceives emotions differently. Therefore in this paper we analyze evaluation variations and propose to use them to get better references for the emotion recognition systems. Also we investigate different speakers' emotion expression variabilities which could be used to build speaker models for model-based speaker-independent automatic emotion recognition.
For the analysis of variations in producing and perceiving emotions, we use a 3D emotion space concept, where emotions are described as a combination of three values along continuous axes of generic attributes ("primitives"). These primitives are *valence* (negative vs. positive), *activation* (calm vs. excited), and *dominance* (weak vs. strong) [4] with a range of values from -1 to +1 each. The assessment results were used to calculate the distributions (centroids, covariances) of the four emotion classes *angry, happy, neutral,* and *sad* in the emotion space spanned by the three primitives. The individual classes have shown to form little-overlaping subspaces within this 3D emotion space [6]. Using these emotion categories, emotion variations can be observed as modified positions of their distributions in the emotion space. In particular, we will focus on the centroids and covariances of the emotion clusters.

Data

For this study, we used the *EMA Corpus* containing 680 sentences of emotional speech in American English [7]. One professional and two non-professional native speakers (1f/1m) produced 10 (14) sentences in each of the emotions *angry, happy, neutral, sad* with 5 repetitions each. The sampling rate was 16 kHz, with 16 bit resolution.
The sentences were evaluated using a text-free method. The evaluators rated the values on a 5-point scale each, as described in [5]. In total, 18 evaluators assessed the database (14m/4f). The average standard deviation was 0.35 for all primitives. The average correlation between the evaluators was 0.63 for valence, 0.79 for activation, and 0.75 for dominance. 9.7% of the sentences were discarded due to too high standard deviation in the evaluation (> 0.5).

*This work was supported by a scholarship of the German Academic Exchange Service (DAAD).

	Centroids								
	Speaker 1			Speaker 2			Speaker 3		
	Val	Act	Dom	Val	Act	Dom	Val	Act	Dom
Angry	-0.49	0.51	0.65	-0.34	0.39	0.49	-0.23	0.47	0.47
Happy	0.58	0.44	0.27	0.23	0.13	0.11	0.12	-0.08	-0.01
Neutral	-0.18	-0.33	-0.08	-0.15	-0.26	-0.14	-0.16	-0.36	-0.20
Sad	-0.49	-0.70	-0.61	-0.31	-0.47	-0.42	-0.48	-0.53	-0.58
	Standard deviations								
	Speaker 1			Speaker 2			Speaker 3		
	Val	Act	Dom	Val	Act	Dom	Val	Act	Dom
Angry	0.23	0.27	0.21	0.13	0.12	0.11	0.14	0.15	0.11
Happy	0.24	0.20	0.14	0.14	0.16	0.10	0.13	0.10	0.07
Neutral	0.09	0.09	0.12	0.09	0.10	0.10	0.09	0.09	0.08
Sad	0.15	0.13	0.17	0.10	0.14	0.15	0.12	0.11	0.13

Table 1: Emotion classes in emotion space – centroids and standard deviations

Speaker Variability

For the analysis of the speakers' variability in expressing emotions we compared the average ratings of all human evaluators, for each of the acted emotion classes individually. The result can be read from Tab. 1. These results indicate that the class distributions vary from speaker to speaker. Not in a general sense, i.e. angry is negative for all for instance, but there are some variations. The most significant differences are found in the centroids of happy, where valence ranges from 0.12 for Speaker 3 to 0.58 for Speaker 1. Activation and dominance are both even smaller than 0 for Speaker 3 (though very little), whereas it is 0.44 (activation) and 0.27 (dominance) for Speaker 1. Similar observations can be stated for angry: Valence varies from -0.23 for Speaker 3 to -0.49 for Speaker 1, dominance from 0.47 for Speaker 3 to 0.65 for Speaker 1. This means, Speaker 3 shows his emotions in a more moderate way.
The standard deviations do not vary to that extent. In general, they are quite small with values being in the range of 0.1 to 0.2. The standard deviation for neutral (Speaker 3: 0.09) is smallest in each component.
Fig. 1 shows the 2σ-region of the class distributions. The clusters are marked with a letter for each emotion class (A,H,N,S) and a digit for each speaker (1,2,3).
To model the emotion expression behavior, we introduce two speaker specific parameters: the *emotion expression bias (EEB)* and the *emotion expression amplification (EEA)*. These parameters define the speaker's indi-

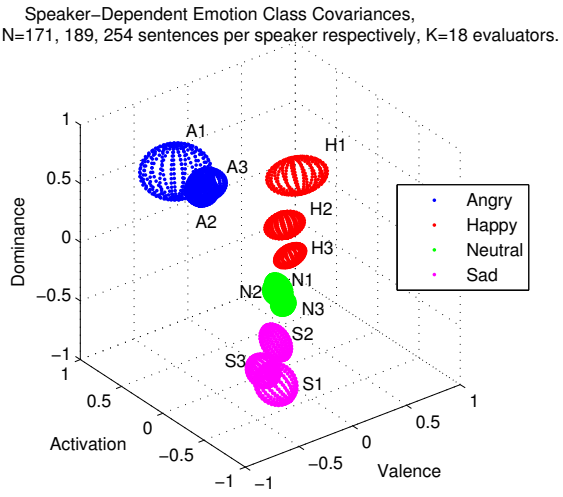

Figure 1: Covariance plot of the emotion classes *angry* (A), *happy* (H), *neutral* (N), and *sad* (S) in 3D emotion space of 3 speakers

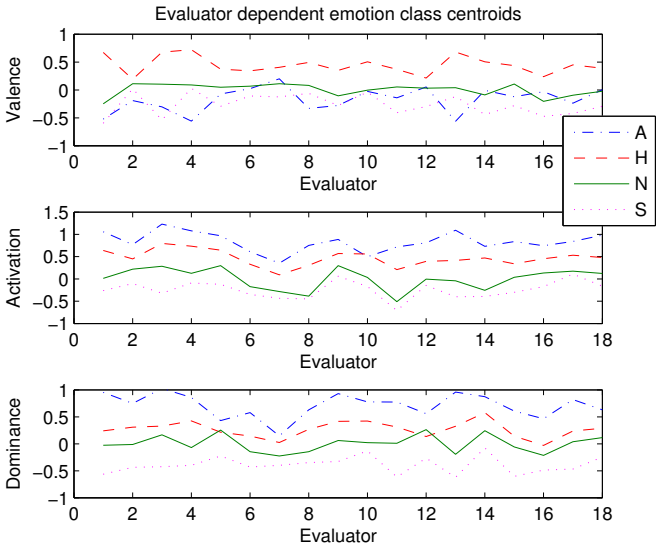

Figure 2: Comparison of evaluator influence on the centroids of normalized emotion classes

vidual deviation from the average emotions, calculated as a mean value of all speakers and all evaluators. The EEB is chosen as the speaker-dependent centroid of the neutral class (cf. Tab. 1). The EEA is chosen as the average quotient of the EEB-subtracted centroids and the initial ones, yielding 1.47, 0.80, and 0.73 for Speaker 1, 2, and 3, respectively.
Modeling the speaker's emotion expression behavior with these 2 parameters and a set of uniform emotions is both very efficient and exact. Reconstructing the initial emotion class centroids from the models yields a mean error of 0.06.

Listener Variability

To assess the variability of the listener side in emotion evaluation, we analyzed the centroids of the four emotion classes as a function of the evaluator. The speaker dependency was dismissed by using the EEB-subtracted and EEA-scaled evaluator-dependent centroids of these classes.
The result is shown in Fig. 2. Although there seem to be some common tendencies, an obvious parametrization is not feasible. However, it can be observed that for most evaluators, happy and angry show opposite extremes on valence, and they go in parallel on activation and dominance. Neutral and sad go in parallel for all emotion primitives. Since these tendencies are emotion-dependent, the evaluator behavior cannot be modelled by few global parameters only.
The best way to overcome the evaluator dependency seems to be in having a big number of evaluators. In our results, the emotion-dependent class centroids in Fig. 2 could be interpreted as noisy estimates of a constant value.

Conclusion

In this paper we discussed the variability in expressing and perceiving emotions in speech. For this task we analyzed the class distributions of four emotion categories (angry, happy, neutral, sad) in the 3D emotion space, spanned by the emotion primitives valence, activation, and dominance.
Speaker modelling was introduced by two simple speaker-dependent parameters, the emotion expression bias (EEB) and the emotion expression amplification (EEA). This parametrization could easily be used to build speaker models for speaker-independent emotion recognition.
Evaluator modelling was shown to be more difficult. Future work will also focus on more sophisticated models for this task.

References

[1] F. Dellaert, T. Polzin, A. Waibel, *Recognizing Emotion in Speech.* Proc. ICSLP, 1996, pp. 1970-1973.

[2] R. Cowie, *Describing the Emotional States Expressed in Speech.* Speech Communications **40**, 2003, pp. 5-32.

[3] C.M. Lee, S. Narayanan, *Towards Detecting Emotions in Spoken Dialogs*, IEEE Trans. Speech and Audio Processing, **13 (2)**, 2005, pp. 293-303.

[4] R. Kehrein, *The Prosody of Authentic Emotions*, in Proc. Speech Prosody Conf., 2002, pp. 423–426.

[5] M. Grimm, K. Kroschel, *Evaluation of Natural Emotions Using Self Assessment Manikins*, Proc. IEEE WSh. ASRU, San Juan, Puerto Rico, 2005.

[6] M. Grimm, K. Kroschel, S. Narayanan, *Combining Categorical and Primitives-Based Emotion Recognition*, Submitted to ICASSP, Toulouse, France, 2006.

[7] S. Lee, S. Narayanan et al., *An Articulatory Study of Emotional Speech Production*, Proc. Eurospeech, Lisbon, Portugal, 2005, pp. 497-500.