PERCEPTION AND ACOUSTICS OF EMOTIONS IN SINGING

Susan Jansens, Gerrit Bloothooft, and Guus de Krom
Computer and Humanities Department / Utrecht Institute of Linguistics-OTS
University of Utrecht, Trans 10, 3512 JK Utrecht, the Netherlands
Tel: + 31 30 2536059, Fax: + 31 30 2536000, E-mail: Guus.deKrom@let.ruu.nl

ABSTRACT

In this experiment, the acoustic correlates of perceived emotions in singing were investigated. Singers were instructed to sing one phrase in a *neutral* way and in the emotions *anger*, *joy*, *fear*, and *sadness*. Listeners rated the strength of the perceived emotions for each fragment. Principal component analyses were performed on the listeners' ratings. The derived factors were interpreted as listening strategies; and a listener's factor loading as an indicator of the extent to which that listener used that strategy. Using the original ratings and the factor loadings, the phrases were assigned composite ratings for each emotion. Acoustic measures of spectral balance, vibrato, duration and intensity were related to the composite ratings using multiple regression analyses. It was found that *anger* was associated with the presence of vibrato; *joyous* phrases had vibrato, a short final duration, and a shallow spectral slope; *sadness* was associated with absence of vibrato, long duration, and a low intensity, whereas *fear* was related to a steep spectral slope.

INTRODUCTION

In this study, we wanted to investigate the acoustic correlates of different perceived emotions in singing.
The results obtained in previous studies have suggested that the perception of different emotions depends in a complicated way on a variety of acoustic parameters [1, 2]. Kotlyar and Morozov [1] used different phrases for the different emotions, which makes it hard to distinguish a possible effect of the type of phrase from the effect of emotion. Sundberg et al. [2] used a professional singer who was asked to sing different phrases in an "emotional" and "neutral" way. The focus of that study was therefore not so much on the difference between the emotions, but rather between neutral end emotional singing.
To eliminate a possible interaction effect of phrase and emotion, we decided to use one phrase that was sung in a number of emotions by a variety of listeners. Our aims were twofold: first, to determine whether listeners can actually perceive the different emotions intended by a singer. Second, if so, to determine the acoustic correlates of the differently perceived emotions.

EXPERIMENTS

Recordings: material, singers and procedure

The material used in this study had been collected previously [3]. For this study, we used recordings of 14 professional singers (7 males, 7 females) who were each asked to sing a part of "*Der Erlkönig*" by Schubert. To ensure that the singers did not deviate too much in their pitch, they listened to a complex signal of the prescribed fundamental frequency (208 Hz, males; 370 Hz, females) before singing a particular fragment.
From this material, we selected the phrase "*Mein Vater, mein Vater, und hörest du nicht, Was Erlkönig mir leise verspricht...*?". The phrase was sung in the following emotions by all singers: *anger*, *joy*, *fear*, and *sadness*. In addition, the phrase was sung in a *neutral*, "emotionless" way. The phrase was considered suitable for the expression of different emotions because it can be given different semantic interpretations.

I. Perceptual evaluation of emotions

Twenty-five non-professional listeners (13 females, 12 males) were asked to rate the strength of the perceived emotions for each of the 70 (14 singers × 5 emotions) fragments. The listeners were paid for their co-operation. The perception experiment was carried out in sound-treated booths.
Stimuli were presented in random order over headphones using an event-driven computer program [4]. During the presentation of the stimuli, four sliders were projected on a computer screen: each was labelled with a particular emotion. By moving the sliders with a computer mouse, the listeners could indicate the strength of the perceived emotion. The leftmost position indicated an emotionless = neutral rating, the rightmost position an emotion that was perceived as maximally strong. The listeners were free to use whatever combination of sliders they considered appropriate. We did not want to restrict ratings to one dominant emotion, because we considered it plausible that certain stimuli give rise to a combined perception of emotions, like *anger* and *fear*, or *fear* and *sadness*. The selected slider positions were logged and converted to ratings with values between 0 (emotionless) and 100. Thus, for every stimulus, all four emotions were rated. Care was taken that the fragments were presented at appropriate

intensities reflecting the SPL differences measured during recording (recording intensities were calibrated, [3]). Stimuli were repeated automatically until the listener pressed a "next stimulus" button. Upon the presentation of a new stimulus, the four sliders were automatically reset to the leftmost "neutral" position. The listeners participated in a short training session using ten other, but comparable, sung fragments to get used to the task.

Results perceptual evaluation of emotions

First, mean ratings were calculated on the basis of the 1750 (25 listeners × 5 emotions × 14 singers) "raw" ratings. As might be expected, individual singers differed markedly in their ability to express the intended emotions; the mean ratings and standard deviations varied widely within the group of singers. Similarly, means and standard deviations of ratings within the listeners also indicated large inter-rater variability.
However, when data were averaged across listeners and singers, the intended emotion always got the highest mean rating of all possible emotions, indicating that -on average- the singers had succeeded in expressing the intended emotion, and that the listeners had -on average- been able to perceive and label these emotions correctly[1]. We therefore felt confident to conclude that the listeners had in fact been able to perceive the emotions intended by the singers, although different rating strategies had probably been used.

II. Multiple regression analyses

The second aim of this study was to determine the acoustic correlates of the perceived emotions by means of multiple regression analyses, with ratings of emotions as dependent variables, and acoustic data as predictors.

Acoustic analyses
A total of 20 acoustic parameters were determined for each of the 70 phrases. The parameters included measures of duration, intensity, vibrato, spectral slope and accuracy of fundamental frequency.
Four duration parameters were determined (in s): the total phrase duration, and the duration of three vowels in the phrase: the vowel /a:/ in the first instance of the word *Vater*, the vowel /ϕ/ in the word *hörest* and the vowel /i:/ in the phrase-final word *nicht*.

[1] Fragments sung in a *neutral* manner got a relatively high mean *sadness* rating, indicating that the perception of emotion was biased in this phrase. Mean ratings on *anger* and *sadness* were relatively high, indicating that the listeners were more confident about these ratings than for the *joy* and *fear*.

Intensity (in dB SPL) was determined for the entire phrase and the aforementioned three vowels /a:/, /ϕ/, and /i:/.
Measures of vibrato frequency and extent were manually determined for the three vowels on the basis of an F0 trace obtained with the signal analysis programme. Vibrato frequency (in Hz) was determined on the basis of the number of observed vibrato cycles in the F0 trace and the vowel duration. Vibrato extent (in semitones) was determined as follows: first, local maxima and minima were determined in the F0 traces. Mean F0 was calculated on the basis of these maxima and minima. The difference between the maximum and mean F0 value was determined in Hz, and subsequently converted to semitones (relative to local mean F0).
For each of the three vowels, a simple spectral slope measure (in dB) was based on the intensity difference of two frequency bands, one ranging between 60 and 2500 Hz, another between 3000 and 8000 Hz.
A measure of fundamental frequency deviation (in Hz) was defined for each of the three vowels as the difference between F0 as prescribed in the musical score and the observed mean F0.

Determination of composite emotion ratings
For each phrase, we had 25 ratings on the different emotions, but only one measured value of a particular acoustic parameter. This discrepancy in the data had to be corrected. The most obvious solution would be to average ratings across listeners, reducing the 25 ratings to one mean rating. However, this would not be appropriate, considering the large inter-rater variability. Instead, we decided to look for patterned rating behaviour in the data, reflecting different rating strategies.

Principal Component Analyses of emotion ratings
First, a 25 × 25 correlation matrix was determined for the listeners' ratings on a certain emotion . Next, a principal component analysis with varimax rotation was performed on the correlation matrix [5], yielding a number of independent factors. Every listener now had a specific loading on each factor: listeners who tended to rate in a similar way had similar loadings on similar factors. Thus, the different factors reflected different rating strategies. The factor loadings of individual listeners could be interpreted as a an indicator of the extent to which the ratings of a particular listener complied with a particular rating strategy.
The results of the principal components analyses indicated that the ratings could be described by at least six independent factors, which further proves that a simple averaging across listeners would have been inappropriate. The first factor accounted for 61, 27, 38, and 26% of the variance in the ratings of *anger*, *joy*, *sadness*, and *fear*, respectively. The combination of all factors explained around 90% of the rating variance for all types of emotions.

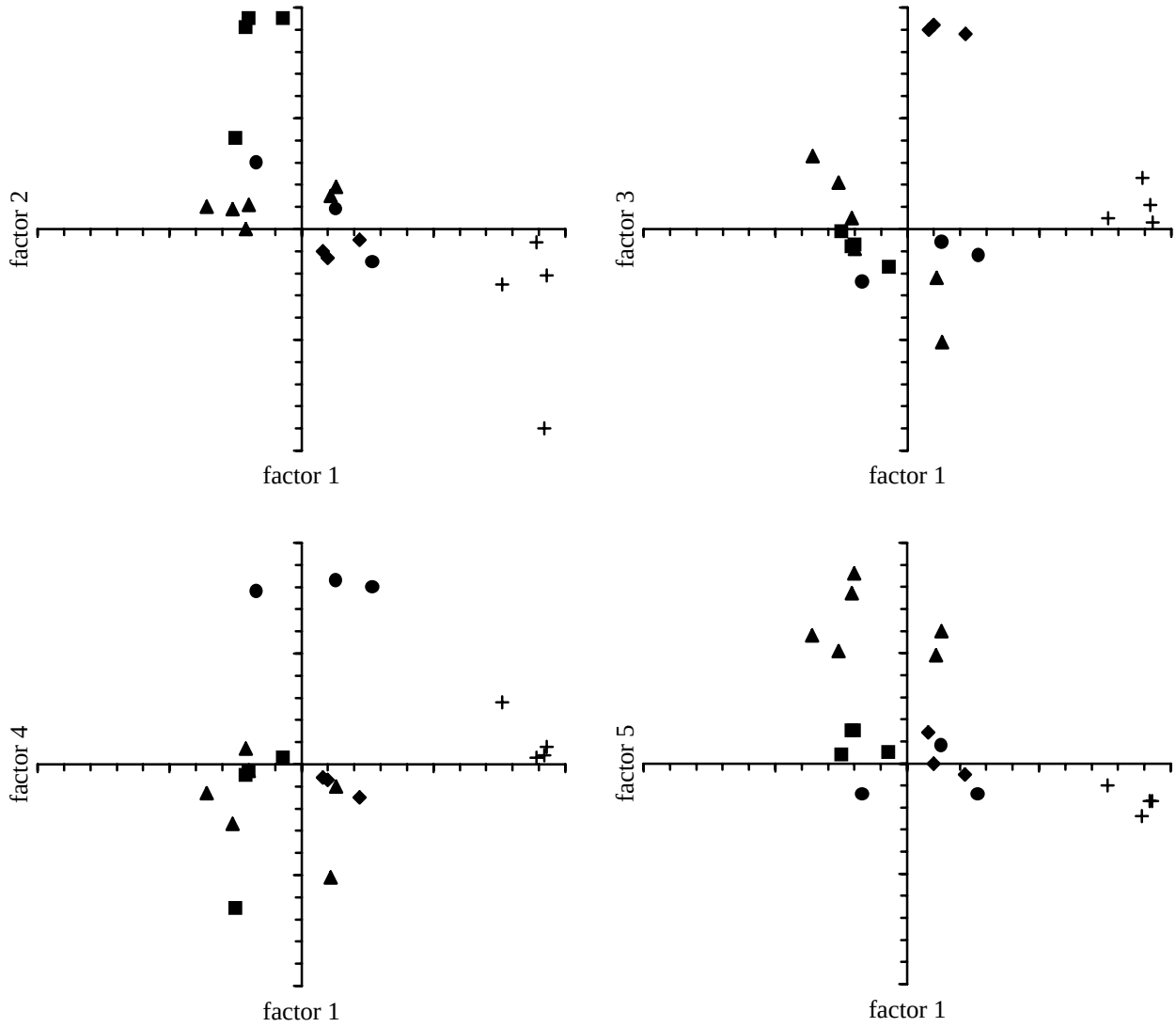

Figure 1. Factor loadings of grouped acoustic parameters on five dimensions. Plusses: duration measures; squares: intensity measures, triangles: vibrato measures; circles: spectral slope measures; diamonds: F0 deviation measures.

For our purpose -the construction of a single composite rating for each emotion on each fragment- we used the loadings on the first factor only. This factor explained most of the variance in the ratings, and was therefore considered to reflect the most dominant rating strategy. The composite ratings were calculated as follows: For each fragment, the product of a listener's "raw" rating and her / his first factor loading was calculated. Next, the products were added for all listeners. The resulting sum was finally divided by the number of listeners. This was done for each emotion separately. The obtained composite ratings weighed a listener's original rating on a particular emotion to the extent that she / he had used the "first factor" rating strategy

Results multiple regression analyses

Using the composite ratings and the acoustic data, stepwise multiple linear regression analyses were performed for each emotion. A combination of five predictors explained 60% of the variance in the composite *anger* ratings. The selected predictors were vibrato extent of the vowels /ɸ/ and /a:/, the intensity of the vowel /i:/, and the spectral slope of the vowels /a:/ and /ɸ/. For *joy*, 24% of the rating variance was explained by three predictors (vibrato extent of /i:/, spectral slope of /a:/, and duration of /i:/). Four predictors explained 67% of the variance in the composite *sadness* ratings (total phrase duration, duration of /a:/, vibrato frequency of /a:/, and SPL /i:/). Ratings of *fear* could be related to just one predictor, the

spectral slope of /i:/, which accounted for 23% of the rating variance. Results are summarised in Table 1.

Table 1. regression equations for the different emotions (composite ratings).

emotion	predictors and coefficients
anger	-80 + 9.2 vibrato extent /ɸ/ + .74 SPL /i:/ + .82 slope /a:/ - .35 slope /ɸ/
joy	6.3 + 4.4 vibrato extent /i:/ -.22 slope /a:/ - 2.6 duration /i:/
sadness	2.4 total duration - .59 vibrato frequency /a:/ - .14 duration /a:/ - .19 SPL /i:/
fear	.26 slope /i:/

Interpreting the signs and weights of the regression coefficients in Table 1 is hazardous. The signs can only be interpreted meaningfully if the predictors in the model are uncorrelated. Second, the magnitude of a regression coefficient can only be interpreted (and then even to some extent) if the different predictors are measured in the same units. The latter was obviously not the case in this study. In addition, it was to be expected that correlations existed between several of the acoustic predictors, as they could be grouped in duration measures, vibrato-related measures, and so forth.

Principal Components Analyses of acoustic data

To facilitate interpretation of the results given in Table 1, the acoustic data were subjected to a principal component analysis. The analysis extracted five orthogonal factors, accounting for 78% of the variance in the acoustic data. Factors one to five explained some 30, 18, 11, 10, and 9% of the variance, respectively. As can be observed in Figure 1, the duration measures had high loadings on the first factor. All intensity measures, except intensity of the vowel /i:/, loaded highly on factor two. The parameters that indicated a deviation in fundamental frequency had high loadings on factor three. Spectral-slope parameters loaded highly on factor four, while all six vibrato measures had high factor five loadings.

Based on the data in Figure 1, we concluded that ratings of *anger* could be meaningfully related to the vibrato measures only; as the other predictors (intensity of the vowel /i:/ and the two spectral slope measures) were not really independent of each other. The positive signs of the regression coefficients of the two vibrato predictors (vibrato extent of the vowels /ɸ/ and /a:/) indicated that *anger* was associated with the presence of vibrato.

The predictors of *joy* ratings (vibrato extent of /i:/, spectral slope of /a:/, and duration of /i:/) did in fact load on different factors. *Joyous* phrases could be characterised by the absence of vibrato, a shallow spectral slope, and a short phrase final duration.

Predictors of *sadness* (total phrase duration, duration of /a:/, vibrato frequency of /a:/, and SPL /i:/) also loaded on different factors. *Sad* phrases were characterised by a long duration, absence of vibrato, and a low phrase-final intensity. Ratings of *fear* were related to a steep spectral slope in phrase final position.

CONCLUSION

The first aim of the experiment was to determine whether listeners could perceive different emotions in sung fragments. The results of the perception experiment revealed that individual listeners differed widely in their ratings of a given stimulus. On average, however, the emotions intended by the singers were correctly recognised ("*angry*" phrases got relatively high *anger* ratings, and so forth). The principal components analyses on the perceptual data provided further evidence that listeners had used different strategies in rating the emotions. Composite ratings were determined on the basis of the principal components analyses. Using these ratings, the different emotions could be related to a fairly distinct combination of acoustic parameters. However, considering the fact that these composite ratings were based on only one of the possible rating strategies, the results of the regression analyses have to be interpreted with some caution. The fact that different listeners may employ different rating strategies could mean that the outcomes of acoustic-perceptual studies strongly depend on the perceptual strategy or strategies employed by the listeners. Further research, preferably involving analysis-by-synthesis techniques, is needed to investigate the full range of possible relationships between the acoustics and perception of emotion for different groups of listeners.

REFERENCES

[1] **Kotlyar, G., & Morozov, V**. (1976). Acoustical correlates of the emotional content of vocalized speech, *Soviet Physics Acoustics*, *22*, 208-211.

[2] **Sundberg, J., Iwarsson, J., & Hagegård, H**. (1995). A singer's expression of emotions in sung performance. in: *Vocal Fold Physiology, voice quality control*, 217-231 (O. Fujimura & M. Hirano, Eds.). San Diego: Singular.

[3] **Bloothooft, G.** (1996). "Vowels in Concert". CD-ROM, Speech Processing Expertise Centre (SPEX), Leidschendam, the Netherlands.

[4] **Veenker, T.** (1997). *FEP 2.0 User's guide and reference manual*. Utrecht Institute of Linguistics-OTS.

[5] **Ferguson, G.A., & Takane, Y**. (1989). *Statistical analysis in psychology and education*. New York: McGraw-Hill.