Modeling Emotion Expression and Perception Behavior in Auditive Emotion Evaluation

Michael Grimm, Kristian Kroschel

Institut für Nachrichtentechnik (INT)
Universität Karlsruhe (TH)
Karlsruhe, Germany

Shrikanth Narayanan

Speech Analysis and Interpretation Lab (SAIL)
University of Southern California (USC)
Los Angeles CA, USA

Abstract

In this paper, we consider both speaker dependent and listener dependent aspects in the assessment of emotions in speech. We model the speaker dependencies in emotional speech production by two parameters which describe the individual's emotional expression behavior. Similarly, we model the listener's emotion perception behavior by a simple parametric model. These models form a basis for improving current automatic emotion recognition schemes such as, for example, for man-machine interaction applications.

For this task, an emotional speech database of the four emotion categories angry, happy, neutral, and sad was evaluated by 18 human listeners. For each of the 680 sentences, the evaluators rated the values of three emotion primitives, valence, activation, and dominance, each on a 5-point scale. The assessment results were used to calculate the distributions (centroids, covariances) of the emotion classes in the space spanned by the three emotion primitives. The individual classes formed separable clusters in the emotion space. Based on these results, we analyzed the variations of the emotion clusters as a function of speaker and listener.

Across different speakers, we found that the main difference in the emotional speech was the position of the neutral cluster and the scaling of the emotions in the emotion primitives space. To capture this effect, we introduced the speaker-dependent parameters *Emotion Expression Bias* and *Emotion Expression Amplification* within this model representation and showed that the original class centroids could be reconstructed fairly accurately. From the perception viewpoint, we found that the listeners' ratings of emotional speech could be described as a realization of a normally distributed random variable. Based on this result, we propose the correlation with the mean value of the ratings to be the listener-dependent parameter, which could be in turn incorporated within the model training for automatic recognition.

1. Introduction

Emotion recognition in speech has gained much attention in recent years [1, 2, 3, 4]. In many applications of man-machine interaction or data mining, it is not only important to determine *what* was said but also *how* it was expressed. To facilitate machine understanding of human emotional expressions, references are usually provided by human evaluation of the emotional speech. However, fine details of human evaluation of emotions has not been exploited to its full extent within the realm of automatic recognition of emotions. For instance, each person perceives emotions differently even if the evaluation context were to be the same and typically, majority agreement amongst the raters is assumed to be provide the reference. In this paper we analyze individual variations in both production and evaluation of emotional speech and propose parametric models of variability to obtain improved data representations and model for automatic emotion recognition schemes.

For the analysis of variations in producing and perceiving emotional speech, we use a 3D emotion space concept, where emotions are described as a combination of three values along continuous axes of generic attributes (referred herein as "primitives"). These primitives are *valence* (negative vs. positive), *activation* (calm vs. excited), and *dominance* (weak vs. strong) [5], each assumed to be defined to take on values, without loss of generality, within the range of -1 to +1.

Using these assessment results, the centroids and covariances of the distributions of the four emotion classes *angry, happy, neutral,* and *sad* are calculated. The individual classes have shown to form fairly minimally-overlapping 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 focus on the trasformation of the centroids of the emotion clusters with respect to speaker and listener effects.

The rest of the paper is organized as follows. Section 2 describes basic concepts of emotion expression and perception, and mentions previous work on incorporating them in emotion evaluation methods. Section 3 introduces the data we used. Section 4 describes the concept of building speaker models to describe speaker-dependent emotion expression behavior, and listener models to describe the emotion perception aspects. Section 5 shows the results achieved using these models with our speech corpus. Section 6 provides our conclusions and an outlook on future work.

2. Expression and perception of emotions

The Brunswick lens model depicts, in a functional manner, how emotions are transmitted from a speaker to a listener [7]: The truly felt emotion is expressed in a speaker-dependent way. It is transmitted by multiple modalities such as voice, mimics, gestures. The listener perceives the emotion as a listener-dependent receiver.

2.1. Production: Speaker-side

The expression of emotions is subject to various influences, such as gender, age, experience, or display rules [7]. They can be summarized as two major factors: *expressivity* and *encod-*

This work was supported by grants of the German Academic Exchange Service (DAAD) and the Collaborative Research Center (SFB) 588 of the Deutsche Forschungsgemeinschaft (DFG).

Contact: grimm@int.uni-karlsruhe.de

ing competence [8]. While expressivity does not necessarily reveal the speaker's feelings, e.g. if he/she generally speaks in an aggressive way, encoding competence describes the speaker's ability to accurately display them [8]. Emotion expression behavior can even be compromised, as found with alexithymia patients [7].

While these two factors are derived from a *descriptive concept* of emotion psychology, an alternative would be using an *explicative concept* of emotion psychology. The most important factors of a human's personality with respect to emotion are *extroversion vs. introversion* and *emotional lability vs. stability* [9].

Emotion recognition systems by now have only included one coarse factor of these influences, namely gender [10, 11, 12]. In the following we propose a concept on how to model a speaker's influence on emotion expression in a more explicit way.

2.2. Perception: Listener-side

The perception of emotions, too, is subject to many person-dependent influences. In addition to those mentioned in Sec. 2.1, cognitive aspects have to be taken into account. The listener's emotional state has an impact on how emotions are perceived. Since these influences impact how exact the initially expressed emotion is determined, they can be called *decoding competence* [7, 8].

Perception factors in emotion evaluation have previously been addressed. In auditive emotion evaluation, listener-dependent confidence scores were proposed [13]. Assessment of emotions was performed by merging weighted evaluations of several human listeners. Incorporation of such information about listener variability is important especially when human assessment serves as basis for training models for automatic emotion recognition systems.

3. Data

For this study, we used the *EMA Corpus* containing 680 sentences of emotional speech in American English [12]. Three native speakers, one professional actor (f), and two non-professional speakers (1f/1m) produced sentences (the professional produced 14 sentences while the others produced 10 sentences) in each of the emotions *Angry (A), Happy (H), Neutral (N), Sad (S)* with 5 repetitions each. The sampling rate was 16 kHz, with 16 bit resolution.

The sentences were evaluated using a text-free method using self assessment manikins as described in [13]. The evaluators rated the values of the emotion primitives on a 5-point scale each [13]. These values were normalized to the range of -1 to +1.

In total, 18 evaluators assessed the database (14m/4f). 9.7% of the sentences were discarded due to too high standard deviation in the evaluation (> 0.5). The average standard deviation of the remaining database 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.

4. Models

4.1. Speaker model

For the analysis of the speakers' variability in expressing emotions the average ratings of all human evaluators were compared. The centroids $\boldsymbol{\mu}(X, s)$ and covariances $\mathbf{C}(\mathbf{X}, \mathbf{s})$ were calculated for each speaker $s = 1, \ldots, S$ and for each of the acted emotion classes $X \in \{A, H, N, S\}$ separately. These speaker-dependent emotion class representations were compared to the average ("normalized") emotion classes

$$\boldsymbol{\mu}_0(X) = \frac{1}{S}\sum_{s=1}^{S}(\boldsymbol{\mu}(X,s) - \boldsymbol{\mu}(N,s)) \qquad (1)$$

$$\mathbf{C}_0(X) = \frac{1}{S}\sum_{s=1}^{S}\mathbf{C}(X,s). \qquad (2)$$

To model the emotion expression behavior, we introduce two speaker specific parameters:

- *Emotion Expression Bias* $\mathbf{EEB}(s)$ and
- *Emotion Expression Amplification* $\mathit{EEA}(s)$.

These parameters define a speaker's individual deviation from average emotion values calculated as a mean value of all speakers and all evaluators. $\mathbf{EEB}(s)$ is chosen as the speaker-dependent centroid of the neutral class,

$$\mathbf{EEB}(s) := \boldsymbol{\mu}(N, s). \qquad (3)$$

$\mathit{EEA}(s)$ is chosen as a scalar amplification factor to minimize the mean square error between the speaker's emotion class centroids and the model-based ones:

$$\mathrm{E}\left\{\left[\boldsymbol{\mu}(X,s) - \left(\mathit{EEA}(s)\cdot\boldsymbol{\mu}_0(X) + \mathbf{EEB}(s)\right)\right]^2\right\} = \min_{\mathit{EEA}(s)}. \qquad (4)$$

It is calculated using the individual matrix and vector elements,

$$\mathit{EEA}(s) := \frac{\sum\limits_X\sum\limits_i\left(\mu^{(i)}(X) - \mathit{EEB}^{(i)}\right)\cdot\mu_0^{(i)}(X)}{\sum\limits_X\sum\limits_i\left(\mu_0^{(i)}(X)\right)^2}, \qquad (5)$$

where $X \in \{A, H, N, S\}$ represents the emotion class, and $i \in \{\text{Valence, Activation, Dominance}\}$ represents the emotion primitive.

The emotion expression bias describes a general tendency in the speaker's emotion expression behavior. The emotion expression amplification captures the range of emotions the speaker produces.

4.2. Listener model

To assess the variability of the listener side in emotion evaluation, the centroids $\boldsymbol{\mu}(X, k)$ of the four emotion classes $X \in \{A, H, N, S\}$ are analyzed as a function of the evaluator $k = 1, \ldots, K$. The speaker dependency is mitigated by using the **EEB**-subtracted and *EEA*-rescaled evaluator-dependent centroids of these classes.

In the listener model we restrict to modeling the listener's emotion decoding competence by a single parameter. Without any further information on the listener (such as demographics), this decoding competence is best described as the similarity to the average of a whole set of evaluators. In particular, this approach is deemed reasonable, if the entity of all evaluators' ratings are normally distributed. Therefore the evaluator-dependent centroids of the four emotion classes are tested on the hypothesis

$$H_0: \text{They are a subset of a normally distributed population.}$$

The mean value and standard deviation of this normal distribution are not specified, since they are different for each emotion class. If these centroids are normally distributed, the similarity to the mean value, expressed by the correlation coefficient $\mathbf{r}(k)$, will be chosen as a parameter for the perception behavior of each evaluator.

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

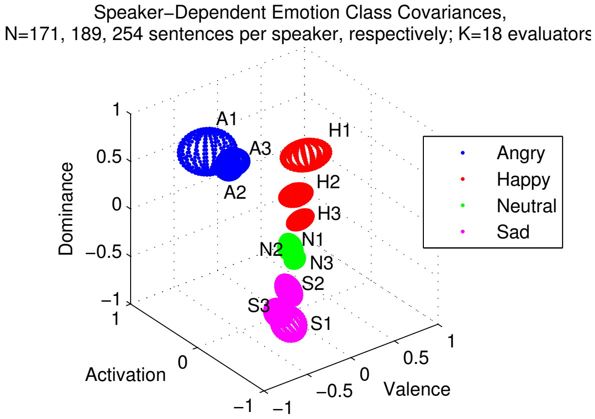

Figure 1: Covariance plot of the emotion classes *Angry* (A), *Happy* (H), *Neutral* (N), and *Sad* (S) of 3 speakers.

	Speaker 1	Speaker 2	Speaker 3
EEB	$\begin{pmatrix} -0.18 \\ -0.33 \\ -0.08 \end{pmatrix}$	$\begin{pmatrix} -0.15 \\ -0.26 \\ -0.14 \end{pmatrix}$	$\begin{pmatrix} -0.16 \\ -0.36 \\ -0.20 \end{pmatrix}$
EEA	1.26	0.84	0.89

Table 2: Results of the speaker model parameters: *Emotion Expression Bias* **EEB** and *Emotion Expression Amplification EEA*.

5. Results

5.1. Speaker variability

The speaker-dependent centroids and standard deviations of the emotions clusters in the 3D emotion space are reported in Tab. 1. These results indicate that the class distributions vary from speaker to speaker. The most significant differences were found in the centroids of happy emotion, where valence ranged from 0.12 for speaker 3 to 0.58 for speaker 1. Activation and dominance were both even smaller than 0 for speaker 3 (though very little), whereas it was 0.44 (activation) and 0.27 (dominance) for speaker 1. Similar observations can be stated for angry: Valence varied 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 showed his emotions in a more moderate way. The standard deviations varied to a lesser extent. Overall, the values were small and were in the range between 0.1 and 0.2. The standard deviation for neutral (speaker 3: 0.09) was 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).

The speaker-dependent parameters $\mathbf{EEB}(s)$ and $EEA(s)$ were calculated using (3) and (5), respectively. The results are reported in Tab. 2, where $\mathbf{EEB}(s)$ consists of the elements valence, activation, and dominance, respectively. The emotion expression bias was non-zero for all speakers, showing minor differences for the individual speakers. Interestingly, none of the speakers had a bias of positive valence, activation, or dominance.

The emotion expression amplification was found to be significantly different for the individual speakers. Speaker 1 showed an *EEA* greater than 1, signifying that her emotions were expressed in a more extreme way than average. Speaker 2 and speaker 3 showed *EEA* values smaller than 1. These speakers expressed their emotions less intense than average.

For validation, we reconstructed the initial emotion class centroids from the normalized emotion classes and the speaker-dependent models:

$$\hat{\boldsymbol{\mu}}(X,s) = EEA(s) \cdot \boldsymbol{\mu}_0(X) + \mathbf{EEB}(s). \qquad (6)$$

The average error,

$$e(s) = \sum_X \sum_i \left| \hat{\mu}^{(i)}(X,s) - \mu^{(i)}(X,s) \right|, \qquad (7)$$

of this reconstruction was 0.07, 0.02 and 0.07 for speaker 1, speaker 2, and speaker 3, respectively. Thus modeling the speaker's emotion expression behavior with these 2 parameters and a set of normalized emotions captured the variations fairly accurately for our corpus.

5.2. Listener variability

For our set of $K = 18$ evaluators, we analyzed the centroids $\boldsymbol{\mu}(X,k,s)$ of the emotion classes for each of the $S = 3$ speakers. The speaker-dependent influence was corrected by **EEB**-subtracting and *EEA*-rescaling,

$$\boldsymbol{\mu}(X,k) = \frac{1}{S} \sum_{s=1}^{S} \frac{\boldsymbol{\mu}(X,k,s) - \mathbf{EEB}(s)}{EEA(s)}. \qquad (8)$$

The result is shown in Fig. 2. Although there seem to be some common tendencies, a simple parametrization is not obvious. 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 modeled by a few global parameters only.

As described in Sec. 4.2, we tested each set of emotion centroids $\{\boldsymbol{\mu}(X,k)\}_{k=1,\ldots,18}$ $\forall X$ on the hypothesis of being normally distributed. Based on a Kolmogorov-Smirnov test for small datasets [14], we found that the null hypothesis cannot be rejected on a level of significance of $1 - \alpha = 0.99$ (maximum of test value: 0.89; critical value: $l^{norm}_{18;0.99} > l^{norm}_{10;0.99} = 0.94$).

	Evaluator																	
k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$r^{(Val)}(k)$	0.74	0.34	0.76	0.59	0.62	0.66	0.54	0.58	0.68	0.40	0.74	0.53	0.70	0.69	0.76	0.54	0.79	0.63
$r^{(Act)}(k)$	0.84	0.80	0.85	0.78	0.76	0.81	0.74	0.85	0.78	0.50	0.87	0.81	0.83	0.77	0.85	0.73	0.78	0.82
$r^{(Dom)}(k)$	0.86	0.86	0.83	0.77	0.57	0.77	0.49	0.79	0.84	0.67	0.88	0.65	0.83	0.62	0.86	0.71	0.84	0.66

Table 3: Results of the listener model parameter: correlation to mean value, $\mathbf{r}(k) = (r^{(Val)}(k), r^{(Act)}(k), r^{(Dom)}(k))^T$.

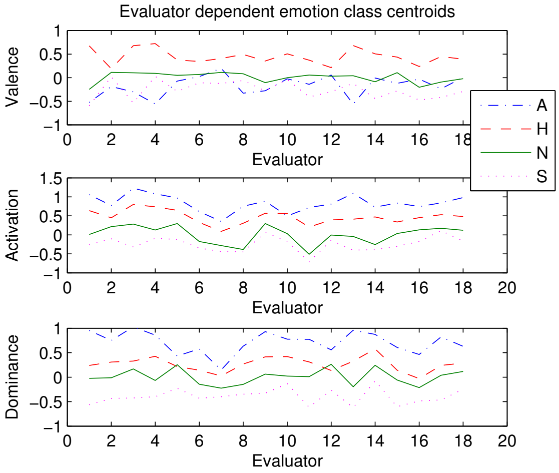

Figure 2: Comparison of evaluators' emotion perception influence on the centroids of normalized emotion classes.

Therefore we found strong evidence that the evaluators' ratings of the emotion class centroids were normally distributed.
As a result, the use of Pearson's correlation coefficient as a similarity measure between the individual evaluators and the average [13] is a good choice for comparative modeling of the evaluators' emotion decoding competence. In our case we found high decoding competence for most evaluators with correlation coefficients greater than 0.7. Valence in general showed smaller correlation coefficients than activation or dominance, c.f. Tab.3.

6. Conclusion

In this paper we discussed issues related to individual 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 modeling was introduced by two simple speaker-dependent parameters, namely the emotion expression bias $\mathbf{EEB}(s)$ and the emotion expression amplification $EEA(s)$ in the 3D emotion primitive space. This parametrization could easily be used to build models for speaker-independent emotion recognition. For the data of 3 speakers that were analyzed in this work, it was found that the two speaker-dependent parameters efficiently captured the influences of individual variations on emotion expression. The reconstruction error was very low, and insignificant. The method of describing personal emotion expression style in terms of speaker-dependent parameters could therefore easily be incorporated into automatic emotion recognition schemes. After determining the parameters of the speaker model, emotion estimates could easily be transformed to speaker-independent values which are necessary for accurately interpreting the emotion estimates.
Listener modeling was reduced to a single parameter, namely the correlation coefficient between emotion class centroids $\mathbf{r}(k)$, to describe the emotion decoding competence. This parameter can be derived from the correlation to the average rating of all evaluators. A comparative parameter can easily be derived to weight individual evaluator's ratings in emotion evaluation. Future work will focus on testing these models on a more exhaustive and realistic data set for emotion recognition as well as consider more sophisticated models for capturing emotion perception differences.

7. References

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

[2] R. Picard, *Affective Computing.* MIT Press, 1997.

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

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

[5] R. Kehrein, *The Prosody of Authentic Emotions*, in Proc. Speech Prosody Conf., 2002, pp. 423426.

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

[7] K.R. Scherer, *Ausdruck von Emotionen*, in K.R. Scherer (Ed.): "Psychologie der Emotion", Hogrefe Verlag Göttingen, 1990, pp. 345-422.

[8] J. Merten, *Einführung in die Emotionspsychologie*, Verlag W. Kohlhammer Stuttgart, 2003.

[9] H.W. Krohne, C.-W. Kohlmann, *Persönlichkeit und Emotion*, in K.R. Scherer (Ed.): "Psychologie der Emotion", Hogrefe Verlag Göttingen, 1990, pp. 485-558.

[10] M. Schröder, *Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, PhD Thesis, Universität des Saarlandes, Germany, 2003.

[11] D. Ververidis, C. Kotropoulos, *Automatic Speech Classification to five emotional states based on gender information*, Proc. EUSIPCO, Vienna, Austria, 2004, pp. 341-344.

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

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

[14] J. Hartung, *Statistik*, 13th ed., Oldenbourg Verlag München, Germany, 2002.