ON THE RELATIONS OF SEMANTIC AND ACOUSTIC PROPERTIES OF EMOTIONS

Kai Alter[1,2], Erhard Rank[4,2], Sonja A. Kotz[1], Erdmut Pfeifer[1],
Mireille Besson[3], Angela D. Friederici[1], Johannes Matiasek[2]

[1]Max-Planck-Institute of Cognitive Neuroscience Leipzig/Germany
[2]Austrian Research Institute for Artificial Intelligence (ÖFAI) Vienna/Austria
[3]Centre de la Recherche National (CNRS) Marseille/France
[4]Institute of Communications and Radio-Frequency Engineering,
Vienna University of Technology/Austria

ABSTRACT

In the present study, the relation between the semantic content and different acoustic parameters of emotional speech are analyzed. The analyzed corpus includes happiness, neutral state, and anger used as semantic content as well as emotional state of the speaker - also comprising mismatch conditions.

We have analyzed voice source quality parameters in a subset of the acoustic signals. The data indicate that the opening quotient and the glottal opening tendentially reflect differences between anger and neutral statements. Breathiness and roughness, estimated by the harmonic-to-noice ratio (HNR) show significantly higher HNR in the sentences expressed with anger than the neutral expressions.

1. INTRODUCTION

Current research on the properties of emotional states of speech has primarily focused on phonetic features, i.e., the realization of speech with a certain intended or perceived emotional state. However, it is important to also consider the semantic content of emotional expressions. We therefore set out to analyze acoustic features of neutral, happy and angry expressions in speech with controlled semantic content. It is expected, that emotion yields a significant change in the parameters affected by speaker arousal.

2. MATERIAL

The following examples (A-C) demonstrate conditions with a match between semantic content and emotional state:

(A) happy:
sie hat den Rekord gebrochen
she has the record broken [literal translation]
(B) angry:
er hat seinen Mitarbeiter entlassen
he has his co-worker fired [literal translation]
(C) neutral:
er hat auf dem Stuhl gesessen
he was on the chair sitting [literal translation]

The semantic content of the stimulus material (148 sentences) was rated by 30 subjects on a 5-point scale indicating whether a sentence expresses happiness (A), anger (B) or is emotionally neutral (C).

Semantic mismatch examples resulting from this scaling are listed below in (D-E):

(D) semantically unhappy, produced with a happy expression
er hat seinen Mitarbeiter entlassen
he has his co-worker fired [literal translation]
(E) semantically happy, produced with an angry expression
sie hat den Rekord gebrochen
she has the record broken [literal translation]

The neutral condition (C) was produced both with an angry and with a happy emotional state as well as the happy condition (A) and the angry condition (B) in a neutral manner.

In addition, the position of the focus accent is varied in all conditions from (A) to (E): (1) with accent on the NP immediately preceding the sentence final verb indicating wide/neutral focus and (2) with focus accent on the verb itself. Note that in (2) the NP preceding the sentence final verb is deaccented. Accented syllables are marked by capitals.

(1) sie hat den ReKORD gebrochen
she has the record broken [literal translation]
(2) sie hat den Rekord geBROchen
she has the record broken [literal translation]

All sentences were produced by a trained female speaker of Standard German in a sound proof room at the University of Leipzig. The combination of the different parameters described above formed a 2x3x3 paradigm with accentuation, semantic content and emotional state resulting in match and mismatch conditions between semantic content and emotional state.

3. ACOUSTIC CORRELATES OF EMOTIONS

The goal of our study is to detect properties of physiological processes during the production of different emotional states. Apart from the influence of emotional state on prosodic parameters (F0, speech rate), influences on the glottal pulse shape, harmonics-to-noise ratio, the mean spectral energy distribution and correlations with some kind of voicing irregularities have been proven [3].

For the choice of the acoustic features, we followed a study on the acoustic correlates of word stress in German by Classen et.al [2] which is essentially based on a method developed by Sluijter [6]. We examined two acoustic features for possible correlations with emotional speech: the Opening Quotient (OQ) and the Glottal Opening (GO), although Classen et.al [2] reported that no significant correlation between these parameters and stressed vs. unstressed vowels could be detected. Our hypothesis was that these parameters could reflect special properties of the vocal tract during the production of emotions, since spectral content is closely related to voice source quality parameters correlating directly to speaker arousal.

3.1 Open Quotient

The OQ is the relation between the glottal opening time and the duration of the whole glottal period ([2],pp.206f). A higher subglottal pressure is accompanied by an increase of the OQ. The variation of the OQ implies changes in the spectral content of the acoustic signal, especially in the low frequency range.

Based on our first observations, a positive emotional state like happiness induces a higher fundamental frequency (F0), a faster speech rate and more breathiness in comparison with a negative emotion and/or a neutral state. Therefore we suppose that happiness is produced with a higher subglottal pressure than angry or neutral statements.

3.2 Glottal Opening

The amplitude value of the first formant (F1) depends on the grade of the glottal opening during an opening cycle. The intensity of F1 therefore also depends on the OQ. We hypothesize that the GO decreases during the production of angry speech in comparison with the two other emotional states.

For the calculation of the OQ as well as for the GO, we followed the formula described in Classen 1998:214f.

The OQ was calculated according to:

$$\begin{aligned}
dH1 &= 20 \log_{10} (F1^2 / ((F1 + F0)\,(F1–F0))),\\
dH2 &= 20 \log_{10} (F1^2 / ((F1 + 2F0)\,(F1–2F0))),\\
H1^* &= H1 – dH1,\\
H2^* &= H2 – dH2,\\
OQ &= H1^* – H2^*.
\end{aligned}$$

These set of formulae depends on measurements of: fundamental frequency F0, frequency of the first formant (F1),. amplitude H1 of F1 (in dB), and amplitude H2 of the second harmonic (in dB).

The GO is computed by

$$GO = H1^* – A1,$$

where H1* can be used from the above formulae for the OQ, and A1 is the amplitude of the first formant F1 (in dB).

Although these two acoustic features do not reflect influences of word stress in German it might well be the case that they reflect other properties - namely emotion related properties.

4. ACOUSTIC MEASUREMENTS

4.1 Measurement conditions

The recorded corpus contains sentences with the typical German past tense verb final word order. The NP immediately followed by the final verb normally attracts the sentence accent. In order to avoid influences of accentuation on the acoustic features related to emotions, the speaker was also asked to move the accent in all conditions on the sentence final verb. Thus, the accented NP per default becomes deaccented and the final verb, e.g., the participle is accented. By this procedure we obtain both accented NP and accented verb.

The lexically stressed accented and deaccented syllables of the NP and of the verb were then extracted from the speech file. We analyzed a subset of 36 speech files in the following way: After recording and digitization at a 44100Hz/16bit sampling rate, the extracted syllables were downsampled at 8000Hz/16bit and filtered at 4000Hz. Via a wide band spectrogram (8ms/125Hz) the mid point in the steady state phase of the stressed vowel was fixed in both the accented and deaccented conditions for all the three states. In a narrow band spectrum (25ms/40Hz) we measured F0, H1 and H2, in a wide band spectrum (8ms/125Hz) we measured F1 and A1.

The analyzed 72 speech files are distributed over all conditions following the 2x3x3 design described above.

4.2 Results

Classen et al. [2] reported that the OQ is difficult to extract in cases where a female voice is producing high vowels because the frequencies of F0 and the F1 are too adjacent to each other. This phenomenon also turned up in our study not only in the case of the lax /i/ but also in the case of the vowel /o/. Around 80% of the OQ are missing values, especially for the high vowels in the happiness condition. This indicates that happiness is produced with an increasing F0.

The results for the GO in all conditions are highly inhomogeneous. Table 1 shows the results for 12 sentence pairs with different vowel types (V), the semantic state (sem), the vowel position in the sentence (VPos), and if the vowel was accented or not (±acc). The last three columns indicate whether the sentence was produced with happiness (pos), in a neutral manner (neut) or with irritation (neg).

V	sem	VPos	±acc	neg	neut	pos
a	neg	noun	+acc	–18.62	–5.40	–8.13
a	neg	noun	–acc	–16.61	–9.29	–7.63
i	neg	noun	+acc	–6.61	9.05	******
i	neg	noun	–acc	–5.11	4.04	–17.64
i	neg	noun	+acc	2.30	0.16	******
i	neg	noun	–acc	–5.33	8.36	–11.93
a	pos	noun	+acc	–18.59	–4.39	3.46
a	pos	noun	–acc	–14.56	2.58	1.89
o	pos	noun	+acc	–14.07	–3.69	0.17
o	pos	noun	–acc	–16.02	–0.74	4.50
i:	neut	noun	+acc	–17.81	–2.65	******
i:	neut	noun	–acc	–17.48	6.16	–4.42
o	neg	verb	–acc	–13.94	–0.03	–9.16
o	neg	verb	+acc	–10.55	–2.95	–5.45
o	neg	verb	–acc	–2.55	–0.93	–5.05
o	neg	verb	+acc	–7.93	–4.58	–26.08
a	neg	verb	–acc	–22.34	2.80	–26.42
a	neg	verb	+acc	–19.72	–8.25	–19.61
a	pos	verb	–acc	–8.38	1.81	–12.27
a	pos	verb	+acc	–10.37	–0.18	12.59
o	pos	verb	–acc	–11.09	–3.51	–11.58
o	pos	verb	+acc	–0.12	–6.56	7.17
i:	neut	verb	–acc	–4.08	–3.03	0.76
i:	neut	verb	+acc	6.08	1.72	–24.08

Table 1: Results for the GO. '*' indicates missing values.

The last three columns indicate whether the sentence was produced with happiness (pos), in a neutral manner (neut) or with irritation (neg).

Missing values for the GO appear in cases when the amplitude of H1 is not calculable because F0 and F1 are too closed to each other.

Comparing the data over the different conditions in Table 1, no clear tendency is detectable concerning the behavior of the GO. This can in fact be due to the high number of conditions. Looking more closely at single conditions, some tendencies of the behavior of the GO can be found.

In Table 2 a subset of conditions are listed comparing the relation of the GO in different acoustic realizations. In Table 2 examples are listed which show evidence that the GO in happily pronounced sentences decreases in comparison to their neutrally spoken variants. The column labeled with 'ne-p' shows the relation of the grade of the GO of sentences between neutral and positive acoustic conditions (=happiness). The '–' indicate that the GO in happily pronounced sentences decreases. As the other listed conditions show there is no relation between the mismatch conditions, the vowel quality, the accentuation and for the accent position. We can further observe that sentences which are pronounced with anger have a higher GO than sentences which are neutrally pronounced. This seems to be a general behavior also for the other sentences listed in Table 1. The label 'n-p' indicates the relation between the sentences spoken with negative (anger) and positive emotional content. Again, neither the whole analyzed subset from Table 1 nor the sentences presented in Table 2 show a clear tendency concerning the GO.

V	sem	VPos	±acc	n-ne	ne-p	n-p
a	neg	noun	+acc	+	–	+
i	neg	noun	–acc	+	–	–
i	neg	noun	–acc	+	–	–
a	pos	noun	–acc	+	–	+
i:	neut	noun	–acc	+	–	+
o	neg	verb	–acc	+	–	+
o	neg	verb	+acc	+	–	+
o	neg	verb	–acc	+	–	–
o	neg	verb	+acc	+	–	–
a	neg	verb	–acc	+	–	–
a	neg	verb	+acc	+	–	+
a	pos	verb	–acc	+	–	–
o	pos	verb	–acc	+	–	–

Table 2: Vowels with lower GO when pronounced happily.

We hypothesized that the comparison between the mismatch conditions should clarify the role of the GO in emotional speech. But again, our data does not confirm this hypothsis. There is no clear observable tendency.

5. FEATURES FOR BREATHINESS AND ROUGHNESS

Breathiness and roughness are often used as parameters in perception experiments concerning speech quality resp. emotional speech (e.g., [3]). A common acoustic correlate for both is the amount of noisy signal components in relation to the harmonic components. As an indicator we used an estimation for the harmonics-to-noise ratio (HNR) on the one hand and the maximum prediction gain computed by means of the mutual information function on the other.

The HNR estimation algorithm [4] uses an assessment for the harmonic components of the speech signal in the cepstral domain and by substraction from the original signal computes the noise component. Examples for estimated noise components of the vowel /a/ in the spectral domain are depicted in figure 1: the spectra show a realization in the verb of a sentence with accented NP (a) with negative emotional content, and (b) with neutral emotional content. The realization with negative emotional content shows a significantly higher HNR, as the statistical analysis proved as general quality.

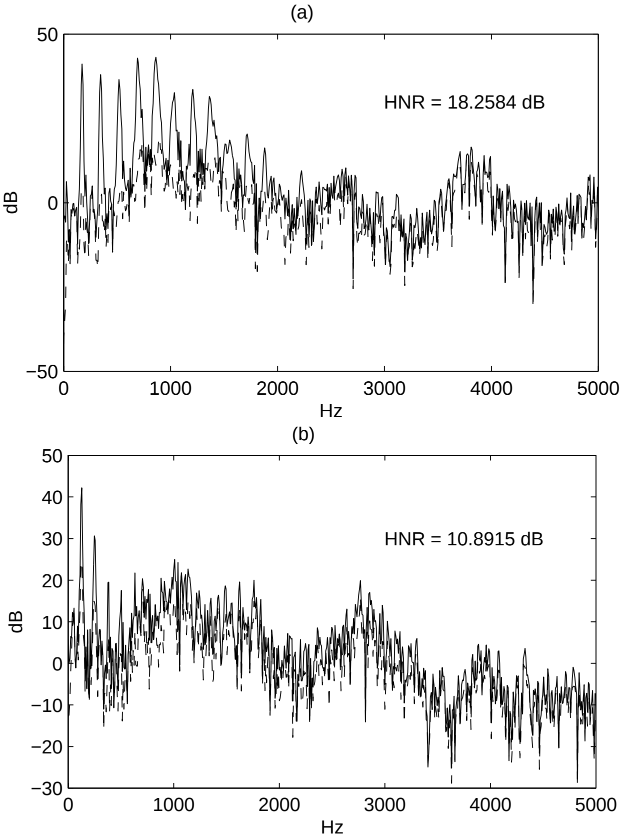

Figure 1: Spectrum of the original signal (solid line) and the estimated noise spectrum (dashed line) from vowel /a/ in the sentence final word. Condition: neutral semantic content, accent on the NP, and (a) negative emotional state, and (b) neutral emotional state.

The mutual information function is a nonlinear signal analysis method and provides a means to calculate the maximum achievable prediction gain—for both linear and nonlinear predictors—from a single signal realization [1]. It thus also yields a measure for the amount of harmonic (predictable) versus noisy (unpredictable) signal components.

It has to be noted that both the HNR estimation and the computation of maximum prediction gain do not solely constitute a measure for the noise present in the speech signal. They are also influenced by other attributes like frequency variations (jitter) or amplitude variations (shimmer) [5]. But it can be supposed that breathiness or roughness perceived by a human listener also depend on several speech attributes. This study thus has to be classified as analysis of acoustic features rather than an analysis of perceptual features.

Both analysis methods were applied to vowel segments of the sentence final verb of all possible conditions of semantic content, emotional state, and accentuation. Analysis was performed on the original signal at 44100 kHz sampling rate and 16 bit resolution. The results were evaluated in dependence of emotional state and accentuation only.

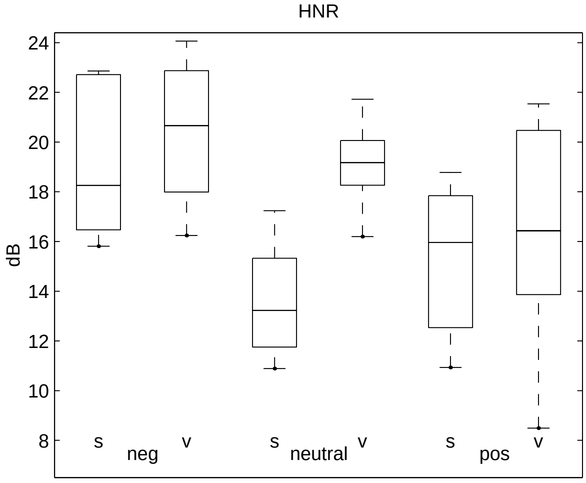

Figure 2: Results of the statistical analysis of the harmonics-to-noise ratios (HNR) for all three emotional states and accented/unaccented vowels in the sentence final verb. For each case the box ranges from the lower to the upper quartile with the median value indicated by a line, and the total value range is indicated by the whiskers. Higher HNR is found for accented (`v') cases, and also for nonneutral emotional state compared to the unaccented neutral state.

The analysis shows that the measure for the HNR (figure 2) and for maximum prediction gain (figure 3) can not be put on a level with each other. The HNR shows the expected higher values in the case of accentuation (labelled `v') for all different emotional states, and for the nonneutral compared to the neutral emotional states. This is due to higher speaker arousal in these cases. Unlike expected, the analysis by the mutual information function does not resemble the properties of the HNR. On the contrary: lower maximum prediction gain values are found for nonneutral emotional state than in the neutral case. The mutual information function also seems to be far less sensitive to accentuation than the HNR (c.f. the median values for each emotional state in figure 3).

6. DISCUSSION

To summarize, the utilized HNR estimation algorithm allows for distinction between accented and unaccented vowels on the one hand and at least between neutral and negative emotional state if unaccented vowels are examined. Computation of maximum prediction gain using the mutual information function also gives a clear distinction between neutral and negative emotional state which is virtually independent of accentuation.

The lack of comparable data for the OQ might be threefold: Either this parameter is not sensitive for emotional content of speech, or the influence of this parameter is not detectable because of the high number of conditions. On the other hand, the parameter could be influenced by the properties of high vowels bearing low F1. Only the great number of missing values in the acoustic happiness condition gives evidence for the fact that this emotional state is produced with a F0 which reaches the range of the F1 especially for high vowels. These circumstances do not allow a calculation of the OQ.

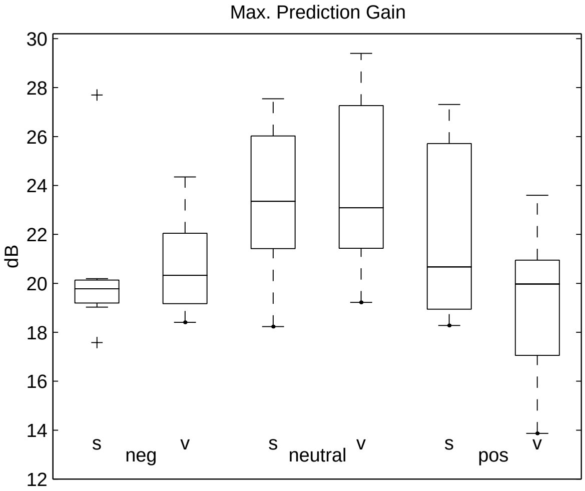

Figure 3: Results of statistical analysis of the maximum prediction gain (c.f. figure 2). Nonneutral emotional state yields a generally lower maximum prediction gain than neutral speech, and accentuation has a remarkable low influence.

The GO allows the interpretation that a pronounced irritation shows higher amplitudes for the F0 and the F1 and that the glottis is more closed for this condition. Further analysis of other spectral parameters such as spectral tilt is certainly needed for a better comprehension of the acoustic properties in emotional speech.

ACKNOWLEDGMENTS

We thank Björn-Helmer Schmidt for his assistance in generating the speech signals. This work was supported by the Deutsche Forschungsgemeinschaft (FR 519/17-1) and by the Austrian "Fonds zur Förderung der wissenschaftlichen Forschung", Grant No. P13224 as well as by the Austrian Federal Ministry of Science and Transport (BMWV), Grant No. GZ 607.521/2-V/B/9/98. ÖFAI is supported by the Austrian BMWV.

REFERENCES

[1] Bernhard, H.-P.: A Tight Upper Bound on the Gain of Linear and Nonlinear Predictors for Stationary Stochastic Processes. IEEE Transactions on Signal Processing, vol. 43, Nov. 1998.

[2] Classen, K., Dogil, G., Jessen, M., Marasek, K., Wokurek, W.,: Stimmqualität und Wortbetonung im Deutschen. in: Linguistische Berichte 174. Westdeutscher Verlag. pp. 202-245. 1998.

[3] Klasmeyer, G.: The Perceptual Importance of Selected Voice Quality Parameters. in: Proceedings of ICASSP'97, Munich, Germany, 1997.

[4] Krom, G. de: Acoustic Correlates of Breathiness and Roughtness. PhD-thesis, published by LEd, Utrecht, 1994.

[5] Pinto, N. B.: Unification of perturbation measures in speech signals. JASA, vol.87, nr.3, pp.1278-1289, 1990.

[6] Sluijters, A.M.C.:Phonetic Correlates of Stress and Accent. PhD Dissertation. The Hague: Holland Academic Graphics. 1995.