THE LONG-TERM ACOUSTIC CHARACTERISTICS OF EMOTION

J. Pittam*, C. Gallois** and V.J. Callan**

* Brisbane College of Advanced Education
** University of Queensland

ABSTRACT - Long-term spectra of recordings of three standard passages differing on perceived dominance and arousal were examined for 30 Australian speakers using three-mode principal components analysis. Results indicated that both affective dimensions were reflected systematically in the spectra, with dominance especially prominent in the upper part of the spectrum, and arousal affecting particularly two bands below 3 kHz.

INTRODUCTION

The long-term acoustic characteristics of emotion remain largely uncharted territory, with only a handful of researchers having worked in the area over the last 25 years. A few studies have utilised the long-term spectrum (LTS) of voice to measure simulated emotions (Williams & Stevens, 1972), or the generalised affect associated with stress (Friedhoff, Alpert, & Kurtzberg, 1962; Popov, Simonov, Frolov, & Khachatur'yants, 1971), depression (Darby & Hollien, 1977; Hargreaves & Starkweather, 1964; Hargreaves, Starkweather & Blacker, 1965), or other neuroses and psychoses (Ostwald, 1963).

One study measured what may have been the spontaneous expression of anger and fear (Roessler & Lester, 1976), but this was for one subject only. In general, all studies point to the importance of the first 1 kHz, although some (Alpert, Kurtzberg, Pilot, & Friedhoff, 1963; Ostwald, 1963; Popov et al., 1971) narrow this down further and suggest that the range up to approximately 400 - 600 Hz is most sensitive to emotional stimuli. In this lower frequency band, fundamental frequency (Fo) and related measures (Fo range, Fo bandwidth) appear to be particularly important (Roessler & Lester, 1976; Williams & Stevens, 1972). More recent work by Scherer (e.g., Scherer, 1986) has confirmed this.

Only two studies suggest the usefulness of higher frequencies (Hargreaves & Starkweather, 1964; Williams & Stevens, 1972). Both present spectra up to 4 kHz. Hargreaves and Starkweather report patients whose LTS show considerably higher energy levels up to 4 kHz after a recovery period from severe depression. Williams and Stevens indicate that the energy level above 1 kHz, relative to that below 1 kHz, was greatest for the simulated emotion of anger and least for sorrow. Both studies, therefore, seem to imply that a rise in arousal levels, from sorrow to anger, or from severe depression to recovery, result in a rise in spectral energy in the frequency range up to at least 4 kHz.

One can see from this that most work has concentrated on measuring a single affective dimension or state such as depression or stress, which is negative in tone. Even where multiple types of affect have been examined, there is a tendency to relate other discrete emotions to depression. As yet, no-one has examined normally expressed affect covering both positive and negative types of arousal in normal voices. Arousal is inevitably present in the voice along with other affective dimensions such as dominance and pleasure. Evidence suggests that these three dimensions, which underlie the expression and perception of many emotions (e.g., see Mehrabian & Russell, 1974), not only usefully describe emotional response domains, but are part of our cognitive set for interpreting emotional stimuli.

The present study examined whether the LTS can be used to differentiate ordinary speaking voices that are characterised by differing levels of arousal and dominance. Although most previous work has been concentrated within the speech band, two studies, as we have seen, have indicated that higher frequencies may be useful. In examining this question, therefore, we decided to look at a frequency range that extended well beyond the speech band.

METHOD

Speakers, recordings and affective ratings

Fifteen male and 15 female Australian speakers, from three ethnic backgrounds (Australian, British

and Italian), each recorded three standard passages in English (90 recordings in all). The content of the passages covered three distinct situations, 1) a job interview (job), 2) an interview with a school headmaster (school), 3) a report of a tennis game (tennis) (see Gallois & Callan, 1988, for details of speaker selection and recording). All recordings were rated on a series of adjective scales by 120 Australian-born subjects. Subsequent factor analysis and analysis of variance revealed that the three passages were perceived as differing on the dimensions of arousal and dominance as follows:

Dominance: School significantly lower than the job and tennis passages.
Arousal: Tennis significantly higher than the job and school passages.

Spectral measurement

All recordings were analysed by a Hewlett Packard HP 3582A digital spectrum analyser. LTS were produced for the range 0 - 10 kHz. A Hanning window was applied, followed by a fast Fourier transform and a root mean square averaging routine. The HP3582A filters the signal at 25 kHz, feeds it into an A/D converter and produces a sequence of samples at a rate of 81.92 kHz (using a 10 kHz frequency range). The signal is then lowpass filtered, retaining 40 kHz. The resulting spectra consist of 256 data points linearly spaced across the frequency range (the equivalent of approximately a 39 Hz cb). All spectra were then normalised by equalising the value of the major peak under 1 kHz.

RESULTS AND DISCUSSION

The LTS data points were entered into a three-mode principal components analysis (Kroonenberg, 1983). This technique allowed examination of the structure underlying the spectra, the three passages and the interaction of both. In addition, it allowed us to examine differences among the speakers and in the between-speaker variables of sex and ethnicity. The strengths of this technique, therefore, are that three modes (in this case, LTS, passage type and speakers) can be dealt with simultaneously, and that latent components for each of these modes can be examined and used in the interpretation of the results.

The analysis reported here was conducted using the program TUCKALS3 (Kroonenberg, 1983), in which separate components are derived for each mode. As a result, several analyses were run to determine the number of components which provided the best description and fit. No formal procedures to decide this are available at present. In this study, mode one represented passage type, mode two represented LTS, and mode three, the speakers. A solution with two components for passage type and three each for LTS and speakers was selected as providing the best fit.

The analysis calculates a multiple correlation between the data and the fitted or estimated data, which in this case was .33. Given the amount of individual variation present in the spectra, it is encouraging that one third of the variance could be explained systematically. The components of each mode partition the multiple correlation into independent contributions, which sum to the multiple correlation. For all three modes, component one was shown to be the most important, in that it explained the majority of the systematic variance. In the case of the spectrum mode, component two (6% of variance) and component three (3% of variance) were less important than component one (24%). For the passage mode, however, both components were important (component one, 19%; component two, 14%), as were the first two components of the speaker mode (variance explained for the three components was 17%, 12%, and 4% respectively).

The analysis calculates joint plots of pairs of modes overlaid on one another for each component of the third mode. The joint plots of passage type and LTS for each speaker component show a clear separation of the three passages relative to the spectrum. Figure 1 presents the joint plot for speaker component one. The area covered by the LTS data points is shaded. As can be seen from the curvilinear arrow, the data points were situated in an approximate horseshoe, with the very lowest part of the spectrum placed in the upper right-hand quadrant, and the very highest part in the upper left-hand quadrant. In moving from the lowest to highest parts of the spectrum, the data points proceed via the lower right-hand and lower left-hand quadrants, in that order.

The joint plots indicate that the whole spectrum was involved in differentiating the three passage types. The school passage is separated from the other two along the horizontal axis. This reflects

the way the three passages differed on the dominance dimension, with school being perceived as low in dominance, and the other two passages as higher. As can be seen, school is characterised by the higher frequency range, indicating the importance of this part of the spectrum to the dominance dimension.

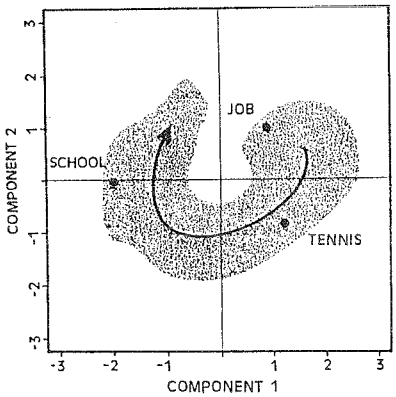

Figure 1. Joint plot for spectrum and passage type for speaker component one

The horizontal axis is approximately equivalent to component one for both passage and spectrum modes. We have seen that this component is the most important discriminator. This suggests, therefore, that the spectra are reflecting dominance more clearly than arousal. The vertical axis separates the tennis passage from the other two passages, reflecting their difference on the arousal dimension. The position of tennis suggests that the low to medium frequency range is important for the arousal dimension.

Having established this, we attempted to determine whether any one part of the spectrum discriminated the passages on one affective dimension only. A number of three-mode analyses were conducted on different frequency bands. All produced similar results, confirming that the whole spectrum was involved in some way. To check this further, a series of discriminant analyses were run on all 256 LTS data points. For each analysis, every tenth value was selected, thus providing 25 values that covered approximately the whole 10 kHz. Ten analyses of this type were conducted to account for all points in the spectrum. For eight of the ten analyses, at least one significant discriminant function appeared, separating school from job and tennis, and involving mainly data points above 3.5 kHz. For two of the analyses, however, a second discriminant function was also significant, separating tennis from school and job. This function involved mainly data points below 500 Hz and in the 2 - 2.5 kHz range. It appeared, therefore, that the whole spectrum was involved to some extent in discriminating both dominance and arousal, but that certain frequency bands were more important for one than the other.

The third mode of the analysis (speakers) was then examined. This mode indicated that all speakers to a greater or lesser extent were doing the same thing as far as the first component was concerned (that is, they all had positive scores on speaker component one). They were split on component two, however; half having negative scores, and half having positive scores. There were no significant differences on either component for sex or ethnicity. At first, we interpreted the speaker space similarly to that for the passages, assuming component one to represent dominance, and component two to represent arousal. Further examination, however, suggested this was not so. This point is taken up later.

It should be repeated that the recordings were of normal voices. No instructions were given on how to present the passages, and no mention was made of affect. One would expect to find much individual variation, therefore, with some speakers tending to separate the passages on dominance

while not emphasising arousal, other speakers doing the reverse of this, while still others emphasised both.

In order to examine the specific ways in which the two affective dimensions were reflected through the spectra, we first looked for clear examples of speakers who, from the evidence of the three-mode analysis, seemed to separate the passages primarily on one component only. In the first instance, we checked component one.

One spectral characteristic was immediately obvious. From about 3.5 kHz, the school spectrum showed considerably higher energy levels than the spectra for the other two passages (this started at about 2.5 kHz in the most extreme cases). This pattern was repeated in the spectra of all speakers who separated the passages primarily on component one. As indicated above, all speakers loaded positively on component one of the speaker space. If, therefore, as we were maintaining, this component could be interpreted as characterising dominance, we would expect the above pattern to be repeated, to a greater or lesser extent, across all speakers. This turned out to be the case. Figure 2 shows the mean spectra for all 30 speakers. The pattern is clear, although somewhat reduced, in comparison to the most extreme cases.

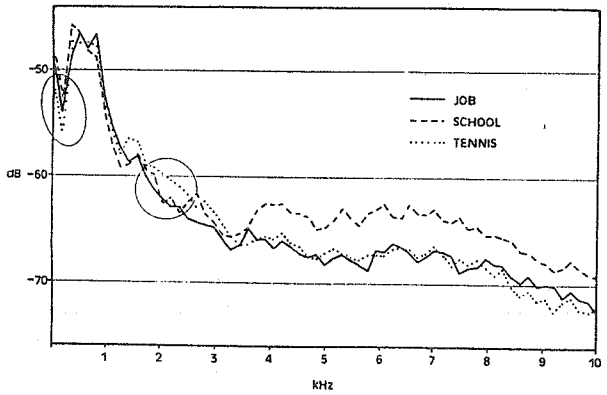

Figure 2. Average spectra for all thirty speakers

We then checked those speakers that loaded highly at each end of component two of the speaker space. Figure 3 presents the average spectra for each of these groups of speakers, with Figure 3a representing nine speakers who were highly positive on component two, and Figure 3b representing seven speakers who were highly negative. It was immediately clear that the two groups differed mainly in the top 5 kHz, well within the frequency range we had suggested was the important disciminator of dominance.

In each case, the spectra for the school passage still showed high energy levels above 5 kHz, but those for the tennis and job passages were reversed across the two groups in this range. To interpret this, we returned to the three-mode analysis. The joint plots for passage type and spectrum on component two, the one we had been examining, showed clearly that the speakers represented in Figure 3a distinguished the job passage from school and tennis, while the group represented in Figure 3b separated tennis from school and job, as illustrated in the figure. In each case, this differentiation involved the spectrum from about 5 kHz up. A check of the mean affect ratings for each of these two groups of speakers revealed that listeners had perceived the speakers in Figure 3a as lower on dominance for the tennis passage, and perceived the speakers in Figure 3b as lower in dominance for job. In other words, component two of the speaker space reflected differences in the encoding of the passages on dominance, not on arousal as we had first thought.

The three-mode analysis had indicated that the lower half of the spectrum is important in separating

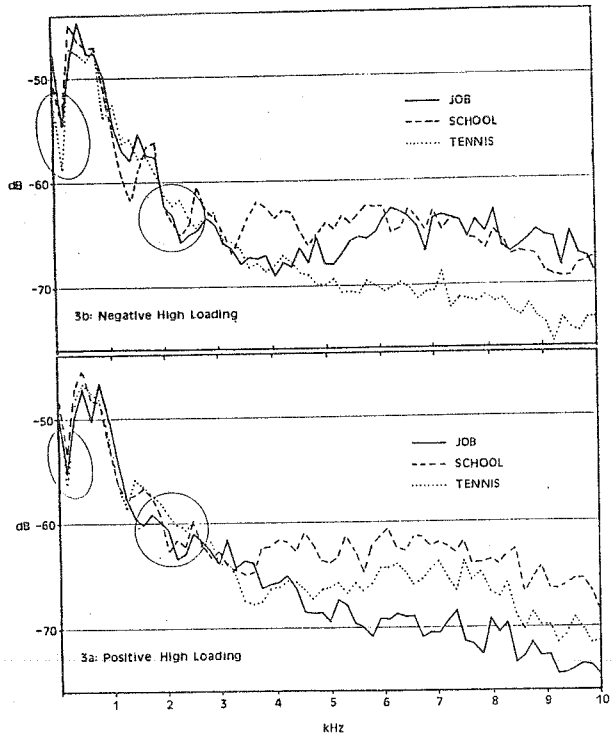

Figure 3. Average spectra for high-loading speakers on speaker component two

This is illustrated in Figure 1. A check of the univariate analyses of variance and the discriminant functions from the discriminant analyses indicated that tennis was being separated from the other two passages in two frequency bands: below 350 Hz. and around 2 - 2.5 kHz. The spectra in Figure 2 confirm this. The average spectra for all 30 speakers show the tennis passage as lower in the band below 350 Hz, and higher in the 2 - 2.5 kHz range. These two bands have been circled in the figures. In addition, the two groups presented in Figure 3 show similar separations, particularly in the higher of these two bands. It seemed, then, that all speakers distinguished the passages similarly on the arousal dimension.

It would appear, therefore, that the LTS can be used to differentiate the level of affect appearing in ordinary conversational-type speech, as well as the very strong negative affect studied in speakers with psychopathology. Our results support earlier work using the LTS to the extent that both the very low spectral range and that above 2 kHz. have been shown to be important in separating levels of arousal. It is likely that previous studies of anger, depression, and stress were finding differences in the LTS that reflected, above all, arousal levels.

None of the earlier studies included the range above 4 kHz. Our results indicate that this higher frequency range, and particularly that above 5 kHz, is important in reflecting dominance. However, as both Figures 2 and 3 show, the distinction between the passages remains much the same across the

5 - 10 kHz. range. A tentative conclusion, therefore, is that above 5 kHz, spectral shape does not change significantly in the way it reflects affect, at least in terms of affective dimensions. It should not be ignored, however. Dominance is one of the most important and robust affective dimensions in interpersonal communication (Mehrabian & Russell, 1974).

We did not find significant differences as a function of sex or ethnicity. It appears that all our speakers, men and women of European language background and long-term residents of Australia, were able to express emotion in essentially similar ways. This supports earlier findings with this group of speakers. Gallois and Callan (1988) report that listeners did not perceive differences in emotion as a function of sex or ethnic group with these speakers.

Future researchers might examine further the contribution of cultural background to the expression of emotion as measured by the LTS. In addition, and related to this, there is considerable scope for examining the long-term acoustic characteristics of discrete emotions, and their perceived features and qualities. In the meantime, it is clear that the LTS can be used to study the affect occurring in ordinary conversational-type speech.

REFERENCES

Alpert, M., Kurtzberg, R.L., Pilot, M., Friedhoff, A.J. (1963) *Comparison of the spectra of the voices of twins*, J. Acoust. Soc. Am. 35, 1877A.

Darby, J.K., Hollien, H. (1977) *Vocal and speech patterns of depressive patients*, Folia Phoniatrica 29, 279-291.

Friedhoff, A.J., Alpert, M., Kurtzberg, R.L. (1962) *An effect of emotion on voice*, Nature 193, 357-358.

Gallois, C., Callan, V.J. (1988) *Communication accommodation and the prototypical speaker: Predicting evaluations of solidarity and status*, Language and Communication.

Hargreaves, W.A., Starkweather, J.A. (1964) *Voice quality changes in depression*, Language and Speech 7, 84-88.

Hargreaves, W.A., Starkweather, J.A., Blacker, K.H. (1965) *Voice quality in depression*, J. Abnormal Psychology 70, 218-220.

Kroonenberg, P.M. (1983) *Three-mode principal components analysis: Theory and applications*, (DSWO Press: Leiden).

Mehrabian, A., Russell, J.A. (1974) *An approach to environmental psychology*, (MIT Press: Cambridge, MA).

Ostwald, P.F. (1963) *Soundmaking: The acoustic communication of emotion*, (C.C. Thomas: Springfield, IL).

Popov, V.A., Simonov, P.V., Frolov, M.V., Khachatur'yants, L.S. (1981) *Frequency spectrum of speech as an indicator of the degree and nature of emotional stress*, Zh. Vysshey Nervnoy Deyatel'nosti 1, 104-109.

Roessler, R., Lester, J.W. (1976) *Voice predicts affect during psychotherapy*, J. Nervous and Mental Disease 163, 166-176.

Scherer, K.R. (1986) *Vocal affect expression: A review and a model for future research*, Psychological Bulletin 99, 143-165.

Williams, C.E., Stevens, K.N. (1972) *Emotions and speech: Some acoustical correlates*, J. Acoust. Soc. Am. 52, 1238-1250.