Proceedings of The Institute of Acoustics

RELATIONS BETWEEN PERCEPTUAL RATINGS OF VOICE QUALITY AND ACOUSTIC MEASURES

Louis Boves and Renée van Bezooijen

Institute of Phonetics, Nijmegen University, The Netherlands

INTRODUCTION

In many situations, both in real life and in research, a precise characterization of the non-verbal properties of someone's speech would be very helpful. Real life situations include, of course, clinical work (both diagnostic and therapeutic) but also speaker verification and testing of the quality of communication channels (including, perhaps, in the near future testing of the quality of the output of text-to-speech systems). Research situations in which descriptions of non-verbal properties of speech are needed occur in phonetics, but also in socio-linguistics and social psychology.

Unlike the extensive tradition of the phonetic description of segmental units in speech, only few attempts to devise a consistent and comprehensive system for the description of non-verbal aspects seem to have been undertaken. One of the few exceptions is the work done in Edinburgh by Laver and his associates, who have developed a description system that is claimed to be based on a solid theory of articulatory phonetics and acoustic speech production [3]. The system consists of scales related to a large number of so-called articulatory settings. The scoring is done auditorily.

Our experience in using Laver's system has revealed a number of problems. Some minor problems, like the lack of a sufficient diversity of scales referring to prosody, can easily be remedied. Two major problems remain to be solved, however, viz. the reliablility of the ratings and the communicability of the results. Obviously, these problems are closely related.

In our present contribution we have approached the problems of the reliability and communicability of the ratings in two ways. Firstly, the reliability problem has been tackled by means of straightforward statistical analysis of the ratings of three transcribers for a fairly large number of adult male speakers. For those non-verbal features which proved to yield reliable ratings it was attempted to find acoustic measures that can explain the ratings. A successful acoustic explanation of a set of auditory ratings would make the communication of the results quite easy, because the acoustic measurements can be described objectively.

AUDITORY DESCRIPTION

Speech material

The speech material which was at the basis of the present study consisted of spontaneous speech produced by 32 males from Nijmegen, a town (150,000 inhabitants) situated in the east of the Netherlands [2]. There were 16 younger speakers, aged between 15 and 18 years, and 16 older speakers, aged between 60 and 74 years. Either age group comprised speakers of varying socioeconomic status. For each speaker, verbally neutral utterances were selected with a total duration of about one minute. These utterances were spliced onto a tape, separated by pauses of 0.5 second.

RELATIONS BETWEEN PERCEPTUAL RATINGS OF VOICE QUALITY AND ACOUSTIC MEASURES

Perceptual parameters

The 32 speech samples, placed in a random order, were rated by three raters, one of whom, the second author, has extensive experience in the auditory description of the vocal characteristics of speech; the other two, graduate students of linguistics, had been trained especially for the purpose of this study. Ratings were given, independently by the three raters, on the 27 scales presented in Table 1. Nineteen of these have been taken from the vocal profile analysis protocol given in [3]; they were chosen because we thought they could be rated reliably and because they pertain to normal voices (Laver et al's protocol also contains some scales specifically aimed at the description of pathological voices). The other eight scales, namely emphasis, varied pitch patterns, deviating pitch patterns, pitch variability, precision of articulation, regional accent, and affectedness, were added because a global listening to the recordings suggested that these scales might differentiate between (groups of) speakers.

Table 1. Reliability of the ratings given by 3 raters for 32 speech samples

7-point scales		4-point scales	
Pitch level	.73	Deviating pitch patterns	.75
Pitch range	.75	Tremulousness	-.03
Pitch variability	.86	Harshness	.83
Emphasis	.82	Creak	.92
Varied pitch patterns	.89	Whisper	.72
Loudness	.92	Nasality	.02
Loudness range	.88	Denasality	.81
Loudness variability	.84	Pharyngeal constriction	.70
Tempo	.87	Breath support	.92
Tempo variability	.72	Lip rounding	.71
Connectedness	.66	Lip spreading	.38
Sonority	.85	Regional accent	.93
Laryngeal tension	.84	Affectedness	.90
Precision of articulation	.83		

Among the 27 scales, two types may be distinguished. The one type consists of scales which form a continuum from the para- and extralinguistic absence of the feature in question to a high degree of presence, and contains four scalar degrees. Thus, speech may contain no creak at all (scale position 0), a little bit (1), a fair amount (2), or a great amount of creak (3). The specification "para- and extralinguistic" serves to indicate that in the rating the presence of a feature for linguistic purposes, such as lip rounding with vowels for which lip rounding is a distinctive feature, has not been taken into consideration.

The other type of scales contains seven scalar degrees and pertains to features which are an intrinsic part of the speech signal; these features can be present in different degrees, going from one extreme to another. An example is the

pitch level scale, in which 1 = very low, 2 = fairly low, 3 = somewhat low, 4 = neutral, 5 = somewhat high, 6 = fairly high, and 7 = very high. The central scale position "neutral" is the reference to which the other scale positions are to be related; it is defined as the "average" manifestation of the feature in question in the spontaneous, non-emotional speech of Dutch speakers of the standard variety. Although it is difficult to give a formal description of this reference, in practice the raters generally experienced few problems in giving it a workable interpretation.

Reliability

The reliability of the ratings was assessed for each scale separately. Use was made of the so-called Ru-coefficient [4]. This coefficient is of the form 1 - MSwithin/MSbetween, and is a measure of the reliability of the means of the ratings of the group of raters. It may be seen from Table 2 that for the majority of the scales the coefficients are satisfactorily high. Especially loudness, creak, breath support, regional accent, and affectedness have been rated very reliably. There are only three parameters which have been rated very unreliably, namely tremulousness, nasality, and lip spreading.
Low coefficients may be the result of a low agreement among the raters (a high MSwithin), a lack of variation in the stimulus characteristics (a low MSbetween), or a combination of these two factors. Inspection of the MSwithin and the MSbetween values revealed that the low coefficients for tremulousness, nasality, and lip spreading were due to little variation in the stimuli, or, more precisely, to a lack of occurrence of these parameters in the material (practically only 0- and 1-scores have been given). Almost complete absence of a parameter may, however, also give rise to moderately high reliability coefficients. This occurs if one or two speakers exhibit the parameter, which is absent from the speech of all other speakers in the set, in a (very) high degree. In our material this proved to be the case with pharyngeal constriction. For this reason this parameter will not be dealt with below.

ACOUSTIC ANALYSES

Only the minority of the vocal features which have been rated auditorily have more or less obvious acoustic correlates. The search for acoustic measures which might predict auditory ratings has been guided by the claim that the ratings do indeed describe settings, i.e., that they describe characteristics that are continuously present in the speech signals. This consideration has motivated the choice of acoustic measures in the form of central tendencies and variances of parameters.
The 32 speech samples were digitized (sampling frequency 10 kHz) and subjected to an autocorrelation LPC analysis, using 12 predictor parameters, a frame length of 25 ms, Hamming windowed at a frame rate of 10 ms [1]. For each frame the frequency and bandwidth of the complex pole pairs were determined by solving the predictor polynomial for its zeros. From these data the frequencies and bandwidths of the first four "formants" were determined and the "formant tracks" were smoothed by means of a simple moving median filter.
Average values of the frequency and bandwidth of the first four formants were next determined for the voiced frames only; variances were also computed. Average formant values were expected to reflect articulatory settings, whereas

the variances might give an impression of the extent of the formant space. In order to get a measure of the speed with which the articulators move, the first and second difference of the formant tracks were computed, regardless of the voiced/unvoiced decision. Average absolute velocity and acceleration were calculated, indicated as v(Fn) and a(Fn), respectively.
In order to keep the number of acoustic variables within reasonable limits, the average formant frequencies were converted into a measure of the Euclidean distance of the center of gravity of the speaker's formant space to the "ideal" neutral formant configuration Fn = (2n-1).500 Hz. This was done for a two-formant plane (ASC-2) and a four-formant space (ASC-4). Data on formant bandwidths were not used at all, mainly because LPC bandwidth measures are not easy to interpret to begin with. The variances of the formant frequencies were also converted into a single measure by multiplying them; this measure is labelled "formant space".
Average F0 and the variation coefficient of F0 (i.e. standard deviation divided by average value) were computed, as well as the variance of the signal energy. Analog equipment was used, that is interfaced to a digital computer. Integration time of the sound level meter was 10 ms and the RMS output signal was sampled after log-conversion. The pitch detector works in the time domain, i.e., it computes the duration of each successive pitch period. In order to compute average F0 the sequence of period durations was interpolated at 10ms intervals and then passed through a smoother of the moving median type. The raw output signal of the pitch extractor was differenced and the standard deviation of the difference signal was taken as a measure of F0 perturbation [5].
The average signal power was, of course, determined as part of the computation of the variance but it was not retained as a useful measure since all recordings had been brought to the same overall level when the stimulus tapes were compiled.
Long-term average critical band spectra of the voiced parts were obtained by means of the procedure detailed in [6,7]. These spectra were converted into a more compact description that essentially consists of the spectral slope in the region below the first formant (slope 1), the slope of the spectrum in the region between the first formant and 1.6 kHz (slope 2), the slope in the upper region of the spectrum (slope 3), and the normalized spectral energy in the region of the first formant (max 400-600) [6,7,8].

The set of measures described above is obviously incomplete. First of all, duration measurements are missing, which might explain ratings of tempo and tempo variability. Neither are there descriptions of F0 patterns that might correspond with ratings on the scales of varied pitch patterns, deviating pitch patterns, breath support, and connectedness, and most likely also emphasis. As for the time measurements, because we are dealing with spontaneous speech they will have to be made by hand. This very time consuming task is under way but not yet finished, so that we cannot report any results. The problem with the description of F0 patterns (or perhaps more accurately: prosodic patterns) is more fundamental. We don't avail of any techniques that allow us to identify those patterns via automatic processing of speech signals nor do we know how to describe their variability in a formal way. Therefore we will limit the discussion to the settings not named in this paragraph.

RELATIONS BETWEEN PERCEPTUAL RATINGS OF VOICE QUALITY AND ACOUSTIC MEASURES

RELATIONS BETWEEN ACOUSTIC MEASURES AND AUDITORY RATINGS

In order to find out to what extent the auditory ratings can be predicted by acoustic measures, a multiple correlation analysis was performed in which the auditory ratings served in turn as the criterion measure. For each scale a subset of the acoustic measures was taken as possible predictor variables. The choice of measures was, of course, based on literature data and experience with respect to acoustic correlates of the auditory parameters. The program used to carry out the analyses performs a stepwise analysis, i.e. at each successive step a new predictor varaiable is entered into the regression equation in order to try to account for the variance in the criterion variable left over by previously entered predictors. Our program allows any combination of free and forced predictors. Forced predictors are entered into the equation at the will of the user; free variables are entered automatically, and their choice is based on the relative amount of variance they account for.
The program continues inserting predictor variables until there are no more left or until some formal criterion is (no longer) satisfied. In the analysis runs on which this paper is based insertion was stopped if the next variable to be inserted explained less than 5% of the remaining variance.

Results

The main results of our analyses are summarized in Table 2. It appears that only for three scales (pitch range, sonority, and creak) more than 40% of the variance is explained by a single acoustic measure. The remaining scales that are related to phonation reach at least a multiple R of .62, corresponding with some 40% explained variance. It should be noted, however, that in the case of loudness range and loudness variability the correlations between acoustic predictors and perceptual criterion measures are most likely due to chance.
One interesting observation that can be made is that there seems to exist a group of auditory rating scales that all have some relation to mean F0. In a factor analysis of the auditory ratings the same cluster appeared, as should have been anticipated given the high mutual correlations of the ratings on the features pitch level, sonority, laryngeal tension and creak. Contrary to what one would expect pitch level is not the scale that is most closely related to mean F0. In fact, it occupies only the fourth rank, preceded by the three other scales mentioned above. The moderately high correlation between average F0 and pitch level scores is a consistent finding throughout our research in this field. Harshness does not fit into the set of scales which are strongly related to pitch level in the ratings. Nevertheless, in previous research into the perceptual and acoustic properties of emotional speech a similar relation between harshness scores and mean F0 was found [7].
Turning to the articulatory scales now, we observe that only in the case of affectedness the multiple R excedes .60. Apparently the acoustic measures used in this study fail to account for the perceived differences in the articulatory characteristics of the speakers.

Discussion

In general it must be conceded that the auditory ratings obtained in the course of this study defeat an explanation on the basis of fairly simple acoustic measures that represent long term averages and variances of a number of popular

RELATIONS BETWEEN PERCEPTUAL RATINGS OF VOICE QUALITY
AND ACOUSTIC MEASURES

Table 2. Results of a multiple regression analysis with acoustic measures as predictors and auditory ratings as criterion. Threshold of insertion: 5%. The signs are borrowed from the regression coefficients.

Criterion	Predictor(s)	Multiple R (r)	Cumulative prop. var. explained
Pitch level	F0 mean	.72	.294
	Max 400-600		.395
	Slope 2		.517 (-)
Pitch range	F0 variation coef.	.73	.408
	F0 perturbation		.527 (-)
Pitch variability	F0 variation coef.	.66	.262
	F0 perturbation		.434 (-)
Sonority	F0 mean	.84	.612 (-)
	Intensity variation		.706
Creak	F0 mean	.76	.572 (-)
Whisper	Slope 1	.70	.273 (-)
	F0 perturbation		.329
	Slope 2		.440
	F0 mean		.491
Harshness	F0 mean	.68	.321
	max 400-600		.465 (-)
Laryngeal tension	F0 mean	.77	.380
	max 400-600		.531 (-)
	F0 variation coef.		.595
Loudness	Slope 1	.66	.320
	F0 mean		.436
Loudness range	ASC-2	.62	.218 (-)
	F0 variation coef.		.337
	Max 400-600		.390
Loudness variability	F0 variation coef.	.80	.287
	max 400-600		.460
	a(F1)		.517 (-)
	% samples voiced		.571
	v(F1)		.644
Precision of art.	v(F2)	.40	.099 (-)
	Slope 2		.160 (-)
Denasality	% samples voiced	.55	.157
	Formant space		.298
Lip rounding	ASC-4	.57	.103 (-)
	Slope 1		.182 (-)
	Slope 2		.325
Regional accent	ASC-4	.54	.137
	Intensity variation		.291
Affectedness	ASC-4	.61	.200 (-)
	% samples voiced		.305 (-)
	F0 perturbation		.377

acoustic phonetic parameters. This type of acoustic measures was deliberately chosen because the auditory ratings are assumed to reflect properties of the speech signals that are more or less continuously present. This makes the failure an interesting one.

The failure cannot be explained away by contending that the ratings are not meaningful. For this the reliabilities are much too high. It is our opinion that the basic assumptions underlying the ratings should be questioned, especially the claim that the ratings pertain to properties of the speech signals that are continuously present. We have a strong feeling that in normal speech many of the features operate intermittently. This applies to both the phonatory and the articulatory features. In such a case ratings depend on two, possibly independent, aspects of the intermittent feature, viz. the frequency with which it is present and its intensity when it occurs. The speech of normal subjects seems to occupy a fairly narrow range of the total scale spanned by a feature if pathological speech is included. Nothing is known yet about the trading relation between frequency of occurence and intensity in the ratings.

If it is true that normal speech occupies a restricted range of the scales because of the intermittent presence of the features, there are two consequences that must be envisaged. The first, and perhaps least interesting, is that high correlations of the scores on a scale with any other measure may only be expected if the measures are very strongly associated, since limiting the range of one or more variables to a small part around the center of a scale tends to diminish the correlation coefficient, even if extremely accurate measurements are available. In this connection it is interesting to remember the high correlations between the ratings on the pitch level, sonority, laryngeal tension and creak scales. Apparently these adjectives refer to some auditory phenomenon that has many facets and therefore is given many different names, but none of the facets is directly accessible to perception, i.e. not without some interference of other facets. This finding is especially surprising for an intuitively simple feature like pitch level. Most likely the explanation of this finding can be found in the physiology of the normal but untrained larynx, but we will not pursue this point here. Note, however, that it does affect the claim that the features in the transcription system represent settings that can be controlled independently.

The second consequence of the intermittent presence of an auditory feature and the assumption that the ratings are influenced by both the frequency of occurrence and the intensity is that considerable doubt is cast upon the appropriateness of acoustic measures based on long term averages and variances. Not all phonetic segments and not all parts of a sentence are equally susceptible to the operation of a specific setting. For instance, a fair degree of lip spreading may affect rounded vowels to a much larger degree than their unrounded counterparts; creak seems to be more likely to occur towards the end of a sentence than towards the beginning, etc. These considerations suggest that the highly global acoustic measures used in this study should be replaced with measures that are much more intelligent. Signal processing strategies have to be developed which first search the signal for the presence of some specific "setting". Especially in the case of articulatory settings such a search presupposes the availability of a phonetic transcription of the utterances, lined up with the speech signal. Without this additional information it is very difficult to imagine a procedure that is able to

interpret e.g. formant values of a vowel in terms of settings. Similar intelligent procedures ought to be constructed for the description and analysis of prosodic patterns.
Intelligent processing strategies that have access to both the signal and its phonetic transcription would enable us to establish both the frequency and the intensity with which a given feature is present. This knowledge, in its turn, would be extremely useful in research into the way in which our perception of the quality of someone's voice operates.

ACKNOWLEDGEMENT

This research was supported by the Foundation for Linguistic Research which is funded by the Netherlands Organization for the Advancement of Pure Research, ZWO.

REFERENCES

[1] J.D. Markel & A.H. Gray Jr, "Linear prediction of speech". Berlin: Springer Verlag, 1976.
[2] R. van Hout, "Sampling problems in sociolinguistic research". in Mededelingen van de NCDN, Vol. 16, 47-93, (1978) (in Dutch).
[3] J. Laver, S. Wirz, J. Mackenzie & S. Hiller, "A perceptual protocol for the analysis of vocal profiles", Work in Progress, Dept. of Linguistics, Univ. of Edinburgh, Vol. 14, 139-155, (1981).
[4] J. Asendorpf & H.G. Wallbott, "Masse der Beobachterubereinstimmung: Ein systematischer Vergleich", Zeitschrift fur Sozialpsychologie, Vol. 10, 243-252, (1979).
[5] A. Askenfelt & B. Hammarberg, "Speech waveform perturbation analysis revisited", STL-QPSR 4/1981, 49-68.
[6] L. Boves, "The phonetic basis of perceptual ratings of running speech". Dordrecht: Foris Publ., 1984.
[7] R. van Bezooijen, "Characteristics and recognizability of vocal expressions of emotion". Dordrecht: Foris Publ., 1984.
[8] K. Elenius, "Long time average spectrum using a 1/3 octave filterbank", STL-QPSR 4/1980, 14-22.