Proceedings of The Institute of Acoustics

# ACOUSTIC CORRELATES OF AUDITORY RATINGS OF SOME VOCAL SPEECH PARAMETERS

B. CRANEN, R. VAN BEZOOIJEN AND L. BOVES

INSTITUTE OF PHONETICS, CATHOLIC UNIVERSITY NIJMEGEN, THE NETHERLANDS

## INTRODUCTION

In various research areas, such as speech pathology, pragmalinguistics and social psychology, use is made of auditory ratings for describing vocal characteristics of running speech. Aim of our research is to investigate to what extent auditory impressions can be represented by acoustic properties of the speech signals. The emphasis in this contribution will be on the acoustic analysis and data processing techniques.

## SPEECH MATERIAL AND AUDITORY RATINGS

The speech data base was comprised of 160 emotional expressions, namely 8 (4 male and 4 female) speakers x 2 phrases ('two months pregnant', /tve: ma:ndə zvɑŋər/ and 'such a big American car', /zo:n ɣro:tə ɑmerika:nsə o:to:/) x 10 emotions (neutral, disgust, surprise, shame, interest, joy, fear, contempt, sadness, anger). The recordings were put in a random order per speaker and per phrase, and rated by six slightly trained judges on 13 vocal parameters. Six of these, namely overall pitch level, pitch range, harshness, loudness/effort, laryngeal tenseness and laryngeal laxness, will be dealt with in this paper.

To collect the ratings use was made of preprinted successive interval scales. All scales were considered as absolute. The only exception to this rule was the pitch level scale which was effectively split up into separate scales for women and men. Every utterance could be given only a single rating on each scale. This means that the scores represent some kind of perceptual average. The reliability of the means of the ratings was assessed by means of the so-called Ebel coefficient (Ref. 1). All coefficients ranged between .90 and .95, values which may be considered to be gratifyingly high. Therefore, in the subsequent analyses use was made of the scores for each of the 160 utterances, averaged over the six transcribers.

## ACOUSTIC MEASURES

If auditory ratings of vocal characteristics of speech are to be explained by acoustic measurements, we must look for long-term average values and their associated dispersions. The nature of the specific auditory parameters examined in this study suggests that we need data on central tendency of $f_o$, $f_o$ variation, and spectral properties of the signal. The acoustic measures used to explain the ratings were obtained as follows. The utterances were fed into a "critical band" spectrum analyser consisting of three filters with a constant 90 Hz bandwidth and center frequencies at 122, 215, and 307 Hz, and 16 1/3 octave filters with center frequencies from 400 Hz up and including 12500 Hz. The rectified and detected outputs of the filters were sampled every 10 ms and fed into a digital computer together with the pitch and voiced/unvoiced output of an analog pitch extractor.

ACOUSTIC CORRELATES OF AUDITORY RATINGS OF SOME VOCAL SPEECH PARAMETERS

The pitch data to which an error correction algorithm was applied were used to obtain one measure of central tendency of $f_o$, viz. the median and two measures of $f_o$ range, viz. the standard deviation of the log-converted (semi-tone) $f_o$ distribution and the variation coefficient, i.e. the standard deviation divided by the sample mean. From the original period duration data, before their conversion to $f_o$ samples at a 10 ms rate, two pitch perturbation measures were derived, i.e. two measures of short-term $f_o$ fluctuations which are supposed to be related to an auditory impression of roughness or harshness. Both measures were based on the distribution of differences between adjacent $f_o$ samples. The relative $f_o$ difference $df_n$ is defined by

$$df_n = \frac{f_n - f_{n-1}}{.5\ (f_n + f_{n-1})}$$

where $f_n$, $n = 1, 2, \ldots$ are $f_o$ values as read by the computer. The pitch perturbation measures are defined by

$$pp = \frac{1}{n} \sum_{n=1}^{N} |df_n|$$

The difference between the two measures employed in our study lies in the presence or absence of an error correction algorithm before the computation of $df_n$. In one version no error correction is employed, in the other a simple logic removes obvious outliers from the data before the $df_n$ are computed.

The spectral samples from the voiced signal segments were averaged to obtain long-term average spectra (LTAS) comprising 19 values, one for each filter output. Since the levels in adjacent filters tend to be fairly strongly correlated some kind of data reduction is called for. We decided to use the spectral slope measures as defined by Hammarberg et al. (Ref. 2). The maximum level in the octave band from 400 to 800 Hz is taken as the reference with respect to which all remaining levels are measured. This band usually contains the first formant, the level of which is decisive for the overall power of the speech wave. The first slope (perhaps one should better use the term spectral "distance", since "slope" would imply a division by frequency distance) is determined by the level in the filter with a centre frequency of 1600 Hz, and the third by the level at 5000 Hz. The last slope depends on the maximum level in the frequency region above 5000 Hz. Since our data are confined to voiced segments this level differs only very slightly from the level at 5000 Hz. Therefore, we have limited our data processing to the first three slope measures.

RESULTS

Ratings of pitch level explained by $f_o$ median

Since pitch level was rated on separate scales for female and male speakers we first computed correlations between $f_o$ median and auditory ratings within either sex group. The correlation was .86 for the males and .80 for the females. When the correlation between rating scores and $f_o$ median was computed for the combined group of 160 utterances the coefficient dropped to .75 which suggests that the scales did indeed differ to some extent. From the fact that the coefficient rose

again to .80 if the $f_o$ medians were converted to z-scores we may conclude that a considerable part of the difference between the pitch level scales for female and male speakers is confined to a linear transformation.

Ratings of pitch range explained by measures of $f_o$ dispersion

The correlation between the variation coefficients and the standard deviation of the log-converted $f_o$ distributions amounted to .88. It is not very surprising therefore, that the correlations of both measures with the rating scores are almost equal: .68 for the variation coefficient and .64 for the semi-tone standard deviation. Apparently, neither measure does too well in predicting the ratings. Since the variation coefficient does not involve the extra computational complexity of a log-transform we decided to prefer it as the measure to be employed in the future.

Ratings of harshness explained by measures of $f_o$ perturbation

For the total of 160 utterances the correlation of the perturbation computed from the raw $f_o$ data with harshness ratings amounted to .42. For the perturbations based on the error-corrected $f_o$ the correlation is as low as .12. From these results it may be concluded that neither measure is a very good predictor of perceived harshness. The one measure that might be considered probably is more an indication of the problems that the pitch extractor had in tracking the $f_o$ in the utterances than a measure of short-time $f_o$ fluctuations.

Ratings of loudness, tension and laxness explained by LTAS measures

Measures defined on the LTAS differ from $f_o$ measures in that we may not expect to find meaningful results from simple product moment correlations between acoustical measures as predictor and auditory ratings as criterion. Rather, we have to take recourse to multivariate techniques. We have chosen an approach based on multiple regression analysis. Since intuitively loudness, tension and laxness seem to be related, and because it appeared that the rating scores were clearly correlated, it is sensible to present the results in combination. It was found that tenseness and laxness showed the expected complementary picture. The order in which the first two slopes enter the equation was reversed, but the differences in variance explained are small enough to blame this to chance fluctuations. The results for tenseness and loudness are partly similar and for another part quite different. The similarity lies in the signs of the regression coefficients, the difference in the relative importance of the slope measures. For the perception of loudness the spectral slope in the frequencies below the first formant is most important, followed by the slope in the frequency region around 5000 Hz. For the perception of tenseness, on the other hand, the slope in the frequencies between the first formant and 1600 Hz is most important. In general, however, it may be concluded that the results seem to confirm the prediction that both loudness and tenseness are characterized by a higher spectral level (or in our measures: a less steep spectral slope) in the upper frequency bands.

Up to now we have considered a single acoustic measure for each setting. This is not contradicted by the multiple regression approach to spectral slope measures since there a number of (correlated) values together define a composite parameter. Such a single parameter approach is motivated by the claimed analytic character of the transcription. Correlations between scores on various setting scales,

however, are one reason for considering the possibility that a number of different acoustic features might interact to determine the auditory ratings. To investigate this we have computed multiple correlations of all rating scales with the complete set of eventually retained acoustic measures, i.e. $f_o$ median, $f_o$ variation coefficient, $f_o$ perturbation computed from $df_o$ based on raw pitch values and the spectral slope measures.

For the pitch level and the pitch range scales the results are extremely easy to describe, since no variables beyond $f_o$ median (for pitch level) and the $f_o$ variation coefficient (for pitch range) appear to make a significant contribution. The same goes for the laxness scale, where no significant contributions are found from other parameters than the slope measures. In the results for the loudness ratings the $f_o$ median takes over the position of spectral slope measure 2 as the third predictor, without increasing, however, the total percentage of variance explained.

The remaining scales, viz. tenseness and harshness, show more interesting results. For tenseness slope 2 is removed from the equation and replaced by $f_o$ median. Slopes 1 and 3 exchange their places and $f_o$ perturbation comes in as a fourth predictor bringing the multiple R to a value of .73. The regression coefficients of both $f_o$ median and perturbation are positive, indicating a tendency in the direction of higher tension scores if either $f_o$ median or $f_o$ perturbation increase. This is what would be expected on physiological grounds. A physiological explanation is corroborated by the regression coefficients of the slope measures which is negative for the high frequency slope and positive for the low frequency slope 1. Turning to the harshness scale we find that $f_o$ perturbation remains the most important predictor but it gets the company of the slopes 3 and 1 yielding a multiple R of .57 compared with a Pearson r of .42 for the perturbation as the sole predictor. Moreover, the signs of the regression coefficients of the slope measures are equal to that of laryngeal tension. This finding confirms the conjecture that harshness is likely to be one of the results of excessive tension of the muscles of the larynx. A more detailed discussion of these results may be found in Ref. 3.

REFERENCES

1. B.J. WINER 1971 Tokyo: McGraw-Hill, 286.
   Statistical principles in experimental design.
2. B. HAMMARBERG, B. FRITZELL, J. GAUFFIN and L. WEDIN 1980 Acta Otolaryngologica 90, 441-451.
   Perceptual and acoustic correlates of abnormal voice qualities.
3. B. CRANEN, L. BOVES and R. VAN BEZOOIJEN 1982 Quantitative Linguistics.
   Acoustic correlates of auditory ratings of some voice quality settings.
   (To appear).