JAPANESE FORENSIC PHONETICS: NON-CONTEMPORANEOUS WITHIN-SPEAKER VARIATION IN NATURAL AND READ-OUT SPEECH

Yuko Kinoshita

Phonetics Laboratory, Department of Linguistics, (Faculty of Arts),
and Japan Centre (Faculty of Asian studies), ANU

ABSTRACT

This paper aims to explore non-contemporaneous within-speaker variation of a Japanese male speaker, focusing on the difference between speech styles, viz.. natural speech and read-out speech. Recordings under forensic conditions are mostly of natural speech. A suspect's recording to be compared are, however, sometimes read-out speech, but not natural speech, in order to obtain the similar phonological conditions to the original criminal speech. This paper aims to examine the validity of such a procedure.

1. INTRODUCTION

This paper aims to answer the question "how does the difference in speech styles affect within-speaker non-contemporaneous variation (WNCV hereafter)?" The recordings from crime scenes are mostly spontaneous speech, but the reference speech obtained from suspects not always so. Suspects may be requested to read out certain texts (e.g. R v.s. Crowther (1992) Victorian County Court DPP W1189). Or, suspects may be requested to repeat a sentence uttered by the examiner or a police officer (Nolan 1983:203). These methods are all to record reference speech with similar phonological conditions to the recordings from the crime scene, but are these valid procedures? Nolan (1983:203) points out the problem of the repeating method saying that "[w]ithout experimental evidence there is no reason to believe that the gain in contextual similarity between the recordings through avoiding the formal context of reading the criminal text will outweigh the problems of convergence, and possibly imitation..."

This study examines the validity of comparison of two different speech styles, focusing on the difference between natural and read out speech. In reality, it is possible to have a readout criminal speech in particular situations such as the criminal reads out prepared material to make a phone call in a kidnapping case. In this study, however, the discussion of WNCV is limited to the more common case of non-contemporaneous variance between natural speech in session 1 and 2, and between natural speech in session 1 and read out in session 2.

It has to be noted that the question above contains two essential factors for the discussion of speaker identification; within-speaker variation and non-contemporaneous data. Why are they so important?

Different speakers differ in their acoustic output, but so does the same speaker. Past research on speaker identification is based on the assumption that between-speaker variations are always greater than within-speaker variations (Tosi, O., H. Oyer, W., Lashbrook, C. Pedrey, J. Nicol, E. Nash, 1972) the investigation of within-speaker variation is indispensable for the discussion on speaker identification. The significance of the research on non-contemporaneous data is also indisputable, since the data to be compared for forensic purposes are always non-contemporaneous.

Thus the experiment in this study was carefully designed to take these two factors into consideration.

2. PROCEDURE

The informant for this paper is a 23 years old male Japanese speaker who is an exchange student at ANU. He is from Tottori prefecture, west part of Japan, but he has spent 10 months in Australia and, before that, spent four years in Yokohama, where people supposedly speak in standard (Tokyo style) accent pattern. His regional accent was not salient enough to be noted by many of his friends, although it is still discernible to a linguist.

Recording was carried out at the Phonetics laboratory at ANU. Two sets of data, separated by two weeks, were obtained from this informant. Exactly the same process was followed in both of the recording sessions.

Two types of recording were made for this study; natural speech and card reading. Tasks were carefully designed for the elicitation of natural speech. In these tasks, the informant was provided with a map and an information sheet on 4 people. The map contains 3 bus roots and names of shops and buildings. The information sheet consists of 4 people's jobs, personality, and favourite foods. The informant was asked questions such as "Where does the root A bus stop?" or "What kind of job is person A doing?," and he had to answer those questions referring to the given materials. The map and information sheet were designed to contain examples of all 5 Japanese short vowel phonemes occurring on the pitch accented syllable, 5 times each. Although a segment that occurs in the same position in repeats of the same word is ideal to determine the nature of within- and between-speaker variance (Rose, 1998: MS4) it is not realistic to expect to have such recordings in forensic situations. Thus, this study deals with the same vowels in the same pitch accent environment, but not necessary the same word. The corpus is summarised in a table 1.

Recording of card readings followed the task above. All the test words consist of two syllables, accented on the first syllable. Two types of test words were included in the cards. One was V̲CV sequence words, and the other was CV̲CV sequence words. Vs underlined are the measured

segments. These two structures were examined separately to see how the phonological conditions affect the target segments. If the difference between VCV and CVCV structures turns out to be a large one, it suggest that what we presume comparable, as they are the same segments, might not be so in reality. The vowels measured were the 5 short vowels, /a/, /i/, /u/, /e/, and /o/, which are the same as the target segments in natural speech. Each word was repeated 5 times. The words included in the corpus are as follows in table 2.

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

Table 1: Words included in the corpus for natural speech. The accented segments are underlined. 'kitadeguchi' for vowel /e/ was uttered twice.

	VCV	CVCV
a	aki (autumn)	shako(garage)
i	iki (breath)	shiki (time of death)
u	uki (rainy season)	shuki (memorandum)
e	eki(station)	seki (seat)
o	oki (name of place)	shoki (the early stage)

Table 2: Words included in the card reading task.

The recordings were digitised at 16 kHz and analysed with CSL. F-pattern and F0 for accented vowels were measured. F-pattern was sampled at 3 points; 25%, 50%, and 75% of the vowel duration, and F0 was sampled at the 50% point of the vowel duration.

For the discussion on WNCV, the values "token X of the natural speech in session 1 subtracted by token Y of natural speech / read out in session 2 were calculated. This indicates the size of the difference between two sessions and between different speech styles. This value will be referred as 'difference values' hereafter.

Devoicing of high vowels, /i/ and /u/, should also be mentioned. Devoicing is a well known phonological phenomenon in Japanese - these high vowels are devoiced when they do not have any voiced segments adjacent to them. Normally this devoicing is avoided when the segments carry the pitch accent. This informant, however, devoiced all /i/ and /u/ in CVCV sequences, consequently the data for these particular words are missing in this experiment. This suggests that the occurrence of devoicing may be idiosyncratic to a certain extent, and thus may be a potential cue to speaker identity. this requires further investigation.

3. RESULTS

3.1 F0

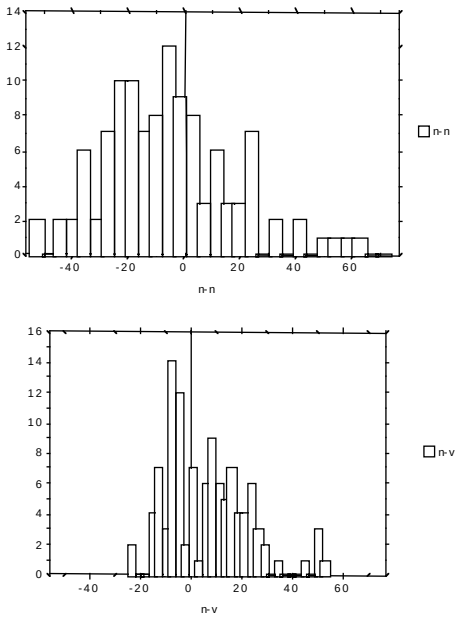
First of all, three kinds of difference values, 'natural 1 - natural 2', 'natural 1 - VCV 2', 'natural 1 - CVCV 2', were calculated. The calculation was done for each vowel separately, and then compared with ANOVA (or t-test for /i/ and /u/, since they do not have data for CVCV). The F ratio and mean differences are summarised in table 3. n-n, n-v, and n-c represent natural 1 - natural 2, natural 1 - VCV 2, natural 1 - CVCV 2 respectively. Table 3 shows, for example, that F-ratio between natural - natural and natural - VCV for vowel /a/ was 9.4.

'n-n vs n-v' seems to have the largest difference. Vowel articulations does not affect on the mean difference of difference values.

	n-n vs n-v		n-n vs n-c		n-v vs n-c	
	F	m. diff.	F	m. diff.	F	m. diff.
a	9.4	-20	2.6	-10.6	2.1	9.4
i	.0278	7.6				
u	.0001	-16.1				
e	2	-6.8	.0003	0.8	2.6	7.6
o	13.3	-14.2	16.5	-15.8	0.2	-1.6

Table 3 Result of ANOVA and t-test. The figures in bold are those which reached the level of significance, 95%.

The difference in recording style appears to have some affect on the magnitude of the difference between two recording sessions. The following figures (figure 1) are the histograms of the difference values for each recording style. Each figure contains the difference values of all five vowels.



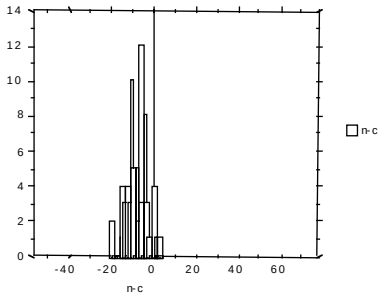


Figure 1: Frequency distribution of different values for each recording styles.

The distribution of difference values are for natural-natural; -55 to 73, for read-outs, natural-vcv; -25 to 54, natural-cvcv; -21 to 3. Natural-natural has a clearly wider distribution than the other two. In the citation of natural speech, the speaker was completely free in the choice of intonation, which directly affect F0 values. This was not the case for the card reading, however. Although there was no particular instruction for reading, the utterance is expected to be acoustically more stable across five tokens, since in card reading, all five tokens for each vowel was the same word, opposed to the natural speech which consisted 5 different words for each vowel. Also in the reading task, there is less variance. For instance, it is not likely that speaker tries to convey any particular communicative intent nor emotional state.

The difference between two read-outs is also noticeable. This might be simply attributed to the fact that CVCV contains only three vowels, missing /i/ and /u/, whereas VCV contains all five vowels. Those high vowels have larger standard deviations (see table 4), ie. larger within-speaker variation.

		a	i	u	e	o	all
N 1	mean	111.2	107.3	128.4	99.6	103.4	110
	std.	16.3	15.1	25.2	10.4	8.8	15.16
2	mean	121.8	121	113	105.2	125	117.2
	std.	11.2	21.3	16.6	12.3	5.1	13.31
V 1	mean	102	108.9	101.6	98.4	110.8	104.3
	std.	4.1	15	3.4	4.6	8.2	7.037
2	mean	108.8	109.2	110.8	104.4	105.8	107.8
	std.	3.5	6.8	6.3	3.4	3.1	4.621
C 1	mean	99.4	0	0	98.8	101.4	99.87
	std.	2.1	0	0	3	2.2	2.433
2	mean	111.2	0	0	106	109.2	108.8
	std.	5.3	0	0	3.5	3.6	4.16
all 1	mean	104.4	108.2	115	98.9	105.2	106.3
	std.	10.8	14.1	22.1	6.3	7.8	12.2
2	mean	113.9	107.2	103.6	105.2	113.3	108.7
	std.	9	28.9	24.2	7.1	9.4	15.74

Table 4: Mean and standard deviation of raw F0

Thus the missing vowels serves an explanation for this difference between natural-VCV and natural-CVCV, although it should be noted that their differences in the phonological environment also might be a part of cause. In VCV structure, tongue as an articulator has more freedom than in CVCV structure, where the tongue movement is more constrained by the preceding consonant.

3.2 F-pattern

The measurements here were limited under 4000 kHz to match the forensically realistic situations where the recording quality is often too poor to pick up those higher frequencies. Much of the natural speech in session 1 does not have values for F4, as some of them are too weak to pick up, and some of them are over 4000 kHz.

ANOVA was carried out to examine the WNCV depending on the recording styles. /i/ and /u/ for F1-3 (missing CVCV) have only two groups to compare, so t-test was employed for those instead. Table 5 are The summary of the Scheffe-F test (and p values for t-test) and mean difference. The figures in lower column are mean differences. nn, nv, and nc indicates the natural-natural, natural-VCV, and natural-CVCV. nn/nv means that natural-natural was compared to natural-VCV. Table 5 shows, for example, that F-ratio between natural-natural and natural-VCV was 7.5 for F1 of vowel /a/.

		a	i	u	e	o
F 1	nn/nv	7.5	.0001	.0001	.4.5	1.5
		96.9	54.5	-42.9	-53.4	-47.3
	nn/nc	4.7			3.2	0.3
		-76.5			-44.6	17.2
	nv/nc	24.2			0.1	3.4
		173.4			8.8	64.5
F 2	nn/nv	0.9	.0002	.0001	35.4	4.6
		-45.7	92.5	156.6	344.8	-147.5
	nn/nc	5.2	-	-	21.3	10.4
		-107.2			256.8	220.5
	nv/nc	1.7	-	-	2.5	28.8
		-61.5			-87.9	367.9
F 3	nn/nv	1.2	.0001	.049	8.2	4.4
		-94.3	243.6	59.8	175.9	111.3
	nn/nc	15.5			1.3	0.1
		-333.3			70.9	-17.8
	nv/nc	8			2.9	6
		-238.9			-105	-129.1
F 4	nn/nv	2.7	.0001	.438	5.2	3.7
		-184.8	224	25.5	-183.9	-111.1
	nn/nc	4.9			0.5	5.5
		-243.7			-49.2	-135.8
	nv/nc	0.3	-	-	2.9	0.2
		-58.9			134.7	-24.7

Table 5: The summary of the Scheffe -F test (and p values for t-test) and mean difference. Figures in bold indicate that those reached at the level of significance; 95%.

The mean difference appears to be larger when the comparison involves natural-natural values. This agrees with the result of F0. Also, it seems that /u/ has smaller mean difference.

The distribution of difference values for F3 are presented in the figures below in order to provide a better view of the variance between three speech styles.

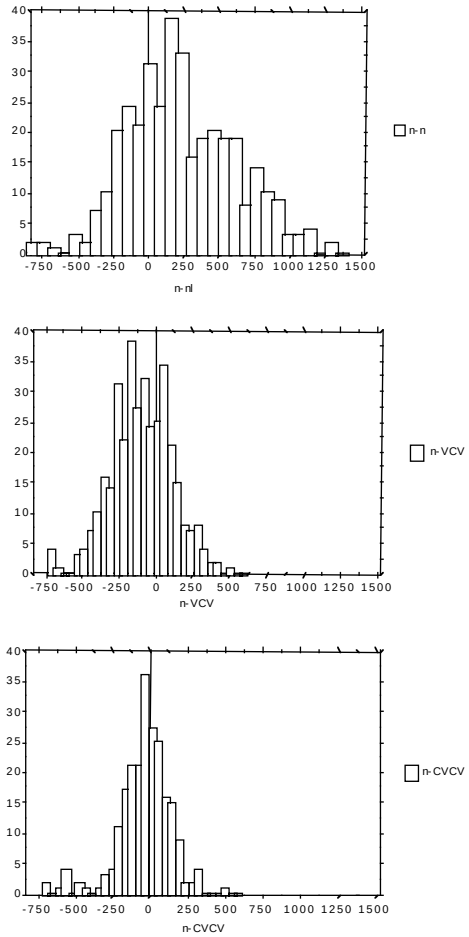


Figure 2: The distribution of difference values for F3.

The difference values range approximately -500 to 1000 for natural speech, -500 to 350 for vcv, and -250 to 250 for cvcv. It is fairly clear that the speech styles make considerable differences in their distribution of the different values. In other words, a single person can exhibit different degree of within-speaker variation when the natural speech and the read-out are compared.

4. CONCLUSION

The difference in speech styles affect the range of non-contemporaneous within-speaker difference greatly. It was found that there is a wider within-speaker variation between two natural speech recordings than between two different styles viz. natural and readout speech. In other words, for the natural and readout speech combination, the threshold for exclusion is lower. Thus the identification with the readout reference speech is valid procedure, at least in the investigation of the segments.

It was also revealed that the difference in the phonological structure of the given words can also affect the size of within-speaker variation.

As an additional finding, the devoicing should be noted. The occurrence and/or frequency of the devoicing is a possible parameter in the discrimination of individual. This parameter, supposedly peculiar to Japanese, needs to be investigated with more speakers from various region since regional dialects affect on this parameter.

5. REFERENCE

1. Nolan, F., The Phonetic Bases of Speaker Recognition, Cambridge University Press, 1983

2. Rose, P.J., “Difference and discriminability in the acoustic characteristics of words in voices of similar-sounding speakers - a forensic phonetic investigation”, manuscript, paper submitted to Language and Speech, 1998

3. Tosi, O., H. Oyer, W., Lashbrook, C. Pedrey, J. Nicol, E. Nash, “Experiment on speaker identification”, 1972