ON THE COMPARISON OF COMPUTERISED PHONETOGRAMS

Gerrit Bloothooft, Margreeth van Dorssen, Kim Koppen, Peter Pabon and Mieke van Wijck
Utrecht Institute of Linguistics OTS, The Netherlands (Gerrit.Bloothooft@let.uu.nl)

ABSTRACT

In automated phonetogram registration, several acoustic voice parameters are computed as a function of fundamental frequency and vocal intensity. We show how such a phonetogram can be described by a set of Hidden Markov Models. This allows a comparison of phonetograms utilizing all underlying data. The method makes it possible to recognize new phonetograms in terms of the most likely phonetogram that could have generated the new data. Promising implications for increased clinical usefulness of phonetograms are discussed.

1. INTRODUCTION

A phonetogram is a presentation of values of acoustic voice parameters plotted as a function of fundamental frequency and vocal intensity. In traditional manual recording the parameters were limited to just fundamental frequency and vocal intensity themselves, but in computerized recordings it is possible to enrich the registration with a series of additional parameters such as jitter [1]. Although phonetogram plots are appealing, their use is limited if we do not have a means to compare them to each other. We want to know, for instance, whether voice therapy has improved a patient's voice, or whether singing training has widened the vocal capacities in range and timbre. Often it is not only the contour of the phonetogram that is informative. This contour, comprising the loudest and softest phonations possible over the whole range of fundamental frequencies, is for instance heavily dependent on the subjects' motivation. Still, this contour is the basis of so-called norm phonetograms [2-7] and is unlikely to be reliable in detail. It can be expected, however, that major vocal changes will manifest themselves more clearly within the normal range of phonation. They show up in acoustic parameters that describe the temporal and spectral characteristics of phonation throughout the phonetogram. No techniques have been described yet that compare phonetograms in this respect. Such a comparison can be put more simply as the question of whether two phonetograms have been produced by the same voice or by different voices. This opens the way to the probabilistic approach that is introduced in this paper. The paper is organized as follows: we first describe computerized phonetogram registration; then our method for comparison of phonetograms is explained. Finally, the method is tested experimentally and discussed.

2. COMPUTERISED PHONETOGRAM RECORDING

In our method of computerized phonetogram recording we measure simultaneously and in real time values of five acoustic parameters: fundamental frequency (F_0), vocal intensity (I), jitter, crest factor, and relative rise time [1]. Jitter reflects the temporal perturbations in the periodicity of phonation, the crest factor measures the ratio between the maximum amplitude in a voice period and the average energy (RMS value), while the relative rise time gives the time from the beginning of the period to its maximum as a fraction of period time. The latter two parameters are linked to spectral properties of the signal. In the extreme case of a pulse-like signal, the crest factor becomes very large and the relative rise time very short, while the spectrum is flat. For a sine wave, with just a single harmonic, the crest factor is at minimum (3 dB), while the relative rise time increases to .25.

Phonetogram recordings are made in a sound treated booth. A subject wears a headset to which a small microphone (B&K 2032) is attached at a distance of 25 cm from the mouth. The microphone signal is amplified and fed into a personal computer with a fast DSP board. At a rate of 70/s the acoustic measures are computed and plotted on the computer screen. The horizontal axis displays fundamental frequency in semitones, the vertical axis vocal intensity in dB. Each sample is displayed as a small rectangle on the screen in a position defined by F_0 and I. A third parameter can be shown in addition by the color of the rectangle. Usually we do not use one of the three remaining acoustic parameters, but the number of hits at a particular combination of F_0 and I, the vocal density. This gives an indication whether sufficient data has been measured. The other acoustic parameters are stored, of course, and can be displayed on command in the same manner.

After a short introductory training session, in which the subject learns the meaning of the axes on the screen and their relation to phonation, the subject sings as many combinations of fundamental frequency and vocal intensity as his or her voice is capable of. A systematic strategy using swell tones, with F_0 increasing semitone by semitone from one swell tone to another is most efficient but often only possible for trained singers. Untrained subjects just experiment with the voice but succeed in the end in realizing all vocal possibilities as well. Most difficult are the controlled production of the loudest and softest phonations and the highest and lowest fundamental frequencies. It depends very much on the motivation of the subject to what extent these extreme values are reached. The total recording takes about 20 minutes. Figure 1 shows an example.

The computer screen can only show a limited part of the data. Besides F_0 and I on the principal axes, only one parameter can be shown, and from this parameter only a single value at some combination of F_0 and I. This could be, for instance, the best (lowest) jitter value (reflected as a color on the screen). In previous phonetogram recordings, all but the best jitter values at some F_0/I combination were actually lost. This means that the distribution of the jitter parameter remains unknown, which is disastrous for any post-hoc statistical comparison. Therefore, we have adapted the data handling in such a way that all sampled data (70/s) are stored in the computer as well. Our investigations on the comparison of phonetograms are based on these full data. A complete phonetogram recording usually consists of between 50.000 and 100.000 samples.

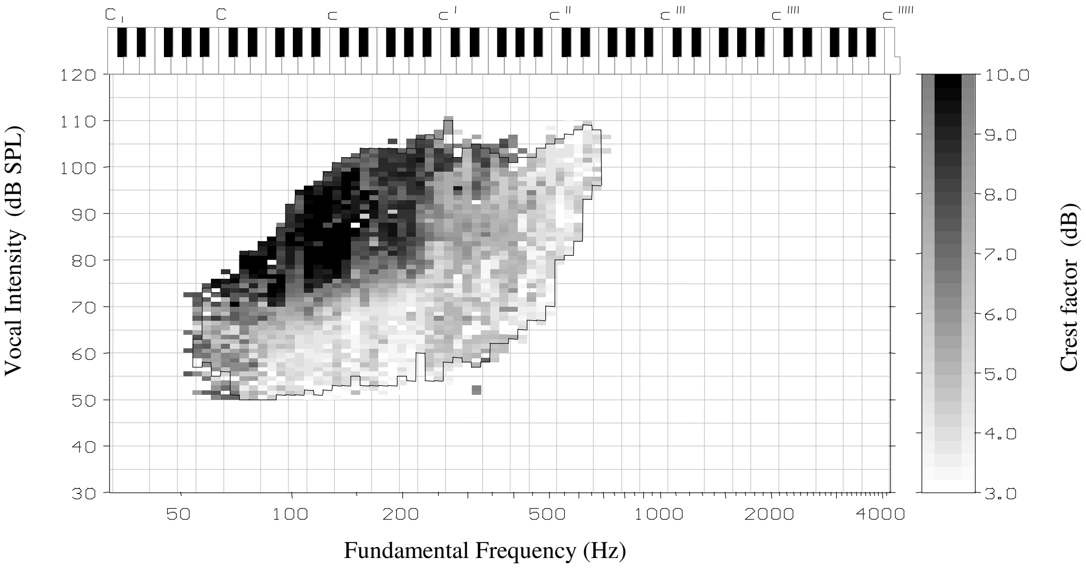

Figure 1. Phonetogram of a male subject showing the value of the crest factor by means of a gray scale.

3. COMPARING PHONETOGRAMS

A comparison of phonetograms should focus on similarities and differences for every combination of fundamental frequency and vocal intensity. The simplest differences between subjects are those where one subject can sing at some combination of F_0 and I and another subject can not. More complex differences may occur when two subjects can both sing at some combination of F_0 and I, but with differences in additional acoustical voice parameters. These differences will be reflected in the distributions of the acoustical voice parameters if these parameters are complete with respect to the description of the voice. With the current stage of knowledge and proposed parameters this is not the case, although the parameters we use give a fair initial indication. Nevertheless, the precise relationship between values of acoustic voice parameters and vocal fold physiology is not known either. This implies that on the basis of the present acoustic parameters we do not always know what mechanism of voice production resulted in some phonation. These mechanisms may be very different as is the case, for instance, with the modal and falsetto registers. At some ranges in voice production, a subject can sing either in modal register or in falsetto register at the same F_0 / I combination, and the precise register is not always known. Under these conditions, it seems that a Hidden Markov Model (HMM) is very appropriate for modeling the observed acoustic parameters in the phonetogram at some combination of fundamental frequency and vocal intensity. We have chosen a simple model with two independent states (figure 2). This means that once the HMM has entered in one state, the other state cannot be reached anymore. This is true for vocal registers where a subject always has to change fundamental frequency and vocal intensity before returning to the same values in a different register. It is assumed that the values of an acoustic parameter are normally distributed in anyone state and that a description with a single Gaussian mixture is sufficient. Three separate streams for jitter, crest factor, and relative rise time are used. We did not use limitations with respect to transitions from one HMM to another, i.e. from one F_0 / I combination to another.

This approach implies that the phonetogram is modeled as a set of HMMs, one for each combination of fundamental frequency and vocal intensity. Since the average range of F_0 is about three octaves (36 semitones) and the dynamic range may extend to 40 dB, a total number of about 1500 HMMs can be envisioned. This is too large given the number of data available for training. We therefore enlarged the ranges of F_0 and I that are modeled by a single HMM to a honeycomb shaped area of 6 dB by 6 ST (see figure 4). The total number of HMMs needed for modeling a full phonetogram is then limited to about 100.

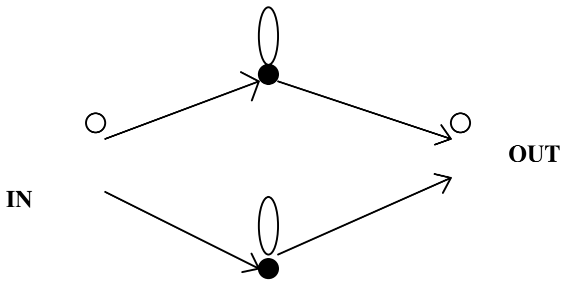

Figure 2. Two-state Hidden Markov Model architecture for the description of acoustic voice parameters within a limited range of fundamental frequency and vocal intensity.

The procedure for comparing phonetograms is the following. The data that were collected during the recording of phonetogram P_1 are used to train a set of HMMs. In the test phase, the probability is computed that the data from another phonetogram P_q were generated by the set of HMMs from P_1. This procedure becomes very interesting once one has a large resource of phonetograms P_1 to P_n. Then for a new phonetogram P_q the probabilities can be computed that the data were generated by each of the phonetograms P_1 to P_n. The phonetogram with the highest probability wins and is the best matching candidate. This becomes particularly useful when the resource of phonetograms describes, for instance, a wide range of voices with a known medical history ranging from vocal fold pathologies to normal voices. The new phonetogram P_q could then be associated with the best matching phonetogram for which extended medical information is already available.

It is not even necessary to have a full phonetogram available for testing. The question what phonetogram best explains new data can be answered for any amount of new data. This can be very useful for the purpose of screening. Of course, the more data are available the more accurate the prediction will be.

It may occur that there are test data with F_0/I values for which no HMM is available. This would immediately drop the probability to zero. This is not always the right approach however. Although a missing model can point to a significant characteristic of a voice in that a certain fundamental frequency is outside the range, it can also be the accidental result of incomplete training. We decided to avoid a possible erroneous interpretation in this respect and only tested data for which an HMM is available.

4. A TEST OF THE MODEL

4.1. Method

We tested the proposed procedure by asking 12 male subjects to make three phonetogram registrations in three different sessions. The subjects were vocally untrained and did not report voice pathologies. They were between 25 and 49 years of age. The data of two phonetograms were used for the training of the HMMs with standard software from the Entropics HTK Toolkit. The data of the third phonetogram were used for testing. The aim was to investigate whether the a subject's HMMs described his own test data best, and whether there were specific combinations of fundamental frequency and vocal intensity that contributed most to that end.

4.2. Distributions underlying the phonetogram

Before actually building the Hidden Markov Models we first looked into some properties of the distributions of the acoustical variables. Figure 3 shows the overall distribution of the crest factor (across all combinations of F_0 and I) for two subjects, both for the training data and for the testing data. It shows that there is good agreement between the distributions for training and testing for a single subject, but that there are clear differences between distributions from different subjects. The latter is of course promising for further processing of the data using the HMMs. Comparable results were found for the other parameters.

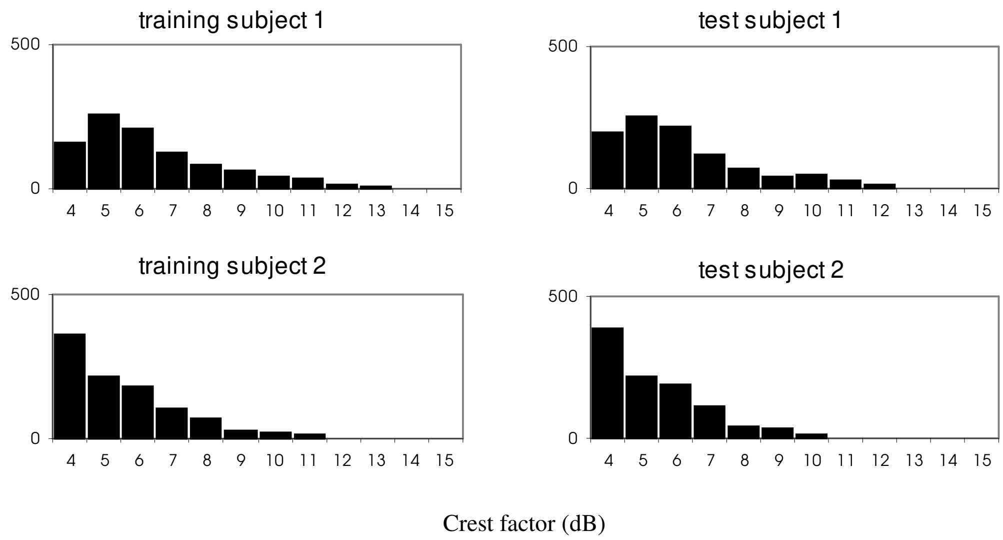

Figure 4. Distribution of the crest factor across all values for fundamental frequency and vocal intensity for two subjects. Data from phonetograms used for training (left-hand panels) and for testing (right-hand panels)

4.3. Results

We tested phonetogram data of each of the 12 subjects against all available phonetogram models, including their own model. For each phonetogram model, this resulted in a probability that that model generated the test data. It showed that for all subjects their own phonetogram model had the highest probability.

In addition, we wanted to know which sub-areas in the phonetogram were most distinctive between subjects. To this end we computed the probability that a subject's own HMM could predict his own test data in the sub-area, and compared this value with the average probability for the eleven other HMMs from the other subjects. This difference across all subjects is plotted as a grayscale within the phonetogram plane in Fig.4. It shows that especially the outer areas of the phonetogram are distinctive between subjects. This concerns the skill (or the lack of skill) to sing softly at higher pitches, to sing loudly in modal register, and to sing in the upper range of the modal register.

5. DISCUSSION

We have demonstrated a new method for the comparison of phonetograms that promises to have a high potential in increasing the (clinical) usefulness of phonetogram registration. It has been shown to be possible to distinguish subjects on the basis of elementary acoustic voice parameters such as jitter, crest factor and relative rise time, in areas of the phonetogram that were common to the subjects. Our approach can be refined by adding acoustic voice parameters from the spectral domain and by extending the modeling with probability estimations of transitions between sub-areas in the phonetogram. In combination with the prospect of a resource of phonetograms with known medical histories this would create exciting new research potentials and applications.

REFERENCES

[1] Pabon, J.P.H. (1991). Objective voice-quality parameters in the computer phonetogram. *J. of Voice*, 5, 203-216.

[2] Böhme, G. & Stuchlick, G. (1995). Voice profiles and standard voice profiles of untrained children. *J. of Voice 9*, 304-307.

[3] Gramming, P. (1988). *The phonetogram, an experimental and clinical study*. Malmö, PhD dissertation.

[4] Schultz-Coulon, H.J. & Asche, S. (1988). Das "Normstimmfeld"- ein Vorschlag. *Sprache-Stimme-Gehör 11*, 5-8.

[5] Sulter, A.M., Wit, H.P., Schutte, H.K. & Miller, D.G. (1994). A Structured approach to voice range profile (phonetogram) analysis. *J. Speech and Hearing Research 37*, 1076-1085.

[6] Hacki, T., Frittrang, B., Zywietz, C. & Zupan, C. (1990). Verführen zur statistischen ermittlung von Stimmfeldgrenzen - Das Durchschnittstimmfeld. *Sprache-Stimme-Gehör 14*, 110-112.

[7] Heylen, L. (1997). Normfonetogrammen voor kinderen (6-11) en leerkrachten met gezonde stemmen. In: *The clinical relevance of the phonetogram*, PhD dissertation, Antwerpen [in Dutch], 86-101.

[8] Bloothooft, G. and Binnenpoorte, D. (1998). Towards a searchable resource of phonetograms. *Proceeding Voicedata'98*, UiL-OTS Utrecht, 112-116.

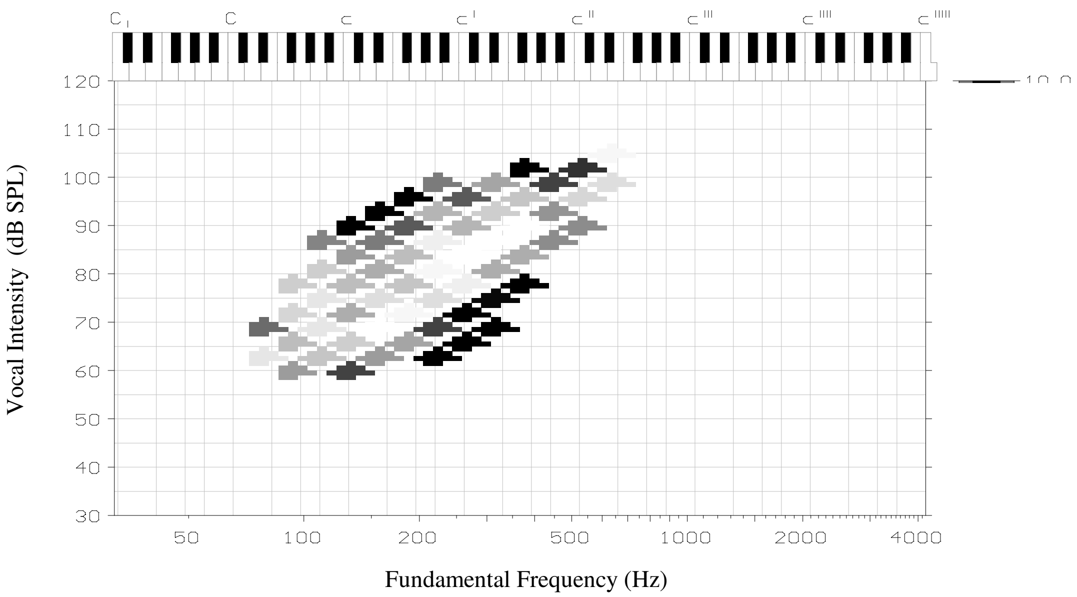

Figure 4. Sub-areas in the phonetogram modeled by a single HMM and common to at least nine of twelve male subjects. The arbitrary grayscale indicates whether the sub-area is capable of distinguishing between subjects. The darker the sub-area, the larger this capacity.