STATISTICAL QUANTIIFICATION OF DIFFERENTIAL VOWEL COMPARABILITY IN FORENSIC PHONETIC SAMPLES

Yuko Kinoshita
Phonetics Laboratory, Linguistics Program,
The Australian National University

John Maindonald
Statistical Consulting Unit of Graduate School,
The Australian National University

ABSTRACT In the phonetic comparison of forensic samples, vowels in different words often need to be compared. This paper discusses to what extent vowels embedded in different words are in fact comparable. The experiment was carried out on natural speech data to simulate forensically realistic conditions, and 11 male native Japanese speakers participated as informants. Multi-level analysis of variance was performed on the F2 of three vowels /a/, /i/, and /e/. The experiment shows that, although the phonological environment, namely the nasality of a preceding consonant, affects the F2 of these vowels, the magnitude of the effect can be discounted. What is shown to be important, however, is the identity of the vowel, and comparisons with /a/ in different phonological environments are strongly disfavoured.

INTRODUCTION

This study was motivated by the importance of the comparability between reference and test samples (incriminating speech and suspect's speech) in forensic phonetics. Künzel (1995:68) says "...non-forensic SR (speaker recognition) is not much of a problem either scientifically or technologically." Forensic speaker identification is, however, far from "not much of problem" even now, five years on from Künzel's statement. The main reason for this lies in the amount of control one can have over the recordings. In the development of non-forensic systems, such as security or automated banking, factors like amount, duration, content, and recording quality are sufficiently controlled, as it is possible to predetermine the structure of both reference and test samples.

In a typical forensic situation, on the other hand, it is not possible to have such a control over recordings, especially over incriminating recordings. As a result, the comparison between incriminating and suspect samples becomes far more complicated than the comparison between reference and test samples in the non-forensic situation. An incriminating recording and a suspect's recording are likely to be recorded with different recording equipment, resulting in different recording qualities. For instance, incriminating speech may have been intercepted from a telephone line, whereas a suspect's speech was recorded directly, but poorly, during police interrogation. The content, duration and formality of speech may also differ between two sets of recordings. This lack of control over data may result in a situation where criminal and suspect recordings do not include optimal words for comparison. Comparability of two vowels is phonologically at its best when the words in which the vowels are embedded are segmentally and prosodically identical, and a segment that occurs in the same position in repeats of the same word is optimal to determine the nature of within- and between-speaker variance (Rose 1998:4). Having the same words in both incriminating and suspect samples enables the comparison of the formant contour, which has been reported to yield a better identification rate than point measurements of formants of individual segments (Ingram et al. 1996, Greisbach et al. 1995). It is well known that segments are coarticulated with the adjacent segments in speech (Farnetani 1997). Therefore, when the same words are compared, the influence of coarticulation would not interfere.

In forensic situations, however, there may not be a sufficient number of the same suitable words or phrases. In such a situation, it is assumed that comparison of single vowels in the same prosodic context (e.g. stress) but different words is the next best thing. This is an attractive alternative, in the sense that it is more or less guaranteed that two sets of recordings contain at least some of the same vowels in comparable environments. One problem here however is that it is not known how and how much the phonological differences in surrounding segments can affect the target vowels. It is well known that phonologically the same vowel may not be phonetically equivalent (Broad 1976). Statistical verification of the assumption is crucial. This study thus performs statistical analyses on the comparability of vowels in different words using multi-levelled analysis of variance.

DATA

The informants for this study were 11 male native speakers of Japanese. In the recording sessions, they performed a set of tasks which were designed to elicit natural speech. In these tasks, the informants were provided with a map and an information sheet on 4 people. The map contained 3 bus routes and names of shops and buildings. The information sheet consisted of 4 people's jobs, personalities, and favourite foods. The informants answered questions such as "Where does the route A bus stop?" or "What kind of job does person A do?," referring to the given material. The map and the information sheet were designed to contain examples of all 5 Japanese short vowel phonemes occurring on the pitch-accented syllable, 5 times each. The linguistic contents of the corpus are summarised in table 1.

/a/	*hanaya* 'florist', *panya* 'bakery', *sakata* '(name)', *sobaya* 'noodle shop', *panyano* 'of bakery'
/i/	*jinja* 'shrine', *jibika* 'otolaryngology', *kobijutsu* 'antique', *sushiya* 'sushi bar', *sanwaginkoo* 'Sanwa bank'
/u/	*nikuya* 'butcher', *tokushima* '(name)', *kaguten* 'furniture shop', *doobutsuen* 'zoo', *kurita* '(name)'
/e/	*Nemoto* '(name)', *terebi* 'TV', *kitadeguchi* 'north exit', *kitadeguchi* 'north exit', *minamideguchi* 'south exit'
/o/	*Kinoshita* '(name)', *toshokan* 'library', *hoteru* 'hotel', *honya* 'book shop', *toposu* '(name of shop)'

Table 1. Words included in the corpus of natural speech. The accented segments are underlined.

Two recording sessions were held for each speaker, separated by two weeks. The sequence of tasks was performed twice in each session. The recording was carried out in the studio of the Phonetics laboratory at ANU.

The recordings were then digitised at 16 kHz and analysed with CSL. F1 to F4 of the short accented vowels were sampled at the middle point of the vowel duration. As both onset and offset of the vowels' F-pattern are expected to be directly influenced by the adjacent vowels, mid point was assumed to be least affected by the adjacent segments (Ladefoged and Maddieson, 1996:287) and, therefore, to be the most stable point across different words. Four formants (F1 to F4) of 5 different vowels were measured, so the data involve 20 vowel / formant combinations. The current study focuses on the F2 of three vowels /a/, /i/, and /e/. F2 of /i/ and /e/ were chosen as they are the strongest parameters in discrimination of speakers (Kinoshita, in preparation), /a/ was chosen for the purpose of comparison. Each of those vowel/ formant combinations consists of 20 samples (5 words * 2 repeats * 2 recording sessions).

STATISTICS

Multi-Level Analysis and Nature of Differences in Acoustic Data Four main factors closely related to the realistic forensic situation are assumed to be contributing to the acoustic differences between samples of each vowel / formant combination in the data for this study. Those factors are: 1) between speaker variation, 2) between words variation, 3) between recording session variation, and 4) between repeats variation within a recording session. Such extraneous factors as state of health or state of mind will of course influence variation between and perhaps within recording sessions. The first variable, between-speaker variation, is obviously the prime interest in forensic speaker identification. The second variable, between words variation, is relevant when the same vowels in different words are compared. The third and the fourth variables are paraphrased as non-contemporaneous and contemporaneous variations respectively. An incriminating recording and a suspect's recording in forensic speaker identification are always non-contemporaneous, and if each of those recordings includes several samples to be compared, contemporaneous variation also needs to be taken into consideration.

Multi-level ANOVA examines multi-level structure in the variation (Goldstein 1987), and it is due to Fisher (1935). As used by other workers, ANOVA was often limited to a single level of variation. Our application requires multi-level analysis of variance to gain an insight into the patterns of variation. The results are in table 2. The mean squares represent how much variation at the respective level contributes to the acoustic differences between samples. 'Df', 'SSQ', 'MSQ' in the top row indicate degree of freedom, sum of squares, and mean squares respectively. The first two shaded rows, "word" and "speaker," show the sums of squares and mean squares for overall between-word and between-speaker variation. For generalisation to a wider population of speakers from whom the 11 speakers have been drawn, one would treat speaker as a random effect. The rest shows the hierarchical structure of random effects in residual when ANOVA was carried out on "speaker" factor. 'A %in% (B/C)' indicates the mean square of variable A in the interaction with the variables B and C. Thus, for instance, the mean square of 'session %in% (speaker / word)' shows the mean square which was obtained when two sessions for each speaker's each word were compared separately. Also, the levels other than 'word' are numbered from 1 to 4, in order to facilitate the discussion later. This 'word' is a fixed effect across speakers.

	Factors	Df	SSQ	MSQ
/a/	word	4	2914719	728680
	1. speaker	10	1958037	195804
	2. word %in% (speaker)	40	554672	13867
	3. session %in% (speaker / word)	55	442427	8044
	4. repeat %in% (speaker / word / session)	95	466509	4911
/i/	word	4	1026898	256724
	1. speaker	10	6175554	617555
	2. word %in% (speaker)	40	810637	20266
	3. session %in% (speaker / word)	55	1147992	20873
	4. repeat %in% (speaker / word / session)	86	1448212	16840
/e/	word	4	761301	190325
	1. speaker	10	4029650	402965
	2. word %in% (speaker)	40	443582	11090
	3. session %in% (speaker / word)	55	524808	9542
	4. repeat %in% (speaker / word / session)	95	407029	4285

Table 2. Results of multi-level analysis of variance.

Table 2 shows that the mean squares for speaker effect of the vowels /i/ and /e/ are larger than the other mean squares by far. For /a/, on the other hand, the mean square for word is considerably larger than other mean squares. For /i/ and /e/, the speaker effect contributes more to the acoustic differences between samples. It should also be noted that, although the mean squares of word effect are smaller than that of speaker effect on /i/ and /e/, the word effect on these vowels seems reasonably large (almost half for /e/ and one-third for /i/).

Hierarchical Structure of Random Effects This section discusses the hierarchical random effects structure. The analysis in the previous section has demonstrated that the fixed word effects can affect acoustics of vowels. In this section, the random effects are examined. Analysis of the speaker mean square shows primarily how well speakers are distinguished from each other. In reality, however, the variables such as words, sessions and repeats are also contributing to the acoustic differences of samples. The interaction of multiple variables (variables 2-4 in Table 2) portrays the hierarchical structure of variances. Comparing variances 1 to 2, 2 to 3, and 3 to 4 gives insight into the source of variation.

Table 3 summarises the comparison of the mean squares of variables 1 to 4. The comparison is expressed in the form of ratios of mean squares. The '95% CI' in the table means the 95% confidence interval for the ratio. The larger the ratio is, the more substantial the effect of the first component of the two factors. The columns of 'Evaluated effect' show which factors were evaluated by the ratio of mean square, and the columns 'Ratio calculation' show which two factors were involved in the ratio calculation.

A ratio = 1 means the first factor does not add further to the difference which is explained by the second variance. Thus for the variable combinations whose '95% CI' include 1, the numerator factor may add nothing to the variance contributed by the denominator factor. Ratios that are significant at the 5% level are marked by shading. For easier interpretation, the bar chart of the mean square ratio of the random effect is also presented (Figure 1).

Evaluated effect	Ratio calculation	/a/		/i/		/e/	
		MS ratio	95% CI	MS ratio	95% CI	MS ratio	95% CI
Speaker	sp/wd	14.1	5.9 - 46	30.5	12.8 - 99.2	36.3	15.2 - 118.3
Word	wd/sess	1.7	NS	1.0	NS	1.2	NS
Session	sess/rep	1.6	1.04 - 2.7	1.2	0.8 - 2.0	2.2	1.4 - 3.6

Table 3. Summary of the comparisons of mean squares and their p-values. The labels "sp", "wd", "sess", and "rep" respectively represent the variables 1 to 4 in Table 2.

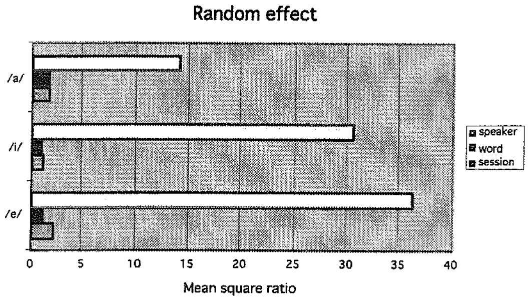

Figure 1. Mean square ratios from the ANOVA table.

Table 3 shows that difference in speaker factor significantly affects the acoustics for all three vowels. For other variables, there are vowel-to-vowel differences. Random word effect was not found significant in any vowels. Session-specific influence is observed for /a/ and /e/ vowels, but not for /i/ vowel. The results suggest that the acoustic data of /a/ contain the least speaker-originated variation.

In sum, these statistics have shown that there were consistent effects of word differences on acoustic data. The size of effect, however, varied vowel to vowel. The /a/ vowel had the largest effect for word by far, followed by the /e/ vowel, and the /i/ vowel was found to be affected least. The fact that word effect had larger impact on the acoustic values than speaker effect for the /a/ vowel suggests that the comparison of an /a/ vowel embedded in different context is not of much use in forensic speaker identification, unless there is an adjustment for the context. The random effect of word was small for all vowels.

WORD EFFECT AND LINGUISTIC INTERPRETATION

The previous sections demonstrated that the word differences affect /a/ vowels substantially, with the /i/ and /e/ vowels less affected. This vowel-to-vowel difference may have been caused by the vowel specific characteristics. The /a/ vowel may not be an ideal segment for forensic speaker identification. It is also possible, however, that the phonological environment of vowels caused the difference. The different articulatory distances between a target vowel and adjacent segments could cause the different degree of coarticulation. These distances may have been larger in one vowel than the others in this study. This section investigates the relationship between the adjacent segments and the results found in the previous sections.

Table 4 presents the comparisons between words, using word 1 as the baseline. Except for 'word 1,' the column in the 'comparison' indicates which words were compared against word 1. The phonological environments of the target vowels are shown in the column headed 'Envt'. The column headed 'Diff.' shows the mean F2 of the vowel for word 1 and, for word 2 to 5, the difference between word (the unit is

Hz). The standard errors for tokens 2-5 are really the standard error of the values shown in the 'Hz' column. 't' and 'p-val' show the results of t-tests for words 2-5 against word1.

	Comparison	Envt	Diff.	Std.Er	Df	t	p-val
/a/	word 1	n _ j	1431	30.0	-	-	-
	word 2 (vs 1)	p _ n	-134	13.0	40	-10.3	<.0001
	word 3 (vs 1)	s _k	-23	7.5	40	-3.1	0.0032
	word 4 (vs 1)	b_ j	-44	5.3	40	-8.4	<.0001
	word 5 (vs 1)	p _n	-21	4.1	40	-5.1	<.0001
/i/	word 1	ʤ _ n	2160	56.0	-	-	-
	word 2 (vs 1)	ʤ _ b	-20	16.3	40	-1.2	<.0001
	word 3 (vs 1)	b_ ʤ	-37	9.2	40	-4.0	0.0003
	word 4 (vs 1)	ʃ_ j	0	6.5	40	0.0	0.9760
	word 5 (vs 1)	g /ŋ _n	29	5.2	40	5.7	<.0001
/e/	word 1	n _ m	1919	45.6	-	-	-
	word 2 (vs 1)	t _ r	-11	23.5	40	-0.5	0.6511
	word 3 (vs 1)	d _ g/ŋ	109	23.5	40	4.6	<.0001
	word 4 (vs 1)	d _ g/ŋ	73	23.5	40	3.1	0.0034
	word 5 (vs 1)	d _ g/ŋ	142	23.5	40	6.0	<.0001

Table 4. Values and standard errors for tokens 2-5 compared to word 1.

If the vowel-to-vowel differences in the word effects are attributed to coarticulation, the coarticulatory effects are expected to be proportional to the distance between the articulatory position of the target vowels and their adjacent segments. A comparison between word 1 and other words, however, does not show this pattern. For instance, words 2 and 5 for /a/ have exactly the same adjacent segments, and yet their difference in relation to word 1 is substantial (-134 for word 2, -21 for word 5). Preceding consonants of words 2, 4 and 5 of /a/, words 1 and 2 of /i/ are the same, words 1 and 4 of /a/, word 1 and 5 of /i/, words 3, 4 and 5 of /e/ all have the same articulatory position for following consonant. The values shown in 'Hz' columns of these pairs (or three) of words, however, vary. The place of articulation for neither preceding nor following segments, thus, seems to systematically affect the target vowels.

We can not conclude, however, that the adjacent segments have no effect. For /a/ and /e/, the words 2-5 have noticeably uneven distributions in relation to word 1, whereas such an obvious tendency was not found for /i/. All words for /a/ are constantly smaller than the baseline. For /e/, three out of four words are larger than baseline. Even though there was one exception of word 2, the difference between this word and word 1 was considerably smaller than the difference between other words and word 1. The phonological environments of word 1 are /a/ - 'n _ j', /i/ - 'ʤ _ n', and /e/ - 'n _ m'. Preceding the target vowels, /a/ and /e/ have a nasal consonant [n], whereas /i/ has an oral consonant, [ʤ]. Both /a/ and /e/ in word 1 are thus most likely to be nasalised. It is known that nasalisation changes the formant structure of a vowel considerably, as when a vowel is nasalised, it comes to have nasal formants and antiformants in addition to oral formants (Fujimura and Erickson 1997:81-3, Johnson 1997:158). The result of this study implies that nasalisation raises F2 of /a/ and lowers that of /e/. The difference in the effect for these two vowels may be due to the difference in the initial tongue location for the vowels. As we know, /a/ and /i/ differ in backness, and the height of F2 correlates with the backness of the body of the tongue.

In the previous section, it was shown that word difference affected /a/ most, then /e/, and /i/ was influenced least. The fact that a preceding nasal consonant had a constant effect on vowel formants suggests the possibility for the large mean square for the random word effect of /a/ to be attributed to this preceding nasal consonant at least partly. If the nasality in the phonological environment is the sole cause for the difference in the size of the word effect, however, the large difference between /a/ and /e/ cannot be explained. The /a/ vowel thus seems to be more susceptible to the surrounding phonological environment.

SUMMARY AND IMPLICATIONS FOR FORENSIC PHONETIC COMPARISON

This study has demonstrated that differences in words in which vowels are embedded could have effect on the formant patterns, and the size of effect differs vowel-to-vowel. /i/ was the least affected among three vowels investigated in this study, and /a/ was influenced the most significantly. In fact, the word effect on /a/ turned out to be even larger than speaker effect, and this clearly suggests that the use of the /a/ vowel in different words for speaker identification is inappropriate, unless it is possible to adjust for the phonological environment.

As the reason for the vowel-to-vowel difference in the size of word effect, the distance between articulatory positions of target vowels and adjacent segments does not provide an explanation. Instead, the results for /a/ and /e/ revealed that a nasal consonant seems to affect the following vowels fairly systematically. Why the nasal affected /a/ more is, however, still unclear. Nevertheless, the results of this study suggest that vowels preceded by nasal consonants (and therefore quite possibly being acoustically nasalised) should not be compared with oral vowels.

Further, it also has to be noted that comparison of vowels occurring in the same word and vowels occurring in different words is not equivalent, in terms of their strength and reliability as evidence. Thus caution is clearly in order in forensic phonetic speaker identification involving cases like this.

ACKNOLEDGEMENT I am very grateful for the reviewers who provided me with valuable comments.

REFERENCES

Farnetani E. (1997) "Coarticulation and connected speech processes", In W. Hardcastle and J.M.D. Laver (eds.): The Handbook in Phonetic Sciences, 371-404, (Blackwell: Oxford).

Fisher, R.A. (1935/1960) The Design of Experiments, 7th edition, (Oliver & Boyd: Edinburgh)

Fujimura O., and Erickson D. (1997) "Acoustic Phonetics", In W. Hardcastle, J. M. D. (ed.): The Handbook of Phonetic Sciences, 65-115, (Blackwell: Oxford).

Greisbach R., Esser O., and Weinstock C. (1995) "Speaker identification by formant contours", Studies in Forensic Phonetics BEIPHOL 64, 49-55.

Goldstein, H. (1987) Multilevel Models in Educational and Social Research, (Charles Griffin & Company LTD and Oxford University Press: London, New York).

Ingram, J., Prandolini, R., Ong, S. (1996) "Formant trajectories as indices of phonetic variation for speaker identification", Forensic Linguistics 3 (1), 129-145

Johnson, K. (1997) Acoustic and Auditory Phonetics. (Blackwell: Oxford).

Kinoshita, Y. (in preparation) Japanese forensic speaker identification based on formants, PhD dissertation, The Australian National University

Künzel, H. J. (1995) "Field procedures in forensic speaker recognition", In J.W. Lewis (ed.): Studies in General and English Phonetics. Essays in Honour of J. D. O'Connor, 68-84, (Routledge: London).

Ladefoged P, and Maddieson I (1996) Sounds of the World's Languages, (Blackwell: Oxford).

Rose PJ (1999) "Differences and distinguishability in the acoustic characteristics of *Hello* in voices of similar-sounding speakers", Australian Review of Applied Linguistics 21, 1-42.