THE LONG-TERM MEASUREMENT OF VOICE QUALITY: A COMPARISON OF ACOUSTIC MEASURES

J. Pittam

Department of English
University of Queensland

ABSTRACT - A long-term acoustic measure of voice quality is developed. Five measures, based on the long-term average spectrum, are compared for their ability to discriminate breathy voice, creaky voice, nasal voice, tense voice and whispery voice. A discriminant analysis technique is used. These measures utilise the Mel and Bark scales as well as 'equal-Hertz' intervals.

INTRODUCTION

Voice quality is essentially a long-term vocal phenomenon. Following the work of John Laver, this study bases the concept on the notion of an articulatory setting (see Laver, 1980) which refers to the tendency of the vocal apparatus to take up specific long-term muscular adjustments that underlie the movement involved in producing a sequence of segments. As such, voice quality may be seen as a marker of individuality. Some voice qualities, however, may be consciously controlled, and have been shown to characterise such social variables as sex, age, regional groups and social class, as well as being carriers of emotion and personality characteristics (see Laver (1979) for an extensive bibliography of voice quality). If an acoustic measure is to be useful in assessing an articulatory setting as a social or personality marker within interactions, that measure should reflect the long-term nature of voice quality. This is particularly so if the physical measure is to be related to perceptual measures.

The present study aimed to provide a long-term acoustic measure that successfully discriminated a number of articulatory settings. Those selected were breathy voice, creaky voice, nasal voice, tense voice and whispery voice. All have been cited in the literature as functioning communicatively (see Laver, 1979). The acoustic measure adopted is the long-term average spectrum (LTAS). This is the averaged intensity or amplitude spectrum across a selected frequency range for continuous speech. It has been suggested that after about 30 seconds of continuous speech, the LTAS will not change significantly regardless of how much more speech is analysed (Li, Hughes & House, 1969). If this is so, it is a potentially useful measure for long-term vocal phenomena.

The data produced by a long-term spectral analysis, however, need to be reduced. Individual variation in the LTAS will confound any attempt to discriminate voice qualities overlaid on the normal voice. In addition, each quality needs to be represented by a small enough set of values for any statistical analysis used to discriminate the voice types. But should the perceptual scales of Bark or Mel be used to reduce the data? If the measure is to be seen as a useful tool for those working in the communications or voice perception areas, it may be that the measure should reflect our perceptual capabilities. On the other hand, voice qualities may affect the LTAS at a sufficiently gross level that they can be discriminated without using such perceptual scales. It is these questions that this study addressed, using the five voice qualities indicated above.

First, an earlier study by the author is referred to, in which the effectiveness of a series of 'equal-Hertz' intervals was considered. This study provided the basis for the present study in that the most useful frequency interval for reducing the data was determined. This frequency interval was then compared to a series of Bark and Mel intervals to establish their ability to discriminate the five voice types.

METHOD

Speakers

Six female and six male Australian-born speakers recorded a 30 - 40 second reading passage. As nasality was one of the qualities measured, the passage was designed to include no nasal consonants. The age range of the speakers was 17 - 67 years. They were asked to read the passage five times producing each of the voice qualities in turn. The differing abilities of the 12 speakers resulted in a series of recordings that ranged from extreme examples of a particular setting to those of minimal auditory presence.

Spectral analysis

All voices recorded were analysed by a Hewlett Packard HP3582A digital spectrum analyser. LTAS were produced covering the frequency range 0 - 10 KHz. A Hanning window was applied, followed by a fast Fourier transform, and a root mean square averaging routine was used. The result was an amplitude rather than intensity spectrum. The Hewlett Packard filters the signal at 25 KHz. and then feeds it into an A/D converter which, using the 10 KHz. frequency range, produces a sequence of samples at a rate of 81.92 KHz. To properly anti-alias the signal, a low pass digital filtering system is then used, retaining only 40 KHz.

Perceptual ratings

To provide independent evidence of the degree to which the settings were present in the voice two trained judges independently rated the 60 recordings. Using a six point scale, they rated each recording for degree of auditory presence on each of the five settings. The reliability coefficient across all ratings for the two judges was .81. In addition to the primary setting (used here to mean the setting that the speakers were trying to produce) most voices were shown to include at least one of the other settings, albeit to a much lesser degree. Not all the five settings considered here, however, can operate at the same time. For instance, breathiness cannot accompany whispery or creaky settings because of the difference in degree of medial compression between breathiness and the other two (Laver, 1980). The presence of the secondary settings, therefore, may simply indicate that a speaker lost control of the primary setting for a moment and produced short examples of a different setting. On the other hand, some settings clearly did operate simultaneously. Eight of the whispery voices, for example, showed some degree of tenseness. Laver (1980) indicates that whisper may accompany certain types of tense voice. The examples here would seem to support this.

Reduction of the data

An earlier study (Pittam, 1985) had indicated that the most useful frequency range for discriminating these five settings was 0 - 2 KHz., and the most effective reduction achieved by taking the mean amplitude of each successive 200 Hz. interval. This was determined by checking the intervals 125, 150, 175, 200, 225 and 500 Hz. in both the 0 - 2 KHz. and 0 - 3 KHz. ranges. This part of the vocal spectrum carries much linguistic and non-linguistic information. It was likely, therefore, to be useful for the discrimination of voice quality.

In addition, the values arrived at by this reduction procedure were then normalised. When the voices were recorded, no stringent controls were placed on constancy of loudness or distance from microphone, although neither varied markedly across the 12 speakers. It is possible that different settings will be associated with differing degrees of loudness. Whispery voice, for example, would normally be accompanied by lowered amplitude. There is also a strong probability that tense voice will be accompanied by a louder loudness-range (Laver, 1980). However, to accommodate any differences in the recordings, the spectral values that had been calculated were normalised by taking the difference in adjacent values. While information was certainly lost by this procedure, it did have the effect of retaining the basic shape of the spectrum by indicating the direction in which the spectrum was moving at that point, and the magnitude of the movement. Thus, if the mean amplitude in the range 0 - 200 Hz. was -44dB, and the adjacent mean in the range 200 - 400 Hz. was -38dB, the difference was 6dB. This shows that the spectrum was gaining amplitude (by the positive polarity) and by how much. It was these normalised values, representing gross spectral movement only, that were used in all analyses.

In the earlier study, each set of normalised values was analysed using a multidimensional scaling (MDS) analysis on a subset of the voices - four from each voice type. These represented the 'purest' voices in that the perceptual rating for the primary setting was high, while secondary settings were either very low or absent. It was argued that these should provide the clearest indication of which interval and which frequency range produced the best separation of the voice types. The 200 Hz. interval, using the 0 - 2 KHz. range, accounted for 97.5% of the variance.

While it is not intended to detail out the earlier analysis here, one important point should be mentioned. The MDS analysis separated the 20 voices, not in terms of underlying dimensions, but in terms of the primary setting and the presence of secondary settings. It was, in other words, an appropriate case for 'neighbourhood' analysis (Kruskal & Wish, 1978) - an analysis which may be preferable to a dimensional approach in that regions of space may have more meaning than underlying dimensions. So while the five voice types did tend to cluster separately, indicating the influence of the primary settings, secondary settings also affected the positioning. To give just one example. Two of the creaky voices contained the secondary setting of tenseness. Both lay to one side of the creaky voice cluster close to the tense area.

In the study reported here, all 60 voices were checked using the 200 Hz. interval across the 0 - 2 KHz. range, and this was compared with Bark and Mel measures.

THE ACOUSTIC MEASURES COMPARED

Adopting the same reduction and normalisation procedures as in the earlier study, sets of acoustic values were calculated for 200 Hz., 1 Bark, 1.5

Bark, 150 Mel and 200 Mel intervals. Barks were calculated using the approximation given by Syrdal (1985); Mels were calculated using the Fant (1973) approximation. The specific Bark and Mel intervals selected were chosen because they included intervals of approximately 200 Hz. at some point across the 0 - 2 KHz. range. The numbers of values representing each interval after reduction and normalisation were as follows: 200 Hz. - 9; 1 Bark - 12; 1.5 Bark - 8; 150 Mel - 8 and 200 Mel - 7. The largest number of values was 12. It was felt that with 60 voices, this would still be acceptable for the statistical analysis. A series of five discriminant analyses was then conducted using the normalised values for all 60 voices as the dependent variables grouped into five voice types. Significant multivariate effects were found for all five analyses, and posthoc Newman-Keuls tests showed that the following voice types were discriminated at p<.05:

200 Hz. - All voice types were discriminated from all others except creaky voice from nasal and tense voices.

1 Bark - Breathy voice was discriminated from creaky and tense voices; whispery voice from creaky, nasal and tense voices, and creaky voice from tense voice.

1.5 Bark - All voice types were discriminated from all others except creaky voice from tense voice.

150 Mel - All voice types were discriminated from all others except nasal voice from breathy and creaky voices.

200 Mel - As with 1.5 Bark, all voices were discriminated from all others except creaky voice from tense voice.

Overall, therefore, of the 10 possible pairwise discriminations, the five analyses accounted for the following: 200 Mel - 9; 1.5 Bark - 9; 200 Hertz - 8; 150 Mel - 8, and 1 Bark - 6. The 200 Mel and 1.5 Bark intervals, therefore, provided the best discrimination based on group means.

Two significant functions were derived for both of these intervals. For the 1.5 Bark interval (F(32,179) = 4.893; p<.001; Wilks' Lambda = .098) the two functions accounted for 90.67% of the variance; for the 200 Mel interval (F(28,178) = 5.184; p<.001; Wilks' Lambda = .117) the two functions accounted for 94.28% of the variance. The centroids indicated that, in each case, function one may be interpreted as a glottal friction/non-friction parameter, separating whispery and breathy voices, both of which are characterised by some degree of glottal friction, from the other three types. Whispery voice, which contains the most friction, was positioned furthest from the other three voice types on this function.

Function two appears to be similar to the 'breathy-overtight' parameter described by Gauffin and Sundberg (1977). They placed creaky voice at the overtight end of their continuum beyond tense voice, with breathy voice at the opposite (lax) end. In the study presented here, creaky voice, whispery voice and tense voice were opposed to breathy voice on this function. Of the 12 whispery and 12 creaky voices recorded for this study, eight of each were characterised by some degree of tenseness.

Scatter plots of the discriminant scores were also checked. As might be expected following the earlier MDS analysis, the position of each voice reflected the presence of both primary and secondary settings. While the voice types clustered, the presence of secondary settings and the degree of auditory presence of the primary setting meant that some overlap occurred. This may suggest that, once again, a neighbourhood type of analysis is the more appropriate approach.

SPECTRAL VALUES RELATED TO VOICE TYPE

The univariate analyses indicated that for the 200 Mel analysis, function one was associated with values two, four and seven, while function two was associated with values one and three. This was very similar to the 1.5 Bark analysis in which values two, four, five and eight were associated with function one, and values one and three, with function two. The Newman-Keuls tests showed that those values associated with function one for both analyses separated whispery and breathy voices from the other three voice types, and the values associated with function two separated breathy and nasal voices from the others. This is as one would expect following the interpretation of the functions given above.

Looking first at the values associated with function two - the function possibly associated with degree of tenseness in the voice. Value one for both the Mel and Bark analyses showed that nasal and breathy voices were characterised by a larger rise in energy in this low part of the range than creaky and whispery voices. This is not unexpected. Value one is the difference between the first two un-normalised mean amplitude values. Creaky voice, with its characteristic low fundamental frequency has more energy in the range covered by the first of these means. The frequency ranges covered are 0 - 149 Hz. for the Mel analysis, and 0 - 161 Hz. for the Bark analysis. Fundamental frequency for the female speakers should drop into this range for creaky voice, ensuring less spectral movement between this first mean value and the second over the 12 speakers. Whispery voice, with its generally flattened spectrum would be expected to produce a similar result. Normalised value three, which was also associated with function two, showed that breathy and nasal voices had the largest drop in amplitude across this range (320 - 741 Hz. for the Mel analysis, and 297 - 631 Hz. for the Bark analysis). This part of the frequency range covers the drop from the first major spectral peak. It is in this area that we would expect the damping associated with these two voice types to affect spectral movement most.

The values associated with function one were two and four for both analyses plus five and eight for the 1.5 Bark interval and seven for the 200 Mel interval. Leaving aside value two for the moment, the others showed that whispery and breathy voices were characterised by less spectral movement than the other three voice types through these frequency ranges. In particular, whispery voice was characterised by a spectrum that was close to being flat. This effect is undoubtedly due to the presence of glottal friction, with the higher levels of friction associated with whispery voice producing the flatter spectrum. Value two meanwhile showed that whispery and breathy voices were characterised by larger drops in energy than the other voice types. For breathy voice this is associated with a damped vocal tract - an effect also noted above for the adjacent value three. It is not so clear why whispery voice should also show a large drop in energy at this point.

CONCLUSIONS

The 1.5 Bark and 200 Mel intervals proved the most successful in discriminating the five voice types, although the equal-Hertz interval proved almost as successful. Generally speaking, the measures developed have proved successful, although more work is needed to extend the range of voice qualities measured to include others that are used communicatively. The ways in which the two most successful measures discriminated the five qualities can be validated phonetically.

The one pairwise discrimination that was not significant was creaky voice from tense voice. Given the interpretation placed on the discriminant functions this is not too surprising. The combination of presence/absence of glottal friction and degree of tenseness will not allow these two voice types to be clearly discriminated from one another, particularly when a number of the creaky voices recorded displayed some degree of tenseness. This, and the overlap noted in the scatter plots, may suggest the need for a neighbourhood approach rather than a dimensional approach. As more voice qualities are added, it is likely that a three-dimensional solution also will be required.

From both the earlier MDS study and the discriminant analyses presented here, it appears that some secondary settings, in addition to the primary settings, also need to be taken into consideration in any attempt to discriminate voice quality acoustically. Using the term 'voice quality' as Laver (1980) does, other individual settings would always be present, of course - even if they are neutral settings. The measures developed here seem capable of overcoming some effects of this individual variation - at least as long as those settings that are present more or less all the time are of minimal auditory presence.

That spectral movement of the type measured here successfully discriminated the five voice qualities, indicates that they are reflected in the LTAS at a rather gross level. As such, when considering the measurement of voice quality as part of the vocal communicative package, it may be useful to think in terms of what might be called a 'macro-acoustic' level, in addition to the 'micro-acoustic' level.

REFERENCES

FANT, G. (1973) "Speech Sounds and Features", (MIT Press: Mass).

GAUFFIN, J. & SUNDBERG, J. (1977) "Clinical Applications of Acoustic Voice Analysis", QPSR - STL, 2-3, 39-43.

KRUSKAL, J.B. & WISH, M. (1978) "Multidimensional Scaling", (Sage: Beverley Hills).

LAVER, J. (1979) "Voice Quality: a classified bibliography", (John Benjamins: Amsterdam).

LAVER, J. (1980) "The Phonetic Description of Voice Quality", (Cambridge University Press: Cambridge)

LI, K.-P., HUGHES, G.W. & HOUSE, A.S. (1969) "Correlation Characteristics and Dimensionality of Speech Spectra", J. Acoust. Soc. Am. 46, 1019-1025.

PITTAM, J. (1985) "Voice Quality: its measurement and functional classification", Unpublished PhD thesis, University of Queensland.

SYRDAL, A.K. (1985) "Aspects of a Model of the Auditory Representation of American English Vowels", Speech Communication 4, 121-135.