THE ACOUSTICS OF VOICE AND ETHNIC IDENTITY

J. Pittam and E.S. Rintel

Department of English
The University of Queensland

ABSTRACT - This paper examines the acoustic characteristics of ethnic identity. In particular, it looks at the long-term acoustic properties of Anglo-, Vietnamese- and Hong Kong Chinese-Australians. The paper is a preliminary report of a larger project, and focuses on long-term spectral features of four speakers from each ethnic group. Three-mode principal component analysis is conducted on long-term average spectra to find group differences among the speakers.

INTRODUCTION

This paper is part of an ongoing project exploring the vocal communication of ethnic identity. That is, it examines the acoustic characteristics of ethnic identity, listeners' perceptions of ethnic identity through the voice, and the attributions that are made about the speakers by listeners from the voice. The project is, therefore, a multidisciplinary one, drawing mainly on speech science and social psychology for its underlying theory and methodology. This particular paper is a preliminary report that looks at some long-term acoustic properties of ethnic identity. To date, little work has focussed on this aspect of identity as it is communicated by the voice.

The literature concerning the vocal communication of ethnic identity, although small, does provide evidence to suggest listeners can distinguish the ethnicity of speakers from the voice alone (Gallois & Callan, 1991). In terms of encoding, however, there is little indication of what aspects of the voice are involved. In a related literature, Laver (1980) points to regional varieties of British English being characterised by long-term settings such as velarisation and denasalisation. The question remains, however, just what vocal features communicate ethnic identity, and how these are encoded in the acoustic signal.

The term "voice" is used here in the sense adopted by Pittam (1994), to cover long-term characteristics of an individual's vocal sounds, such as fundamental frequency and amplitude, and measurements of these, including complex measures based on the combination of these features. Pittam also includes tension and temporal factors, as well as articulatory settings (see also Laver, 1980). It is the long-term acoustic measurement of voice, however, that is of concern here. In particular, we are concerned with the long-term spectrum of voice (LTS).

The long-term spectrum has attracted interest from researchers in many disciplines (see Pittam, 1987; Pittam & Millar, 1989, for reviews). It is the averaged intensity spectrum computed across a selected frequency range for continuous speech. It averages the contribution of individual speech sounds in describing the spectral characteristics of the voice as a whole. As such, the LTS tends to stabilise in shape after averaging approximately 30 seconds of speech, becoming a measure of the averaged spectral energy underlying speech. This is the primary measure adopted here.

The statistical method adopted to compare the spectra is three-mode principal components analysis. This technique allows examination of the latent component structure underlying, in this case, the ethnic groupings of the speakers, the spectral values, and the speaking tasks subjects engaged in, thereby providing a measure of any separation among groups of speakers and/or tasks on the basis of the spectral values, as well as indicating how the three modes interact. One strength of the method is that it can handle large amounts of data, thus allowing all the spectral points in the LTS to be entered into the analysis.

This study sets out, then, to examine differences in the characteristics of the LTS across groups of speakers from different ethnic backgrounds. Three ethnic groupings were selected for study: those of English background, Vietnamese background, and Hong Kong Chinese background. Generally, Asian cultures differ from western cultures in the way they use non-verbal behaviour, including voice. These

differences can be seen on a number of cultural communicative dimensions such as levels of collectivism, synchronisation and intensity of behaviour, which may be reflected in the long-term acoustic characteristics of voice, and may impact upon the attributions made by listeners about the speakers they hear.

METHOD

Subjects

In the larger project, 10 female and 10 male speakers from Anglo-Australian, Vietnamese-Australian, and Hong Kong Chinese background took part in the study. Two female and two male speakers from each ethnic group were analysed for the present paper. Speakers were aged between 16 and 35, and, in the case of the Vietnamese group, bilingual in English and Vietnamese, and in the case of the Chinese group, bilingual in English and Cantonese. Speakers attended the recording sessions in pairs of same-sex acquaintances.

The four Vietnamese-Australian speakers were born in Vietnam of Vietnamese parents. Each had spent at least half their lives in their home country, and none had spent time in a third country. The Anglo-Australians were all born in Australia of Anglo-Australian parents, had not spent time in another country and were all monolingual in English. The Chinese speakers were split between Chinese-Australians born in Australia of Hong Kong parents, and Hong Kong Chinese born in Hong Kong of Hong Kong Chinese parents. None had spent time in a third country, and the Chinese Australians had not spent time in any other country including Hong Kong.

No speaker had a pathological condition that might affect the voice, and none had had surgery relating to the vocal system. None experienced respiratory problems or hearing problems, and none had medical problems affecting the voice at the time of recording. None were smokers. All speakers indicated that they believed their voice sounded as it usually did.

Speaking Tasks

After completing a questionnaire giving demographic information, details of language background and usage, and other information concerning factors that may influence voice quality, such as those noted above, subjects took part in three recording tasks. Each separately recorded in English a standard reading passage, then with their partner engaged interactively in two other tasks. The latter were included to invoke more "naturalistic" speech. Firstly, each speaker described to the other a route on a map, such that the partner could reproduce it on a blank map. Speakers were given five minutes each to complete this task which was conducted in English. Secondly, each speaker constructed a small Lego assemblage and then described this to the other, once again to allow the partner to reproduce it. Five minutes were allowed, and this task was conducted in the first language of the speakers (Vietnamese for the Vietnamese-Australians; Cantonese for the Chinese; English for the Anglo-Australians).

Recording and Digitisation

Recordings were made using a Marantz CP430 analogue recorder and digitised for each speaker using the program Signalyze 3.12. Sampling frequency was set at 44000 Hz. Digitised signals were stored on a Power Macintosh 8500. The full reading passage was digitised for each speaker, and a minimum of 30 seconds was digitised from each of the other two tasks. As the latter were interactive, speakers' partners were able to ask questions and make comments throughout the tasks. Partners' talk was edited out of the digitised signals, leaving only the primary speaker in each case.

RESULTS

Acoustic Analysis

As an initial check of the digitised recordings, an FFT-Comb pitch extraction routine (Signalyze 3.12) was used and average fundamental frequency (F0) and standard deviation calculated for all speakers.

Standard deviation was used as a measure of F0 variance. A series of t-tests was computed for the averaged F0 for males and females separately for all possible pairings of sex by ethnic group and for sex by ethnic group by task. No significant differences were found. Similarly, t-tests showed no significant difference between any pairing of ethnic group by task for F0 variance.

Long-term average spectra were then produced for each recording (that is, three spectra for each speaker) across the frequency range 0-22KHz using the program Signalyze 3.12. This produced 128 equi-spaced points of averaged amplitude across this range. As no spectrum displayed energy above 17.7KHz (103 values), the 25 values above this frequency were discarded from the subsequent analyses.

Three-mode Principal Components Analysis

The spectral values were entered into a three-mode principal components analysis (Kroonenberg, 1983). The 12 speakers were entered as mode 1, the 103 spectral values as mode 2, and the three tasks as mode 3. In the present study, it was possible that the latent components underlying mode 1 would represent the three ethnic groups. The assumption of three-mode analysis is, then, that the latent components describe the systematic variance in the spectrum for each mode.

The program used here was TUCKALS3 (Kroonenberg, 1983). Several analyses were run to determine the number of components in each mode that provided the best description and fit. No formal procedures to decide this are currently available. A solution with two components each for speaker mode and spectrum mode and one for task was selected as providing the best fit for the data.

The single component for the task mode suggested that there were no significant differences in the spectra across the tasks and, therefore, across the language types. To check this, three-mode analyses were run on each ethnic group separately with only the map and Lego tasks analysed (the latter task was conducted in the first language of the speaker, while the former was in English). Once again, in each case, a single component provided the best fit, and the component weights for the two tasks in each case displayed the same polarity and were not significantly different from one another. In other words, neither the language used nor the task significantly affected the contour of the spectra.

One methodological point coming out of this is that the lack of significant difference across the tasks for the Anglo-Australian group indicates that the LTS did stabilise in the 30 seconds of speech digitised. The language used in both the map and Lego tasks for this group was English. The lack of significant difference in the spectra for these two tasks, then, was not confounded by language type.

The three-mode analysis calculates a multiple correlation between the data and the fitted or estimated data, which in this case was 0.58. This is slightly better than many three-mode applications (Kroonenberg, 1983). Given the amount of variation in the 103 spectral values, it was encouraging that more than 50% of the variance could be explained systematically. The components of each mode partition the multiple correlation into independent contributions (which, therefore, sum to the multiple correlation). For all three modes, component one was shown to be the most important (there was only one component for mode 3), in that it explained the majority of the systematic variance in each case. For both speaker and spectrum modes, component one explained 45% of the variance.

Figure 1 shows the 12 speakers graphed out in two-dimensional space (corresponding to the two components). As can be seen, component one (horizontal) most clearly separates the ethnic groups, particularly the Vietnamese-Australian speakers from the Chinese speakers. The Anglo-Australian speakers are neither clearly separated from the other groups nor do they clearly group together, although they do tend to sit between the other two groups. This accords with component one explaining the most variance. It can be noted that the two Chinese speakers born in Hong Kong are C1 and C3. As can be seen, they are not grouped separately from the two Chinese-Australian speakers (C2 and C4).

Component two (vertical) of the speaker mode is less easy to interpret in terms of ethnic identity. As can be seen from Figure 1, there is a tendency for this component to separate Chinese speakers to some extent from the other two groups. This is by no means a clear separation, however. It is hoped

that when the full set of 60 subjects are analysed, the function of a second latent component, in terms of ethnicity, will be clarified.

One factor that may help explain component two is sex of speaker, which may also be influencing the spectral contour. In the small subset of speakers presented here, however, the sex distinction is not clear for all three groups. Thus, in Figure 1, V3 and V4 are female, while V1 and V2 are male. Similarly, C1 and C4 are female, and C2 and C3 are male. As can be seen, component two separates the latter on gender lines reasonably clearly, but the former less so. The Anglo-Australian speakers are, once again, not well grouped. A1 and A2 are female, while A3 and A4 are male. The last speaker, particularly, is not positioned appropriately if the main function of component two is to separate the speakers on gender. Sex of speaker, then, may help explain component two, but only in part, and only as a function of ethnic group.

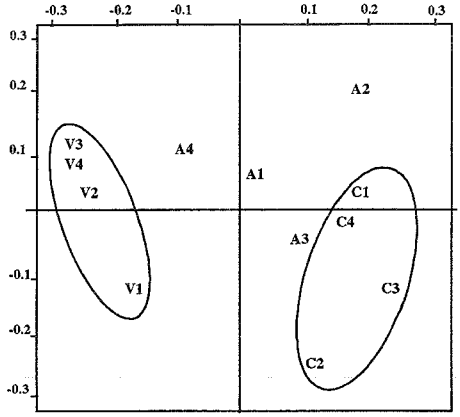

Figure 1. Speaker mode: component one (horizontal) vs component two (vertical)

A = Anglo Australian; C = Chinese; V = Vietnamese Australian

To illustrate the distinction between Vietnamese-Australian and Chinese speakers spectrally, Figure 2 shows examples of long-term spectra for these two ethnic groups. For clarity, one spectrum (that for the map task) is shown representing each group. The Vietnamese-Australian speaker is the one labelled V3 in Figure 1 above, while the Chinese speaker is C3. While only one spectrum is illustrated in each case, the differences between the Vietnamese-Australian and Chinese spectra are clear, particularly above approximately 1600Hz. We can note that no spectrum for the Vietnamese-Australians displayed energy above approximately 5KHz, while all 12 spectra for the Chinese speakers displayed the "hump" in energy in the higher frequencies to a greater or lesser extent. The spectra illustrated in Figure 2, then, are quite representative of their respective groups.

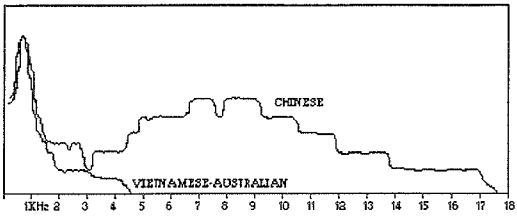

Figure 2. Example spectra of a Chinese and Vietnamese-Australian speaker

The two components underlying the spectral mode are less easy to interpret. Essentially, component one separates most of the higher part of the spectrum (approximately 3800Hz-13900Hz) from both the low (<3800Hz) and very high (>13900Hz) parts of the spectrum. Component two is even less clear. It is likely that future analyses would be best served by a single dimensional solution underlying the spectral mode. Component one in part, then, seems to separate out that part of the spectrum where we would expect to find the most similarity across speakers - the part of most importance to speech - from the higher levels. The two dimensional space is not graphed here because of the difficulty of showing these separations clearly with 103 spectral values plotted on the same graph.

The interpretation of the relationships between each pair of modes is achieved by calculating joint plots of each pair for all components of the remaining (third) mode. Thus, in this case, speaker and spectral values could be examined together in a joint plot for the single component of the task mode. This allowed us to see whether particular parts of the spectrum, in general terms, were important for each ethnic group.

The clearest and most continuous separation of all four Vietnamese-Australian speakers from all four Chinese speakers was in the range above approximately 3500Hz. This can be seen illustrated in Figure 2 for the two examples shown. Although the Anglo-Australian speakers did not separate out as clearly as the other groups when all spectral values were considered, they were quite distinct as a group in the frequency range 6500Hz to 8250Hz, with spectral values well below that of the Chinese speakers. Similar to the situation with the spectral mode, the interaction of speaker and spectral modes is not graphed out due to the difficulty of representing so many values on a single plot.

DISCUSSION

This has been a preliminary report of an ongoing examination of the communication of ethnic identity by the voice. This particular paper has presented results of an acoustic analysis conducted on a subset of speakers from Anglo-Australian, Vietnamese-Australian, and Chinese backgrounds. Long-term spectra on recordings made of three tasks undertook by four speakers from each of the three ethnic backgrounds were examined using three-mode principal components analysis.

Although only a preliminary study, several interesting findings were reported, some not related to the main aims of the project. Thus, we have provided evidence of the stabilisation of the LTS after approximately 30 seconds. The comparison of the map and Lego tasks for the three groups separately effectively eliminated both language type and the specific speech sounds of a single language as confounding variables on the spectrum. Beyond this, however, the findings are only suggestive and leave us with more questions than answers.

The major finding of the analysis, the clear separation of Vietnamese-Australian speakers from Chinese speakers, is of considerable interest. It is also not what was expected. From what is known of the differences in the encoding of nonverbal behaviour between Asian and Western cultures, we had anticipated differences between the Anglo-Australian speakers and the other two groups. This finding above all will be examined once the full set of recordings from all 60 speakers has been digitised and analysed. Related to this finding is the fact that the Anglo-Australians did not group together as clearly as did the other two groups, particularly when the groupings are viewed in terms of the latent components. This also needs to be examined further. In particular, the one part of the spectrum in which they were grouped together, at the same time as being quite different from the other groups (6500-8250Hz), will be considered in more detail with analyses conducted on just that part of the spectrum.

Of some interest is the lack of significant difference between the Chinese-Australian speakers and the Hong Kong born Chinese. This will be examined further when all speakers have been analysed, thus providing sufficient numbers of each type of Chinese speaker to allow one to be compared directly with the other. The lack of difference may suggest that accent features learned early within the family setting have a major impact on long-term aspects of accent that continue into adult life. In this case, the Chinese-Australian speakers may well have developed long-term accent features from their

Chinese families which remained after short-term features were changed as they learned English within the larger Australian context. These long-term accent features would then be more similar to the Hong Kong Chinese speakers than the Anglo-Australian speakers, something that is reflected in Figure 1.

The underlying assumption of this study is that ethnicity is manifested vocally by a speaker's accent. This, however, is to use the term accent to cover not simply the short-term speech sounds and the paralingual phenomena such as intonation and stress patterns, but much longer term vocal features that we are here calling voice quality. The findings of this study suggest the LTS may be capturing long-term accent features of voice quality that are linked to a speaker's ethnic identity. Component one of the speaker mode captured 45% of the systematic variance in the spectra, and the prime function of this component seemed to be to separate at least two of the ethnic groups.

The LTS, however, may capture other features of identity. Sex of speaker was also separated, this time on component two of the speaker mode, but once again most clearly for the two Asian groups. In addition, while both of the two groups separated the spectra on sex of speaker, they each did so on that part of component two on which they as an ethnic group were situated. In other words, component two was not simply a gender component, but a component on which sex and ethnicity interacted. That ethnicity was implicated in the second component as well as clearly dominating the first, suggests that ethnic identity is indeed captured by the LTS.

Until we have analysed the full data set, it would be premature to interpret these findings too closely. Once the groupings have been repeated for the larger set of speakers, and once the spectral contours have been replicated, we can set about linking this back to ethnic identity as manifested in the accent. With such small numbers of speakers, then, these results must remain speculative. It does seem, however, that we can claim that, in general terms at least, the LTS has been shown to be useful for capturing long-term aspects of voice. It also seems that it is the higher spectral levels that are most important for this type of information. This supports earlier work by Pittam, Gallois, and Callan (1990) that showed a similar finding for the encoding of emotion. Similarly, one can point to the usefulness of the statistical analysis, allowing not only large amounts of data to be analysed - a necessity when working with the LTS - but providing the possibility of finding a meaningful dimensional structure underlying the variables under study and the interaction of those dimensions.

REFERENCES

Gallois, C. & Callan, V.J. (1991) "Interethnic accommodation: The role of norms", In H. Giles, J. Coupland & N. Coupland (eds.), *Contexts of accommodation,* pp. 245-269, (Cambridge University Press: Cambridge).

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

Laver, J. (1980) *The phonetic description of voice quality*, (Cambridge University Press: Cambridge).

Pittam, J. (1987) "The long-term spectral measurement of voice quality as a social and personality marker: A review", Language and Speech 30, 1-12.

Pittam, J. (1994) *The voice in social interaction: An interdisciplinary approach*, (Sage: Thousand Oaks).

Pittam, J., Gallois, C. & Callan, V.J. (1990) "The long-term spectrum and perceived emotion", Speech Communication 9, 177-187.

Pittam, J. & Millar, J.B. (1989) *Long-term spectrum of the acoustics of voice: An annotated and classified research bibliography*, (Indiana University Linguistics Club: Indiana).