# Effects of Fine Phonetic Detail on Speaker Identification from Japanese Nasal Consonants

Ai Mizoguchi[1,2], Mark K. Tiede[3,4] & D. H. Whalen[3,4,5]

*[1]Maebashi Institute of Technology (Japan), [2]NINJAL (Japan), [3]Haskins Laboratories (USA), [4]Yale University (USA), [5]City University of New York (USA)*
aimizoguchi@maebashi-it.ac.jp, mark.tiede@yale.edu, douglas.whalen@yale.edu

The acoustic characteristics of nasal consonants are more complicated than those of oral consonants due to involvement of both the oral and nasal vocal tracts. Theoretically, the oral cavity acts as a side-tube because the oral articulators block the airflow for a nasal consonant and the airflow is emitted only from the nostrils. As a result, the acoustic information of the place of articulation (PoA) achieved in the oral cavity appears only as anti-resonances or anti-formants in nasal acoustics [1]. The formants, not anti-formants, of nasal consonants are considered to be rather stable, reflecting the shapes of the pharynx and nasal cavity, which are largely dependent on the anatomy of each speaker. This is supported by research that indicates nasal consonants were better perceived for speaker identification [2, 3].

However, acoustic analyses have revealed some tendencies of nasal formants depending on the PoA in several languages: frequency values of the first formant (N1) are the highest for [ŋ] and lower for [ɲ], [n], and [m] in that order [4 for review]. [3] assumed the larger role for the oral cavity and examined the acoustics of /m/ and /n/ produced by Dutch speakers. It was shown that phonetic context and syllabic position affected the nasal acoustics, especially on the second formant (N2) and the spectral center of gravity (CoG) (N1 was unmeasurable due to the loss of frequencies below 300 Hz in telephone utterances).

The inter-speaker variability of the articulation of the nasal consonant /N/ (moraic nasal) in Japanese has been reported [5]. In this study, the acoustics and articulation of nasal consonants in Japanese were investigated to determine whether speaker variability predicts speaker identity.

Ten native speakers of Japanese (6 females, 4 males) participated in the experiment. The target phonemes were intervocalic nasal consonants in the words, /amata/ ‘many’, /anata/ ‘you’, /aŋa/ ‘lay down’, and /kaNaN/ ‘consideration’. The participants read aloud the words displayed on the screen one at a time in a random order and each word was repeated 10 times. The audio signal, ultrasound video showing the midsagittal image of the oral tract, and motion measurement data were recorded simultaneously. The Haskins Optically Corrected Ultrasound System (HOCUS) [6] was used to align tongue contours obtained from ultrasound images to palatal hard structure.

N1, N2, N3, and CoG were measured at the midpoint of each target phoneme using Praat [7]. The tongue contour of the midpoint was traced using GetContours [8]. The highest point of the tongue contour was identified and the values on the horizontal axis (highest_X) and vertical axis (highest_Y) were used for statistical analyses. Multinomial logistic regression analyses were performed using the ‘nnet’ package [9] in R [10].

Table 1 shows the result of the multinomial analysis predicting speakers from acoustic variables (N1, N2, N3, and CoG) and articulatory variables (highest_X and highest_Y). All the variables except N2 showed multiple significant effects on identifying speakers. The multinomial analysis was also carried out for phoneme prediction, using the same variables as for the speaker prediction. Fig. 1 shows the correct prediction rates for speakers and phonemes by each variable. The correct prediction rate was 89.1% for speakers and 79.0% for phonemes when all variables were used. The acoustic variables were more relevant for the speaker prediction than the articulatory variables and vice versa for the phoneme prediction. N1 seems to reflect some PoA information as it predicted 40.8% of the phonemes correctly. Speaker specificity seems to be most evident on N3, as N3 itself predicted speakers better than the other variables. Relationships between acoustics and articulation will be investigated in future analyses.

**Table 1.** Statistical output for multinomial analysis of speaker predictions.

*p<0.1; **p<0.05; ***p<0.01

| | *Dependent variable: Speaker* | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Coefficients:* | JF03 | JF04 | JF05 | JF06 | JF07 | JM01 | JM02 | JM03 | JM04 |
| N1_Hz | 0.099*** | 0.058*** | 0.063*** | 0.016 | 0.153*** | 0.017 | 0.079*** | -0.006 | 0.457*** |
| N2_Hz | -0.001 | 0.003 | -0.006 | -0.005 | -0.029*** | -0.0003 | -0.006 | 0.006 | -0.017 |
| N3_Hz | -0.045*** | -0.042*** | -0.058*** | -0.140*** | -0.165*** | -0.046*** | -0.052*** | -0.051*** | -0.278*** |
| N_cog | -0.123*** | -0.023 | -0.058*** | 0.081 | -0.535*** | -0.005 | -0.249*** | -0.092*** | 0.259*** |
| highest_X | 0.375*** | 0.405*** | 0.310*** | 0.637* | 0.954*** | 0.515*** | 0.405*** | 0.412*** | 3.783*** |
| highest_Y | -0.961*** | -0.531*** | -0.770*** | -1.104** | -1.343*** | 0.374 | 0.085 | 0.318 | 3.703*** |

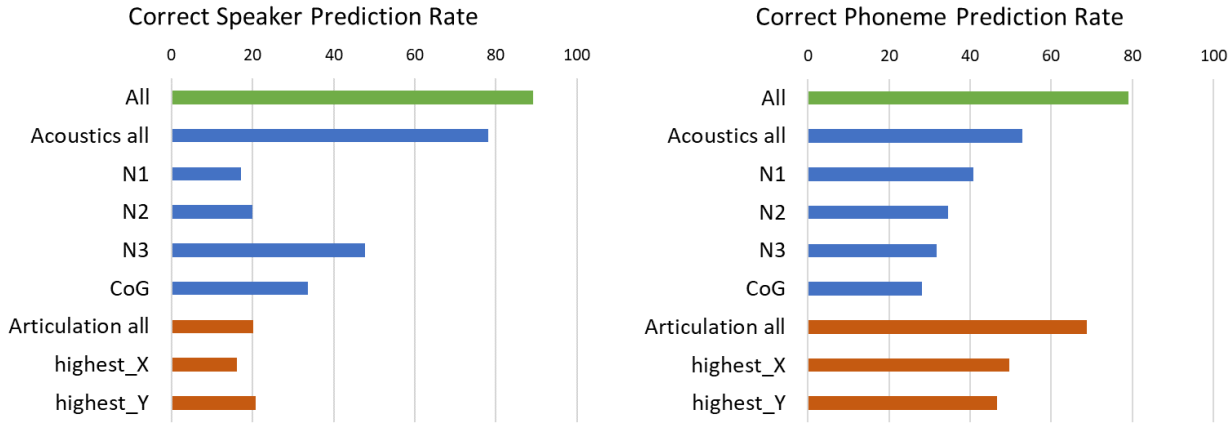


**Fig.1** Correct prediction rate for speakers (left) and for phonemes (right).

References


[1] Johnson, K. (2012). *Acoustic and auditory phonetics, 3rd ed*. Malden, MA: Wiley-Blackwell.

[2] Amino, K., & Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic Science International, 185(1–3)*, 21–28.

[3] Smorenburg, L., & Heeren, W. (2021). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *The Journal of the Acoustical Society of America, 150(2)*, 979–989.

[4] Recasens, D. (1983). Place cues for nasal consonants with special reference to Catalan. *The Journal of the Acoustical Society of America, 73(4)*, 1346–1353.

[5] Mizoguchi, A., Tiede, M. K., & Whalen, D. H. (2022). Inter-speaker variability of articulation for the Japanese moraic nasal: An ultrasound study. *Phonological Studies, 25*, 121–132.

[6] Whalen, D. H., Iskarous, K., Tiede, M. K., Ostry, D. J., Lehnert-LeHouillier, H., Vatikiotis-Bateson, E., & Hailey, D. S. (2005). The Haskins optically corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research, 48(3),* 543–553.

[7] Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer. Version 5.2.46, retrieved 7 October 2011, http://www.praat.org/.

[8] Tiede, M. K. (2016). GetContours. GitHub repository, https://github.com/mktiede/GetContours.

[9] Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S, 4th ed*. New York: Springer.

[10] R Core Team. (2019). R: A language and environment for statistical computing.