Are Nasals Better? Likelihood Ratio-based Forensic Voice Comparison with Segmental Cepstra from Cantonese and Japanese Syllabic/Mora Nasals

Alex Chi Sing Yim [1] *& Phil Rose* [1,2]

[1] Divison of Humanities, Hong Kong University of Science & Technology
[2] School of Language Studies, Australian National University
alexycs@ust.hk, philip.rose@anu.edu.au

Abstract

Two likelihood ratio-based forensic voice comparison experiments are described using long nasal segments from Japanese and Cantonese. The performance is similar for nasals in both languages, with log-likelihood ratio costs of between ca. 0.5 and 0.8. This demonstrates some FVC potential, but with no indication that they are superior to other segments.

Index Terms: Forensic voice comparison, likelihood ratio, mora nasal, syllabic nasal, Cantonese, Japanese, segmental cepstrum.

1. Introduction

Nasals are often thought to be potentially very useful segments in speaker recognition - provided segments are being used, of course - because of probably the most important factor in speaker recognition: a high ratio of between- to within-speaker variance [1, 2]. This results firstly from the potential for considerable between-speaker differences in the usually asymmetrical internal structure and complicated dimensions of the human nasal cavity. Much of its internal volume is taken up by three convoluted bony protrusions extending from the lateral walls of the nasal cavity and partially dividing it into three passages on either side of the often asymmetrical septum. The nasal cavity is also connected bilaterally to several auxiliary cavities, or sinuses, and these, again, are often asymmetrical. The relative rigidity of the nasal cavity structures, on the other hand, contributes to a low within-speaker variation. Together, these anatomical features should yield an advantageous ratio of between- to within-speaker variance in the acoustic features associated with the nasal cavities' acoustic resonances.

But of course nothing is ever remotely totally invariant in speech. The acoustic properties of the sound radiated from the nostrils and throat during a voiced nasal consonant depends not only on the nasal cavity but also on what the other vocal tract structures coupled to them are doing: the acoustics of [m] in *imitate*, *Armagh*, *a merchant*, *you moved*, will all differ a little due to the differing tongue body position during the nasal hold; and in *army*, *Kumar*, the acoustics will actually be changing as the tongue body moves from one vocalic target to another. The area of the velic port, which tunes the nasal cavity resonances, can also be expected to vary depending on adjacent speech sounds, and an [m] said with raised larynx will have different acoustic structure to one said with the larynx low.

Surprisingly, not much seems to have been done on the Forensic Voice Comparison (FVC) potential of nasal consonants within the LR-based paradigm, which is specifically designed to test the strength of evidence that can be expected from a given system that might be used in FVC [3]. The Japanese mora nasal was investigated in [4], but this was early on in the development of the LR-based method – the LRs were not calibrated for example, nor channel-corrected, and no C_{llr}s were reported: we remedy this here. A LR-based study using 20 female Chinese speakers' [m] and [n] from good quality recordings [1] reported poor performance for [n] and better for [m] (C_{llr}s of ca. above 0.75 & 0.5 respectively). It was not clear whether the [n] tokens were syllable-initial or codas; given the phonotactics of Mandarin, the [m] tokens would have to have been syllable-initial.

The aim of this paper is to present two experiments which investigate further the potential for nasal segments in FVC. The first, which is on a slightly larger scale, uses data from the Japanese mora nasal reported in [4], but reanalysed with current methods. The second is a pilot experiment using the syllabic nasal in Cantonese. Both experiments thus make use of nasals with a possibly beneficial greater expected duration than, for example, syllable-initial nasals in English or Chinese. They also use a 'whole-spectrum' segmental-cepstrum approach and, most importantly, use non-contemporaneous data for testing. Our main interest is in any indication that nasals are indeed superior as segments for FVC.

2. Japanese Mora Nasal

2.1. Corpus & Elicitation

The Japanese mora nasal is a phoneme whose occurrence is restricted to the syllable coda. It counts as a separate timing unit and can carry a pitch accent. Its basic, prepausal, allophone is a nasal sonorant ranging between velar [ŋ] and uvular [N], but it also shows place assimilation with following, heterosyllabic, segments, and is often realised without closure as a nasalised vowel [5]. It might be expected to have a fairly long, stable spectrum due to its phonological properties *qua* separate mora and pitch-accent carrier.

Mora nasal tokens were taken from the first 60 speakers of an early Japanese *National Research Institute of Police Science* database, which contains recordings, digitized at 10 kHz with 12 bit quantization, of 297 adult male Japanese from 11 different prefectures around Japan. All speakers were members of the Japanese Police Force and were uncontrolled for age (which ranged from ca. 20 to 50 years). The recordings were made centrally, on the same equipment, from landline telephone calls. Most importantly, two non-contemporaneous recordings were made for each speaker, separated by three to four months. Each recording for each speaker contains about 70-80 seconds net speech comprising two repeats of the set of five Japanese vowels and several single-word and many-word read-out utterances like kodomo wa daijōbu da *the kid's safe*, or bakudan o shikaketa *I've planted a bomb*.

The words and expressions from which the nasal tokens were extracted are given in table 1. As there were two repeats per recording session, there were 14 nasal tokens in each recording.

Table 1. Corpus. N = mora nasal

saN	*three*
*yoN_*	*four*
bakudaN	*bomb*
bakudaN_o shikaketa	*I've planted a bomb*
gyakudaN suruna	*don't try and trace the call*
deNwa	*telephone*
moo ichido deNwa suru	*I'll call you again*

2.2. Processing

Although vocalic centre-frequencies remain the feature of choice in traditional LR-based FVC, it was decided instead to adopt a 'segmental-cepstral' approach [7], and quantify the nasal spectra cepstrally. This was because it was anticipated that formant estimation and extraction might be problematic with nasals, not the least because it could not be assumed that a corresponding set of poles would be extracted for each subject, and it is not clear how to handle the sort of discrete situation, where for example one recording has an extra pole, in a LR framework. Moreover, it has been shown that, not surprisingly, a whole-of-spectrum model delivers stronger LRs *ceteris paribus* than one based just on formant centre-frequencies [4]. An additional possible argument, that formants are usually extracted from all-pole models, and nasals (at least bilabials) have zeros, does not apply as we used all-pole LPC CCs.

Using dynamic programming, the mora nasal segment was located, and 12th order LPC CCs extracted from as long a section as possible of its quasi-stationary wave-form. The first 20 speakers were selected for training, the remaining 40 for testing. Using the kernel-density version of the generative multivariate LR model in [6], LRs for same-speaker and different-speaker comparisons were estimated for the 40 test speakers against the between- and within-covariance data from the 20 speaker reference sample. This formula estimates multivariate LRs (MVLRs) taking into account any correlation between variables within a segment. (Although the variables in this case are CCs, which are orthogonal by definition, there is always the chance that correlation will arise by virtue of the spectral shape of the actual sound being modeled, as was shown for [ɕ] in [4]). With two non-contemporaneous recordings, two independent non-target trials are possible. Only one was tested, thus giving in all 40 target- and 780 non-target trials. In an attempt to at least partially compensate for the inevitable spectral distortion caused by all aspects of the phone transmission, a set of cepstrally-mean subtracted CCs (cms-CCs) was obtained by subtracting each speaker's CCs for a given nasal token from the mean cepstral vector obtained from the speech in their whole repeat, and these were also used to estimate MVLRs. MVLRs from both raw and cms_CCs were then logistic-regressively calibrated.

3. Japanese Mora Nasal Results

Both raw CCs and cms-CCs showed the usual good-discrimination but bad calibration expected from a large number of automatic variables (here, 12 CCs), with subsequent calibration resulting in a significant drop to a C_{llr} below unity, showing that the system was delivering some information. Surprisingly, cepstral mean subtraction did not improve the performance, and the cms-CC data had a C_{llr} worse than that of the raw CCs by ca. 0.15. Fig. 1 shows the Tippett plots for the calibrated MVLRs. As can be seen, the C_{llr} is quite high, at about 0.5, and the greatest strength of evidence obtained for same-speaker comparisons rather low: only about one hundred times more likely. If one wants to regard this as an experiment discriminating same-speaker speech samples from different-speaker speech samples, then the EER is about 16%. Although the C_{llr}s are comparable, these results must be considered a little better than those in [1], which was obtained with clean, as opposed to telephone speech, and fewer speakers. This difference presumably reflects a small contribution from the extra signal duration in the Japanese mora nasal. Clearly, however, although these results and those in [1] show that nasals contain a reasonable amount of speaker-individualizing information, they do not appear to be particularly special in that regard: better C_{llr}s have been reported from other segments under comparable circumstances [7]. Nasal segment LRs could therefore serve as a normal part of a segmentally based FVC system when combined/fused with LRs from other segments.

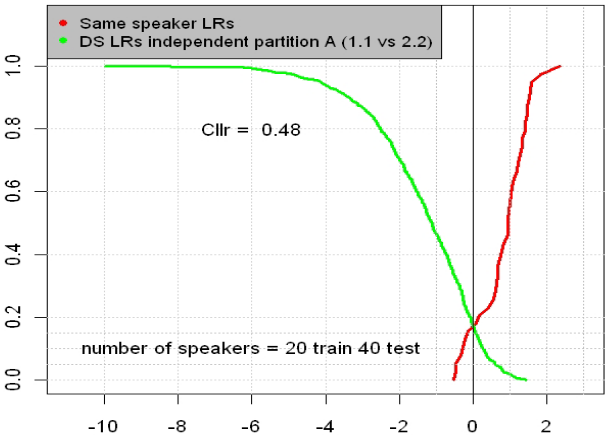

Figure 1: *Tippett plot for Japanese mora nasal. Horizontal axis = $Log_{10}LR$ greater than ...; vertical axis = cumulative proportion of non-target trials ~ 1-cum. prop. of target trials.*

4. Cantonese Syllabic Nasal

Like many Chinese dialects [8], Cantonese has morphemes whose sound shape is just a syllabic nasal consonant. Conservative Cantonese contrasts two places: bilabial and dorso-velar, for example /m̩ ˨˩/ 唔 *negative particle*, /ŋ̩ ˨˩/ *Surname Ng* 吳. One of the few syllabic nasal morphemes is the numeral {five 五}, which makes it fairly easy to elicit natural speech with syllabic nasals. In Hong Kong Cantonese, the vast majority of younger speakers are said to have a bilabial for this morpheme, whereas older speakers may preserve the conservative velar nasal [9].

The suitability for FVC of the Cantonese syllabic nasal was tested with the speech of 20 young Cantonese male speakers from the MTR forensic database. This is a small, quasi map-task database primarily designed to elicit natural but controlled speech for testing likelihood-ratio based approaches to forensic voice comparison. Subjects were given a map of the Hong Kong Mass Transit Railway and were asked various questions about it, for example how to get from station x to station y, or how many stations there are between station a and station b. For this paper we used the response to the second type of question, where the number of stations was five. Speakers were encouraged to give answers repeating the question, and to count out loud if they wanted, thus a typical question-response was:

Q: *Taaigú tòhngmàih Waānjái jīgaān yáuh géidōgo jaahm a?*

A: *Taaigú tòhngmàih Waānjái jīgāan yáuh ... yat, yih, sāam, sei, ḿh, ... ḿhgo jaahm.*

Q: *How many stations are there between Taiku and Wanchai?*

A: *Between Taiku and Wanchai there are ... one, two, three, four, five, ... five stations.*

The data-base was designed to elicit up to 10 replicates per recording session, although if speakers counted-out loud, there were usually several more. It is an essential component of a forensic voice comparison database to include non-contemporaneous recordings [10]. For this experiment speakers were recorded on two occasions separated by about a month.

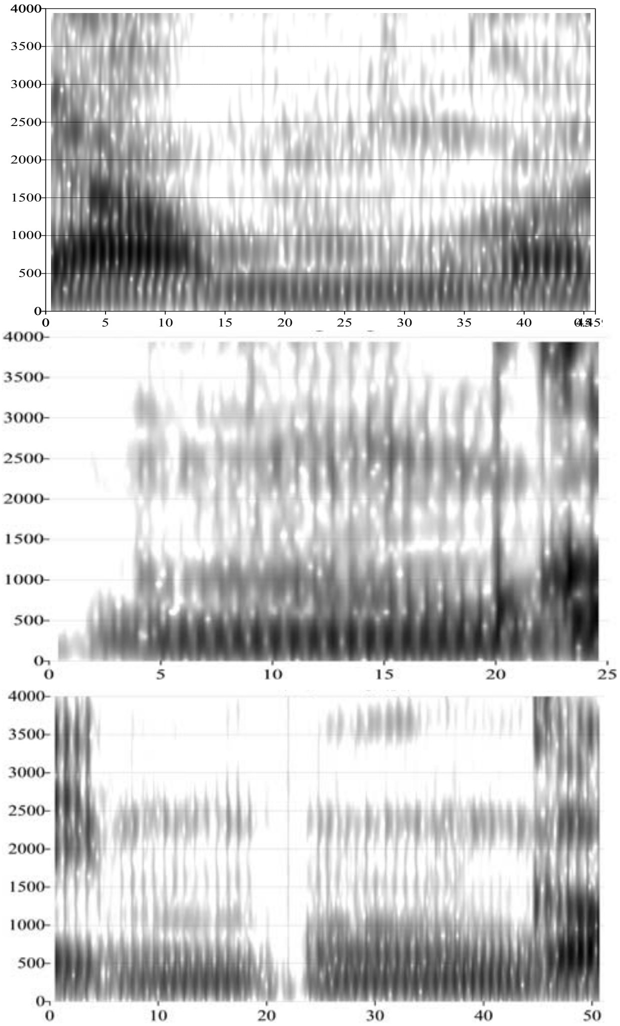

Figure 2: *Wideband spectrograms illustrating between-speaker variation in Cantonese* /syllabic nasals/. *x axis = frequency (Hz), y axis = duration (csec.)*

Examination of the nasal acoustics showed, as expected, some clear between-speaker variation in the database. In addition to the frequencies of spectral peaks, there were differences in the anticipation of the place and orality of the velar stop in the following /k/. Figure 2 shows bilabial nasals in the phrase /m̩23 ko33/ 五個... *five* ... . In the top panel, the portion corresponding to the nasal segment is clear: from ca. csec. 14 to csec. 35, and long: about 25 csec. From the detailed analysis in [11], the low frequency spectrum of nasals is expected to show a low quasi-Helmholz resonance at ca. 250 – 300 Hz, and this is clear in the spectrogram. The next highest pole is expected to be a half-wavelength resonance associated with the nasal cavity tuned by the open velopharyngeal port, at about 1000 Hz, and this is also clear, but at a lower frequency of ca. 800 Hz, and at a lower amplitude. Depending on how much of the oral cavity is available from the nasal's place of articulation, the lowest zero will be present from above 1 kHz in bilabials to ca. 3 kHz in velars. Zeros do not usually show up well in spectrograms, but there is certainly a drop in amplitude in the expected region, above the 2nd pole. The bilabial nasal should also show an F2 in the 1.5 kHz region reflecting the oral cavity resonance, and this is clear, albeit weak, at ca. 1400 Hz. This pole can be seen to stop rather abruptly at csec. 27, which indicates that the speaker has opened their lips, thus removing the front cavity. It is likely that this is done at the same time, or slightly after, a dorso-velar occlusion is effected (thus: m > m͡ŋ > ŋ), and the following 8 csec. show acoustics more typical of a velar nasal, as would be expected from anticipatory coarticulation with the following velar stop in /ko/. There is a weak transient at about csec. 35 indicating the release of the dorso-velar occlusion, but it is clear that at this time the velic port is still lowered, as the acoustic nasalization can be seen from the Helmholz resonance to continue well into the following /o/ vowel, so there cannot have been much pressure build-up behind the occlusion.

In the middle panel of Fig. 2, which shows [m̩] from another speaker, one can see the effects of the velic port closing slightly earlier, such that a stronger dorso-velar release is effected. In the bottom panel of figure 2 are shown two syllabic nasal tokens of {five}, differing in place, from a third speaker. The first is [m̩], the second is [ŋ̩], and they were said in the utterance *yat yih sāam sei ḿh … ńghgo jaahm. one two three four five ... five stations*, with about a 35 centisecond pause between them (this has been edited out in the spectrogram). Portions of the preceding /ei/ and following /ko/ can be seen. It seems here that this speaker has fully assimilated to the velar place of the following consonant. It is interesting to note that both tokens show an extra pole of unclear origin at ca. 660 Hz – stronger in the [ŋ̩] – between the two poles associated with the nasal cavity resonances. It can be seen from the tokens in Fig. 2 that the nasals have durations of ca. 15 – 20 centiseconds, which would be great luxury for a single segment if not for the latter part being counter-indicated because of spectral changes induced by anticipatory coarticulation. The ca. 10 csec. of quasi stationary signal left for acoustic analysis is nevertheless quite a lot and, one would hope, enough for the nasal to demonstrate its forensic potential.

4.1. Procedure

For each speaker, up to ten segmental nasal tokens were selected from both their non-contemporaneous recordings, all tokens of the morpheme {five} in the phrase /m̩ 23 ko33 tsa:m 22/ 五個站 *five + classifier + station*, and edited out. As mentioned above, despite their simple phonological representation, it is likely that a bilabial nasal will have spectral changes before the following velar stop in the general classifier /ko/. It was therefore important to eliminate such portions. This was done by identifying a quasi steady-state portion from the start of the nasal by inspecting wide-band spectrograms in *Praat* of the type illustrated in Fig. 2. This was usually of about 10 csec. duration. The steady-state portion of the nasal was then edited out and saved as a .wav file for further processing, which involved downsampling to 10k, and then extraction of 12th order LPC CCs, to capture an expected six poles below 5 kHz. Cms-CCs were also obtained using other, much longer portions of a speaker's database recordings, where they described how to get from one station to another. The CCs and cms-CCs were used to derive MVLRs in the same way as with the Japanese data, except that leave-one-out cross-validation was used instead of training/testing because of the small number of speakers. The MVLRs were then calibrated with logistic regression.

5. Cantonese Syllabic Nasal Results

The same badly calibrated LRs were obtained as with the Japanese mora nasal – even more so, in fact. After calibration, however, the C_{llr} decreased to 0.68, which is worse than the Japanese nasal, but still shows that the system is giving useful information. Fig. 3 shows the Tippett plot of the results. As with the Japanese mora nasal, no improvement was observed with cepstral mean subtraction. Perhaps this is related to the overall low amplitude of nasals: this is something to look into.

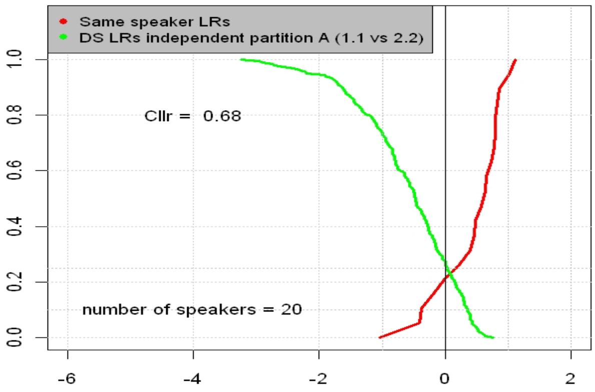

Figure 3: *Tippett plot for FVC with Cantonese syllabic nasal. Horizontal axis = Log*$_{10}$*LR greater than ...; vertical axis = cumulative proportion of non-target trials ~ 1- cum.prop. of target trials.*

6. Conclusion and Summary

This paper has described two forensic voice comparison experiments investigating whether the longer nasal segments in Japanese and Cantonese yield the greater than normal strength of evidence that might be expected from their articulation. The C_{llr} magnitudes obtained are comparable to those from other non-nasal segments obtained in the same way, and thus, contrary to expectations, there is no indication that they are superior to other segments, even when there is an unusually large amount of signal to furnish an acoustic spectrum. Future work might look into quantifying differences in pole frequencies after all: this would make us face the problem of estimating and combining discrete and continuous LRs.

Not all crimes are committed in English, so it is still useful to know about nasals like these. Fig. 4, for example, shows spectrograms of syllabic nasals from suspect and offender in a real case involving FVC in Cantonese, where the results from the Cantonese nasal in this paper were useful in providing an idea of the kind of strength of evidence to be expected under the prosecution hypothesis (and thus indicating the kind of prior odds that would be necessary to give a posterior of the desired amount). This reminds us that the kind of research reported in this paper has a serious use.

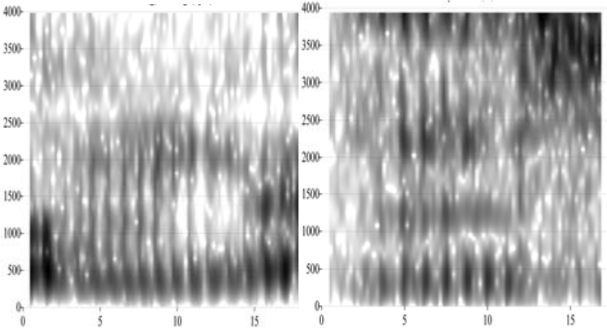

Figure 4: *Spectrograms of suspect and offender Cantonese* [m̩] *from real case-work. Axes = duration (csec.) & frequency (Hz).*

7. Acknowledgements

This paper was written using research findings from *Australian Research Council Discovery Grant* No. DP0774115. We heartily thank all the speakers who contributed their voices to the study, and our colleagues who obtained and furnished us with their own MTR database recordings. We also thank the Hong Kong University of Science & Technology for putting on their Humanities postgraduate course *Topics in Chinese Phonetics: Forensic Voice Comparison in Cantonese*, which made it possible to do this experiment. We also thank our two reviewers for their comments, a few of which we have been able to incorporate.

8. References

[1] Enzinger, E., & Zhang Cuiling, "Nasal spectra for Forensic Voice Comparison", Paper at Special Session on Forensic Acoustics, 162nd ASA Meeting, San Diego, 2011.

[2] Nolan, F. *The Phonetic Bases of Speaker Recognition*, Cambridge Studies in Speech Science and Communication, CUP, 1983.

[3] Morrison, G.S. Forensic voice comparison. In I. Freckelton, & H. Selby [Eds.], Expert Evidence (Ch. 99), Thomson, 2010.

[4] Rose, P., Osanai, T., Kinoshita, Y. "Strength of forensic speaker identification evidence: multispeaker formant- and cepstrum-based segmental discrimination with a Bayesian likelihood ratio as threshold", The International Journal of Speech Language and the Law 10(2): 179-202, 2003.

[5] Vance, T.J. An Introduction to Japanese Phonology, State University of New York Press, 1987.

[6] Aitken, C.G.G. & Lucy, D. "Evaluation of trace evidence in the form of multivariate data", Appl. Statistics 53(4): 109-122, 2004.

[7] Rose, P. "Forensic Voice Comparison with Secular Shibboleths – a hybrid fused GMM-Multivariate likelihood-ratio-based approach using alveolo-palatal fricative cepstral spectra". *Proc. International Conference on Acoustics Speech & Signal* Processing, IEEE: 5900-5903, 2011.

[8] Shen Z. "Syllabic Nasals in Chinese Dialects", in Li Fang-Kuei [Ed.], Bulletin of Chinese Linguistics, 1(1): 77-104, 2007.

[9] Bauer, R.S., & Benedict, P.K. Modern Cantonese Phonology. Mouton de Gruyter, 1997.

[10] Morrison, G. S., Rose, P., & Zhang C. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Australian Journal of Forensic Sciences, 12, 1-13. 2012.

[11] Stevens, K. Acoustic Phonetics, MIT Press, 2000.