VOICE PRINTS AS A TOOL FOR AUTOMATIC CLASSIFICATION OF VOCAL PERFORMANCE

Claus Weihs
Uwe Ligges
Fachbereich Statistik, Universität Dortmund, Germany

ABSTRACT

In order to find objective criteria for the assessment of the quality of vocal performance, time series of voice generated vibrations (so called waves) were measured in a standardized experiment (Weihs et al., 2001).

We are interested in properties of such time series related to performance quality aspects of single tones like purity of intonation, vowel purity, vibrato intensity, solidity of tone, and softness / brilliance of tone.

Based on a segmentation of the waves into notes (Ligges, 2000 and Ligges et al., 2002), the individual tones are judged from the estimated periodogram of the sung notes. Intonation accuracy is estimated by the halftone distance from the ideal tone. In order to analyze the other above quality aspects, we consider the widths and the heights (amplitudes) of the periodogram peaks corresponding to the fundamental frequency and the first twelve overtones.

From these measures we derive spectral characteristics of voices (and instruments) (Güttner, 2001) and cluster voices and selected instruments into homogeneous groups. This way, it is, e.g., possible to separate the voice types.

Moreover, voice prints for vocalists are derived and compared to corresponding instrument prints of selected instruments.

1. INTRODUCTION

Sound characteristics of orchestra instruments derived from spectra are currently a very important research topic (see, e.g., Reuter, 1996, 2002). The sound characterization of voices has, however, many more facets than for instruments because of the sound variation in dependence of technical level and emotional expression (see, e.g., Kleber, 2002). Particularly the emotional component of singing drastically extends the possible sound properties, whereas the sound properties of instruments are much more restricted by their physical properties. This paper presents first steps of the sound analysis of voices compared to some selected instruments.

The paper is structured as follows. In Section 2 the data our results are based upon are introduced. Section 3 compares voices and instruments by means of characteristics of their spectrum. Section 4 defines voice and instrument prints. Finally, the paper is concluded in Section 5.

2. DATA PREPARATION

In this paper we analyse time series data from an experiment with 17 singers performing the classical song “Tochter Zion” (Händel) to a standardized piano accompaniment played back by headphones (cp. Weihs et al., 2001). The interpreters could choose between two accompaniment versions transposed by a third in order to take into account the different voice types.

Voice and piano were recorded at different channels in CD quality, i.e. the amplitude of the corresponding vibrations was recorded with constant sampling rate 44100 hertz in 16-bit format. The audio data sets were transformed by means of a computer program into wave data sets. For time series analysis the waves were reduced to 11025 Hz (in order to restrict the number of data), and standardized to the interval [-1,1]. Since the volume of recording was already controlled individually, a comparison of loudness of the different recordings was not sensible anyway. Therefore, by our standardization no additional information was lost.

Since our analyses are based on measures derived from single tones, we used a suitable segmentation procedure (Ligges et al., 2002) in order to get data of segmented tones.

The periodograms (cp. Brockwell and Davis, 1991) used for the analyses described in this paper were calculated from overlapping sections of 2048 observations, overlap starting in the middle of the preceding section. This way, we get roughly 11 (= 2 * (11025 / 2048)) periodograms per second of sound, whereas the duration of the whole song is roughly 60 seconds.

3. CLUSTERS AND TREES

Let us now consider the clustering of the singers concerning global differences in their voices. The periodogram itself is not appropriate to serve as a basis for clustering, since the differences in the height of the sung tones of the different voice types hinder comparability. What is needed are characterizations of the periodogram, or a better estimator of the spectrum, corresponding to location deviation, weight and form of the peaks which are independent of the voice type.

$$\mathrm{DHT} = 12 \log_2 \left(\frac{\text{estimated frequency(tone)}}{\text{ideal frequency(tone)}} \right)$$

In order to determine the location of peaks it is sufficient to estimate the fundamental frequency since then the locations of the overtones are determined also. Therefore, as a first measure the half tone distance DHT of the estimated and the ideal frequency is determined:

ISBN 3-931852-67-9
ISSN 1617-6847

Note that the ideal frequency has to be adapted to the actual frequency of the diapason a' which was estimated to be 443.5 Hz in our case.

In order to measure the size of the peaks in the spectrum, the mass (weight) of the peaks of the fundamental frequency and the first 12 overtones are determined as the sum of the percentage shares of those parts of the corresponding peak in the spectrum, which are higher than a pre-specified threshold.

The shape of a peak can often not easily be described. Therefore, we only use one simple characteristic of the shape, namely the width of the peak of the fundamental frequency and the first 12 overtones. The width of a peak is measured by the half tone distance between the smallest and the biggest frequency of the peak with a spectral height above a pre-specified threshold.

Overall, every tone is characterized by the above 27 characteristics which are used as a basis for clustering. For details on the computation of the measures see Güttner (2001).

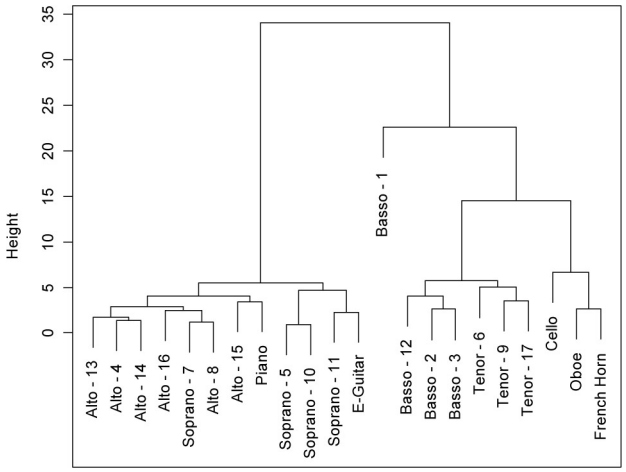

Figure 1: Hierarchical clusters of voices and instruments (Ward method)

These characteristics are also computed for selected instruments, namely cello, e-guitar, French horn, oboe, and piano. The tones analysed for these instruments are standardized versions of the tones in "Tochter Zion", alto version, taken from the "McGill University Master Samples". The Euclidean distances of the tones of the singers and the instruments corresponding to the above 27 characteristics build the basis for the so-called 'Ward clustering method' (see Ward, 1963). This method is aiming at clusters as homogeneous as possible, i.e. at clusters of minimum variation. In order to give all characteristics the same weight, they are mean centred and normalized by their standard deviations before hand. The results are illustrated by means of a dendrogram in Figure 1, where the vertical axis (Height) indicates the distance between the various voices and instruments.

Obviously, the voice types are clustering nearly ideally together, and the male professional opera singer (Basso 1) builds a 'singular branch' in the dendrogram. Thus, the spectral characteristics appear to build a reasonable first step for the discrimination of voices. Concerning the instruments, piano and e-guitar cluster together with female voices, whereas the other instruments built an individual cluster nearer to the male voices. Note that the tone height was eliminated by the way computing the characteristics, and the location of the tone should be ideal (corresponding to diapason a') for the instruments. Thus, similarities are only based on peak heights and widths.

Cluster analyses are called unsupervised since the methods do not have any background knowledge to judge the quality of grouping. On the other hand, there are also supervised classification methods trying to reproduce a pre-defined grouping by means of so-called classification rules. For such methods the classification quality can, e.g., be measured by means of the so-called misclassification rate, i.e. the ratio of the wrongly classified cases to the overall number of cases. We applied the easily interpretable so-called 'classification tree' (more specifically RPART by Therneau and Atkinson, 1997) to our data, estimating the error rate by the 'leave-one-observation-out method'.

In most applications the quality of the outcomes depends on the relation of the number of groups to the number of cases. Using mean characteristics over the involved notes we, obviously, have 22 cases (17 singers and 5 instruments), and we consider two different situations concerning the grouping. First we try 7 groups, the 4 voice types, bowed instruments (cello), wind instruments (oboe, French horn), and plucking instruments (e-guitar, piano). Second we try 2 groups only, female voices together with all instruments, and male voices. Note that instruments 'play' the alto version of the song. In contrast to the cluster analysis, we used non-normalized characteristics in order to get interpretable rules in the classification tree.

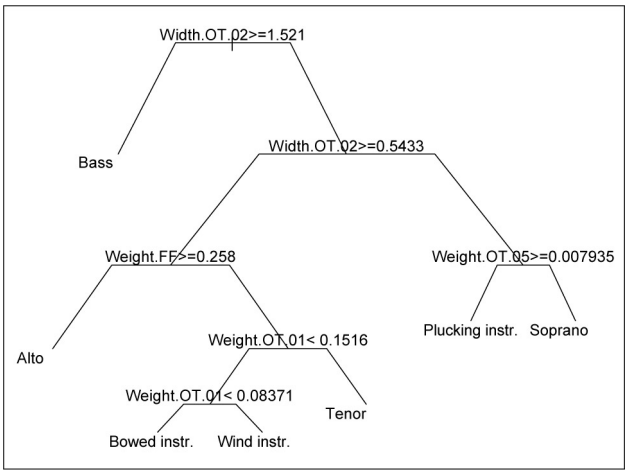

Figure 2: Classification tree (rpart) for seven classes

Unfortunately, for the 7 groups (cp. Figure 2) the error rate was estimated to be 50%, much too high for accepting the tree to be useful. Here the widths of the second overtone's peak (accordingly labelled 'Width.OT.02') are relevant in the upper 2 nodes, whereas weights (e.g. 'Weight.FF' denotes the weight at the fundamental frequency) are relevant in the other nodes.

For the 2 groups, however, the tree shown in Figure 3 produced no error at all. The width of the first overtone's peak is sufficient

ISBN 3-931852-67-9
ISSN 1617-6847

to classify males into the group 'males', and females together with instruments into the group named 'females'. Obviously, covering more than 2.7 halftones, that width is quite large for tenor and bass singers.

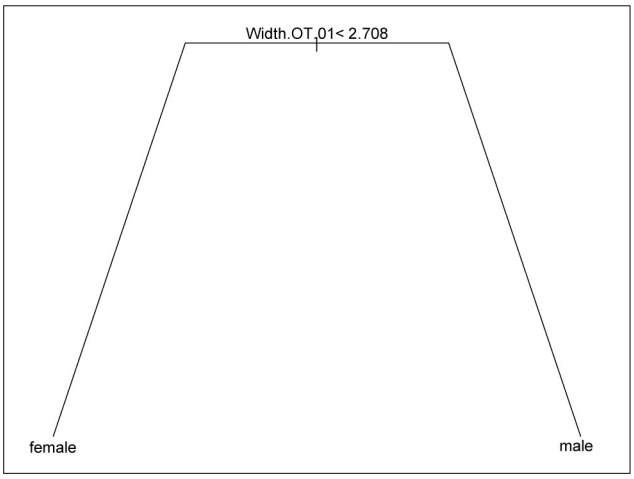

Figure 3: Classification tree (rpart) for two classes

4. VOICE AND INSTRUMENT PRINTS

In order to fully characterize a voice, one possibly should not restrict oneself to characteristics like in section 2, but should consider at least the whole spectrum. On the other hand, in order to be able to compare different voices, the spectrum should be independent of the voice types, as in section 2. This led us to the construction of what we call voice (and instrument) prints.

A voice (or instrument) print is derived from the spectrum in the following way. The locations of the fundamental frequency and the overtone frequencies are determined, and the centers of the peaks are shifted to a pre-defined location at the x-axis. Also, the width of the peaks is adjusted to take into account the logarithmic halftone distance scale. This can be achieved by simply dividing the frequencies by the known fundamental frequency of a note.

Obviously, the this way adjusted spectra, called voice (or instrument) prints, can be compared, even between voice types. Figures 4 to 7 show examples: the male and the female professional singer, a female amateur singer, and the French horn.

It appears that the wind instrument has more important overtones than the female singers. The amateur female singer emphasizes the fundamental frequency much stronger than the professional one. Wider peaks regularly indicate more vibrato rather than narrow peaks. The peaks produced by the bass singer appear to be rather flat, because weight is spread over more than nine overtones (the plot is cut off at the tenth overtone for visualization reasons).

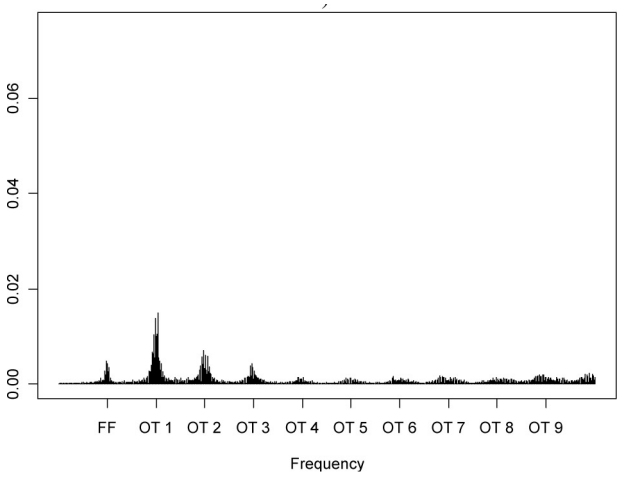

Figure 4: Voice print of the male professional (basso)

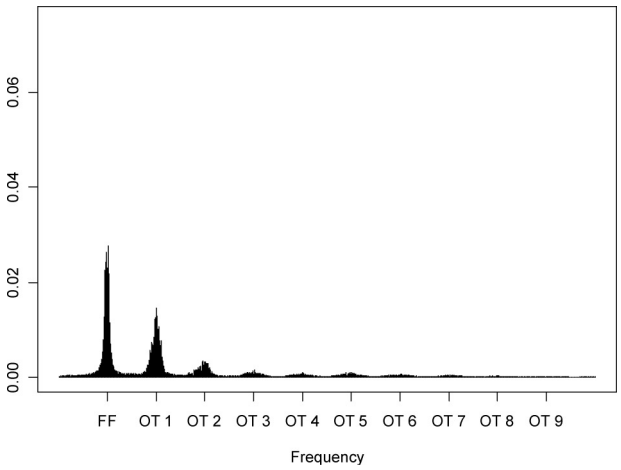

Figure 5: Voice print of the female professional (soprano)

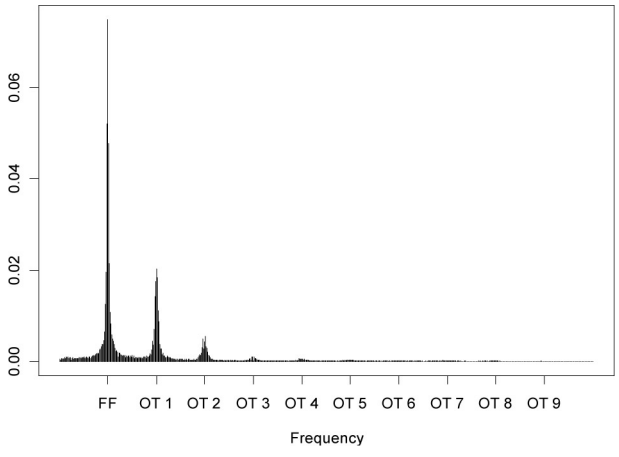

Figure 6: Voice print of a female amateur (soprano)

ISBN 3-931852-67-9
ISSN 1617-6847

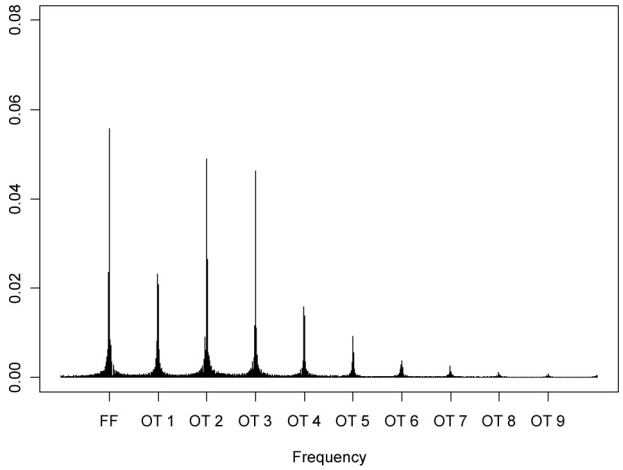

Figure 7: Instrument print of the French Horn

5. CONCLUSION

We succeeded in finding first sound characterizations of voices and instruments on the basis of spectra, namely weight and width of peaks corresponding to the fundamental frequency and the first 12 overtones, as well as so-called voice and instrument prints. As next steps we are planning to derive quality characteristics concerning vowel purity (for voices only), vibrato intensity, solidity of tone, and softness / brilliance of tone from the voice and instrument prints and the spectrum in general. For this purpose formant intensity and amplitude variation of the different overtones is planned to be studied next.

6. REFERENCES

1. Brockwell, P. J. and Davis, R. A. (1991). Time Series: Theory and Methods. New York: Springer.

2. Güttner, J. (2001). Klassifikation von Gesangsdarbietungen. Diploma Thesis, Fachbereich Statistik, Universität Dortmund, Germany.

3. Kleber, B. (2002). *Evaluation von Stimmqualität in westlichem, klassischen Gesang*. Diploma Thesis, Fachbereich Psychologie, Universität Konstanz, Germany.

4. Ligges, U. (2000). Identifikation lokal stationärer Anteile in Gesangszeitreihen. Diploma Thesis, Fachbereich Statistik, Universität Dortmund, Germany.

5. Ligges, U., Weihs, C. and Hasse-Becker, P. (2002). Detection of Locally Stationary Segments in Time Series. In W. Härdle and B. Rönz (eds.), *COMPSTAT 2002 – Proceedings in Computational Statistics – 15th Symposium held in Berlin, Germany* (pp. 285-290). Heidelberg: Physika Verlag.

6. *McGill University Master Samples.* McGill University, Quebec, Canada. URL: http://www.music.mcgill. ca/resources/mums/html/index.htm

7. Reuter, C. (1996). Die auditive Diskrimination von Orchesterinstrumenten - Verschmelzung und Heraus-hörbarkeit von Instrumentalklangfarben im Ensemblespiel. Frankfurt/M: Peter Lang.

8. Reuter, C. (2002). Klangfarbe und Instrumentation - Geschichte - Ursachen - Wirkung. Frankfurt/M: Peter Lang.

9. Therneau, T.M. and Atkinson, E.J. (1997). An Introduction to Recursive Partitioning Using the RPART Routines. *Technical Report,* Mayo Foundation.

10. Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, pp. 236-244.

11. Weihs, C., Berghoff, S., Hasse-Becker, P. and Ligges, U. (2001). Assessment of Purity of Intonation in Singing Presentations by Discriminant Analysis. In J. Kunert and G. Trenkler (eds.), Mathematical Statistics and Biometrical Applications (pp. 395-410). Köln: Josef Eul.

ISBN 3-931852-67-9
ISSN 1617-6847