AUDITORY AND F-PATTERN VARIATIONS IN AUSTRALIAN *OKAY*: A FORENSIC INVESTIGATION

Jennifer R. Elliott
Phonetics Laboratory
School of Language Studies
The Australian National University

ABSTRACT. An understanding of the acoustic properties, as well as the nature of within- and between-speaker variation, of words which occur with high frequency in natural discourse, is of great importance in forensic phonetic analyses. One word which occurs with relatively high frequency in natural discourse, including telephone conversations, which are often a source of data in forensic comparisons, is *okay*. This paper presents the initial findings of a study of auditory and F-pattern variations in *okay* in a natural telephone conversation spoken by six male speakers of general Australian English. Seven pre-defined sampling points are measured within each token to determine the most efficient sampling points and formants for distinguishing between-speaker variation from within-speaker-variation in *okay*. F-ratios at these seven sampling points are calculated as a mean of ratios of between- to within-speaker variation. The greatest F-ratio is shown to be for F_4 at voice onset of the second vowel. Forensic implications are discussed.

1. INTRODUCTION

When phoneticians are asked to compare samples of speech for forensic purposes they are faced with a specialised case of speaker verification which involves comparing a sample of speech which is known to be associated with a crime, with another sample of speech from a known person who is suspected of being involved in the crime. This forensic application of speech analysis is based on the assumption that there will be greater variation between speakers than within a speaker.

Nolan (1983: 6-14) notes that forensic speaker verification or identification tasks are inherently more complicated than other forms of speaker identification, where a sample of speech is compared against another predetermined sample for the purpose of authenticating or verifying a speaker is who he or she claims to be. Apart from the obvious difficulties inherent in comparing speech samples recorded at different times and usually under very different conditions, the speech samples used in forensic phonetics are invariably both uncontrolled and restricted in content, leaving a minimal amount of speech for analysis and comparison. The recording of the criminal, for example, may constitute only a few short words. It is desirable that the linguistic data from both samples used in a forensic comparison are, if possible, linguistically equivalent, and the best results are likely to be obtained when the same lexical items are compared. For this reason words which occur frequently in conversation are likely candidates for analysis and comparison.

One word that is used frequently as a tool of negotiation in conversational English is *okay*. This word functions both as a response such as agreement, acceptance or confirmation to preceding talk, and/or as a transitional device between two stages of a conversation (Merrit 1984; Condon 1986). Furthermore, as Schegloff (1979, 1986) and Schegloff and Sacks (1984) have demonstrated, *okay* occurs frequently in both openings and closings of telephone conversations, which in turn are the most common source of recordings used in forensic comparisons. The question therefore arises: is *okay* an appropriate word to use in forensic analysis, and if so, how useful is it for distinguishing between speakers?

Research by Rose (1997, 1999), in which the within- and between-speaker differences in *hello* spoken by six speakers were examined, demonstrated that even similar sounding speakers "can be distinguished on the basis of significant differences in their acoustics" (Rose 1997: 35). Based on these findings, a similar hypothesis was proposed for the present study: that there will be greater variation in the acoustics of *okay* between speakers than within a speaker. If this hypothesis was confirmed then a secondary question would arise: which parts of the word *okay* provide the clearest evidence of between-speaker differences? The research was designed both to test the hypothesis and, if the null hypothesis was disproved, to seek an answer to this question. Although both auditory and acoustic analyses are indispensable in forensic analysis, one of the key measures of comparison of forensic phonetic acoustic analysis is the formant- (F-)pattern of short-term segments. This paper describes briefly the auditory variations in the phonetic realisations of ten tokens of *okay* from each of six different speakers of general Australian English (as described by Mitchell & Delbridge 1965, Burridge & Mulder 1996, for example), and reports the F-pattern variations of these same tokens when examined from an acoustic phonetic perspective. This study represents the first stage in a broader research project on the subject of auditory and acoustic within- and between-speaker variations in Australian *okay*.

2. EXPERIMENT DESIGN AND DATA COLLECTION

In keeping with the nature of data used in forensic phonetics, a premium was placed on the data being collected from natural conversation. A map task was devised to engage pairs of participants in a conversation requiring negotiation, potentially leading to the elicitation of several tokens of *okay* from each speaker. In order to encapsulate each conversation as a closed speech event, the task was carried out by telephone, thus providing a distinct beginning and end to each interaction. Recording a speaker engaged in a telephone conversation had two additional advantages. Firstly, it enabled a clean speech signal of a single speaker conversing with someone else to be recorded without the attendant confusion of overlapping talk from the other speaker (a common characteristic of natural conversation). Secondly, since there was no eye contact between the speakers, all communication had to be verbal, thus increasing the opportunity for negotiation, and hence the likelihood of eliciting numerous tokens of okay. The recordings used in acoustic analysis were made directly, and not through the telephone.

The study involved six native speakers of general Australian English working in pairs, as indicated in Table 1. All participants were aged between 16 and 20 years, and were from similar socio-economic backgrounds. In order to minimise the effect of convergence of linguistic styles between the participants (Giles & Coupland 1991: 60-93), each pair was also well acquainted. In addition, a number of the participants were from the same family (they were either brothers or cousins), and although they were not necessarily paired together, it was hoped that this would impose a slightly higher level of control over the possibility of confounding sociolinguistic variables.

Table 1. Pairs of participants

Caller	DL	EO	GO	MO	JE	PE
Recipient	JE	PE	PE	JE	MO	GO

The map task involved two similar, but not identical maps. The caller was required to guide their partner (the recipient of the telephone call) through a predetermined route marked on the map. The negotiation of the differences between the maps would provide the opportunity for the elicitation of tokens of okay. The caller was recorded directly in the recording studio of the Phonetics Laboratory at the Australian National University, using a Nakamichi 500 stereo cassette deck and a Sony ECM-909A microphone. From this recording the ten tokens of okay which could be most easily isolated from the surrounding talk, and which had the least excess noise, were extracted for acoustic analysis.

3. AUDITORY ANALYSIS

The Australian Oxford Dictionary (published in 1999) suggests that the Australian English pronunciation of *okay* /oʊ'keɪ/ has three phonemic segments, consisting of two diphthongs (V_1 and V_2), separated by a voiceless velar stop (C). Auditory analysis of each of the sixty tokens studied

Table 2. Occurrences of different phonetic realisations of Australian *okay* segments by each speaker

V_1 /oʊ/	DL	EO	GO	MO	JE	PE
ə	3	4	8	0	6	0
ɐ	6	5	0	7	3	0
ɑ	1	1	2	1	1	0
ɵ	0	0	0	0	0	9
ɛ	0	0	0	1	0	1
ɛ°	0	0	0	1	0	0
C /k/						
kʰ	10	10	6	10	9	4
k	0	0	4	0	1	1
g	0	0	0	0	0	4
x	0	0	0	0	0	1
V_2 /ɛɪ/						
ɛɪ	8	3	9	8	3	9
ɛ°	0	1	1	0	0	0
ɛ	0	6	0	0	7	1
æ	2	0	0	1	0	0
ʲɛɪ	0	0	0	1	0	0

showed considerable variation in the phonetic realisations of these segments, both within and between speakers. Phonetic realisations of each of the three segments from auditory analysis are set out in Table 2.

Table 2 shows that V_1 was realised as a diphthong only once out of the 60 tokens analysed. Interestingly, this particular token was also irregular in that V_2 was palatalised ([ɛ°kʰʲɛɪ]). Thus the generalisation can be made that V_1 of *okay* in conversational general Australian English is usually realised as a monophthong. Moreover, this monophthong was in the majority of cases, centralised to [ə] (a typical realisation of unstressed vowels) or centralised and lowered to [ɐ]. One token of the low back vowel [ɑ] was also elicited from each speaker except PE, whose V_1 was realised 90% of the time as the slightly raised central rounded vowel [ɵ]

The /k/ was most commonly realised as an aspirated voiceless velar stop. For example this was the case 100% of the time for DL, EO and MO, and 90% of the time for JE. The stop was aspirated in six of GO's tokens, while the remaining four were unaspirated voiceless stops. PE again differed the most, with only four tokens being aspirated, while one was a voiceless unaspirated stop, four were realised as voiced stops, and in one token the consonant was fricated throughout, without an audible hold phase.

With V_2, 43 of the 60 tokens were realised as diphthongs. In keeping with the findings of previous studies of Australian English (for example Harrington et.al. 1997,) the first target for this vowel was consistently lowered, and was realised as [ɛ] rather than [e]. In two instances, the offglide was more central than high, but in one of these cases, this may have been due to anticipatory coarticulation (Laver 1994: 379) for a bilabial approximant, /w/, which followed in the next word, however this requires further investigation. In a number of instances, V_2 was not realised as a diphthong at all, but was realised simply as an open-mid front [ɛ]. Six instances of this were elicited from EO, seven from JE and one from PE. Extreme lowering of V_2 to [æ] was also occasionally heard, twice by DL and once by MO, and in each of these instances V_2 was also realised as a monophthong. The incidence of both /ɛ/ and /æ/ in V_1 of *okay*, suggests that there is possibly a choice of phonemes for this syllable in Australian English (c.f. Rose's (1997, 1999) findings for V_1 of Australian *hello*).

While the suprasegmental structure will not be discussed in detail in this paper, it should be noted that there is also considerable variation in the realisation of stress. In all but one instance, the major stress fell on the second syllable: EO provided the only token where the stress fell on S1, and the general lenition and centralising of V1 noted above may well be accounted for in terms of stress.

4. ACOUSTIC ANALYSIS

Tokens were digitised at 16 kHz, and the F-pattern was analysed on a CSL 4300 by generating wideband spectrograms, and using the FFT power spectrum facility overlaid with the LPC filter response at selected sampling points. A filter order of 20 kHz was used, with hamming window and 100% preemphasis. The first four peaks were measured to extract an estimate of the centre frequencies of the F-pattern, based on the expected frequencies for each given phonetic segment.

The primary aims of the experiment were to determine whether or not it is practicable to use *okay* in forensic comparisons, and if so, which part of the word okay provides the best F-pattern for determining between-speaker differences. Since the tokens were to be used for comparing both within- and between-speaker variations, it was essential that the sampling points were also comparable across all tokens. To ensure the integrity of measurements between all the tokens, seven sampling points were chosen at which to measure the first four formants. The decision to use these particular sampling points was motivated by the goal of extracting as much acoustic information as possible which could highlight significant differences between speakers.

The seven sampling points, illustrated in figure 1, were identified as follows:

S_1 1. within the first three regular glottal pulses of V_1 (V_1 onset);
2. within the last three regular glottal pulses of V_1 (V_1 offglide);

S_2 3. at consonant release (C release);
4. at phonation onset follow the release phase (PO);
5. within the first three glottal pulses of V_2 (V_2 onset);
6. at the lowest point of F_2 within V_2 (V_2 mid); and
7. at the highest point of F_2 within V_2 (V_2 offglide).

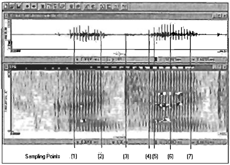


Figure 1. Wideband spectrogram showing sampling points of *okay* tokens

Table 3. F-ratios for each formant at each sampling point in order of magnitude.

Sampling Point	Formant	F-ratio	Confidence level
V2 onset	F4	32.367	.000
V2 offglide	F4	29.937	.000
V2 onset	F3	25.791	.000
V2 onset	F1	21.631	.000
PO	F3	19.439	.000
PO	F4	19.363	.000
V1 offglide	F3	18.102	.000
V2 mid	F4	16.581	.000
V2 mid	F1	14.419	.000
V2 offglide	F3	12.665	.000
V1 onset	F1	13.662	.000
V2 onset	F2	10.836	.000
V2 mid	F3	10.239	.000
V1 offglide	F2	9.694	.000
V1 onset	F3	9.635	.000
PO	F2	9.093	.000
C release	F3	8.872	.000
V1 onset	F4	8.036	.000
PO	F1	7.012	.000
C release	F1	5.124	.001
V2 mid	F2	4.193	.003
V2 offglide	F2	3.850	.005
V1 offglide	F4	3.656	.006
C release	F4	3.185	.014
V1 onset	F2	2.903	.022
C release	F2	2.802	.025
V2 offglide	F1	2.142	.074 (n.s.)
V1 offglide	F1	1.514	.201 (n.s.)

The estimated centre frequencies of the first four formants for each sampling point were collated for statistical analysis. One method which has been shown to be effective in determining the most efficient parameters for distinguishing between speakers is the analysis of variance, in which the ratio of variance of speaker means to the mean within-speaker variation is calculated (the F-ratio) (Pruzansky & Mathews

1964; Wolf 1972; Nolan 1983; Rose 1999, 1997). The greater the magnitude of the F-ratio, then correspondingly, the greater between- to within-speaker variation can be expected. A series of univariate ANOVAs was performed to calculate the F-ratio for each formant at each of the seven sampling points. The sampling points with the highest F-ratios were deemed to represent the most promising parameters for distinguishing between speakers. The results in order of magnitude of the F-ratio are set out in Table 3.

The results indicate that the most efficient sampling point for distinguishing between speakers in Australian *okay* is F_4 at V_2 onset, with an F-ratio of 32.367. This is followed closely by F_4 at the V_2 offglide (F=29.937), while the next most efficient sampling points are F_3 at V_2 onset (F=25.791) and F1, also at V_2 onset (F=21.631). The magnitude of these F-ratios is sufficiently high to suggest that these sampling points are acceptable for distinguishing between speakers, although higher F-ratios have been found to occur in a range of other parameters which have not been considered here. For example, Wolf (1972: 2048) found "individual fundamental frequency parameters had the highest F ratio of all the parameters investigated" in his study, with F-ratios for F_0 ranging from as high as 84.9 down to 30.9. In Wolf's study, the only formant measurements taken were F_1 and F_2 for vowels /æ/, /a/ and the schwa /ə/, and F-ratios for these ranged from 46.6 (for F_2 of /æ/) down to 15.5 for F_1 of /æ/. The highest F-ratio in the present study falls at around the median result of Wolf's study, while the four highest F-ratios noted above for the present study all occur within the top two-thirds of Wolf's values.

A further comparison could be made with Nolan's (1983) study in which F-ratios were calculated for 15 speakers for F_1, F_2 and F_3 of the two English liquids, /l/ and /r/. Nolan found that F_3 provided the highest F-ratios (F=216.9 for /r/ and F=77.8 for /l/). Although, as Nolan (1983: 102) notes, the high value for /r/ may be due in part to "an artefact of the formant extraction process", these values are still considerably higher than the F-ratios obtained from Australian okay, which compare more closely with Nolan's lowest F-ratios, which were recorded for the two lower formants of /l/ (for F_1 F=17.7, and for F_2 F=21.6). Nevertheless, Nolan (1983: 115) concludes that "Spectral information from initial allophones of /l/ and /r/...yield moderate identification rates...[and] are worth incorporating in a speaker identification scheme making use of segmental information." The comparability of the top 25% of F-ratios found in Australian *okay* (set out in Table 3) suggests that the formants at these sampling points are also worthy of incorporation in a forensic analysis, particularly as this data was recorded from natural speech events, rather than having been obtained from read-out speech, as was the case for both the Wolf and Nolan studies. (Greater within-speaker variation would be expected from natural speech than from read out speech, thus lowering the F-ratios.)

Just over 50% of the F-ratios were below 10, indicating that these parameters are the least efficient formants and sampling points in Australia *okay* for distinguishing between speakers. Nevertheless, with the exception of the two lowest F-ratios (for F_1 of the offglides of each of V_1 and $V_{2)}$) they were still statistically significant, and could be used. It should also be noted the the highest F-ratio for each formant was always found at voice onset of V_2, that is, within the first two or three glottal pulses of V_2.

Further analysis of the data in this study using a Bonferroni post hoc test for the analysis of variance, showed that an average of 8 out of a possible 15 between-speaker distinctions were found in each of these top 25% formant X sampling points. The highest number of between-speaker distinctions occurred in F_4 at V_2 onset, where 9 statistically significant differences between speakers were found. The more conservative Scheffé post hoc test (which may be preferable to use in a forensic analysis) indicated that on average, 7.3 distinctions were made in the top 25% of F-ratios, with 8 out of 15 speakers showing a significant difference for F_4 at V_2 onset.

One point which should be made is that the integrity of using the higher formants (and particularly F_4) in the context of telephone recordings is highly questionable, due to the bandpass limitations which affect the acoustic properties of the transmitted signal (Rose & Simmonds 1996). When this is taken into account, the actual sampling points which may prove useful in forensic analyses, where data has been gathered from recordings of telephone conversations, is further reduced.

5. CONCLUSION

The analysis of F-pattern variations of *okay* in natural conversation has shown there is greater between-speaker variation than within-speaker variation in the F-pattern of *okay* in Australian English, making this frequently occurring word potentially useful in forensic comparisons. Given the questionable reliability of F_4 in speech samples recorded over the telephone, it would appear that the most efficient formants and sampling points for measuring between-speaker differences are likely to be F_1 and F_3 at voice onset of the second vowel, while F_3 at both PO and V_1 offglide should also be useful. Additional measurements for F_1 at V_1 onset and midway through V_2, and for F_3 at V_2 offglide may also be valuable in distinguishing between speakers. F_2 has not shown itself to be a particularly efficient parameter at any sampling point in okay. In directly recorded data (as opposed to data collected via telephone), the most efficient sampling point for distinguishing between speakers is unquestionably at voice onset of V_2, where a significantly high F-ratio is obtained for all of the first four formants.

No forensic analysis should rely on F-pattern alone for determining likelihood ratios. While auditory analysis is also clearly important, ongoing research on the potential value of using the frequently occurring word, okay, in forensic investigations will consider other acoustic parameters, including fundamental frequency and duration, and will attempt some form of quantification of coarticulatory effects, such as the extent of "velar pinching" in V_1 triggered by the following consonant. In addition a survey will be made of intonational and stress patterns of each token, and how these relate to their discourse function. Forensic phonetics would also benefit from similar studies of other high frequency words, such as *yeah*, *so*, *well* and *y'know*, as well as other discourse markers such as *oh*, *ah* and *um*, and these could be the focus of future research.

NOTES

This paper was first presented at the Eighth Australian International Conference on Speech Science and Technology, Canberra 5-7 December 2000. I would like to thank Phil Rose for his readiness to provide guidance and advice while undertaking this project, and the two anonymous conference reviewers for their very constructive comments.

REFERENCES

Burridge, K. & J. Mulder. (1996) *English in Australia and New Zealand. An Introduction to Its History, Structure, and Use*. Melbourne: Oxford University Press.

Condon, S. (1986) "The Discourse Functions of OK." *Semiotica* 60, 73-101.

Giles, H. & N. Coupland. (1991) *Language Contexts and Consequences*. Milton Keyes: Open University Press.

Harrington J., F.Cox & Z. Evans. (1997) "An Acoustic Phonetic Study of Broad, General, and Cultivated Australian English Vowels." *Aust. J. Linguistics*, **17**, 155-184.

Laver, J. (1994) *Principles of Phonetics*. Cambridge: Cambridge University Press.

Merrit, M. (1984) "On the Use of OK in Service Encounters." In J. Baugh & J Scherzer (eds.), *Language in Use: Readings in Sociolinguistics*. New Jersey: Prentice-Hall Inc.

Mitchell, A.G. and A. Delbridge. (1965) *The Pronunciation of English in Australia*. Sydney: Angus & Robertson.

Nolan, Francis. (1983) *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Pruzansky, S. & M.V. Mathews. (1964) "Talker-Recognition Procedure Based on Analysis of Variance". *J. Acoust. Soc. Am.* **36**(11), 2041-2047.

Rose, Philip J. (1999) "Long- and short-term within-speaker differences in the formants of Australian hello." *J. Internat. Phonetic Assoc.*

Rose, Philip J. (1997) "Differences and Distinguishability in the Acoustic Characteristics of Hello in Voices of Similar-Sounding Speakers - A Forensic Phonetic Investigation". *Austr. Rev. Appl. Linguistics* **22**(1), 1-42.

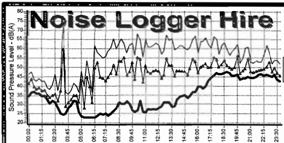
Rose, Philip J. & Alison Simmonds (1996) "F-pattern variability in Disguise and Over the Telephone - Comparisons for Forensic Speaker Identification" In Paul McCormack & Alison Russell (eds.) *Proc. 6th Aust. Internat. Conf. Speech Science and Technology, Aust. Speech Sci. & Tech. Assoc*. 121-126.

Schegloff, E.A. (1986) "The routine as achievement." *Human Studies* **9**, 111-152.

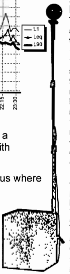
Schegloff, E.A. (1979) "Identification and Recognition in Telephone Conversation Openings." In G. Psathas (ed.), *Everyday Language: Studies in Ethnomethodology*. New York: Irvington.

Schegloff, E.A. & H. Sacks. (1984) "Opening Up Closings." In Baugh & J Scherzer (eds.), *Language in Use: Readings in Sociolinguistics*. New Jersey: Prentice-Hall Inc.

Wolf, Jared J. (1972) "Efficient Acoustic Parameters for Speaker Recognition". *J. Acoust. Soc. Am.*, **51**(6), 2044-2056.



- Type 1 or Type 2 noise loggers available for hire.
- Expecting noise levels below 35dB(A)? Don't take a chance with a Type 2 logger, use an iM3 logger with Type 1 accuracy down to 25dB(A).
- Full hire and placement service available, you tell us where and we do the rest - call for a free quote.
- Free data review at the end of each job, including hardcopy of daily noise level graphs.
- Central metropolitan location in Ryde
- Store Leq,0.1s for over a week to capture aircraft flyovers, train passing events etc.



Sales

Infobyte

The Infobyte iM3 Type 1 precision noise monitor is a unique instrument designed specifically for the task of long term noise monitoring in hot climates. *Compare these features ...*

16MB of nonvolatile storage providing secure space for over a year of data.
95dB dynamic range allows the use of a single range setting minimising setup errors.
Peak Overload Detector to ensure Leq results are reliable as required by AS1259.2-1990.
15dB(A) microphone self noise allows accurate results down to 25dB(A).
+65ºC internal temperature rating combined with a custom engineered environmental case allows operation in hot Australian conditions.
iM3 software allows virtually any Leq or Ln to be calculated from the binary survey data and produces report quality graphs at the click of a button.

MICROTECH GEFELL

The Microtech Gefell range of microphones and accessories are sold by Infobyte. These high quality German microphones are cost effective replacements on most name brand sound level meters and analysers. Contact us for a price and specification for your application.

Call Geoff Veale on 9807-8786 to make your booking or enquiry

Infobyte Pty Ltd 19 Curtis Street, Ryde, NSW 2112
Ph. +61-2-9807-8786 E-mail: gveale@f1.net.au http://www.f1.net.au/users/infobyte