NASAL COARTICULATION IN L1 AND L2 ENGLISH SPEECH: A LARGE-SCALE STUDY

Jiahong Yuan, Hui Lin, Yang Liu

LAIX Inc.

ABSTRACT

We investigated nasal coarticulation in English produced by native English and Mandarin speakers, based on hundreds of hours of speech. A novel method was proposed to measure nasality, using softmax-based features from deep neural network models for broad phonetic classes. Segment durations were calculated from the results of forced alignment. Our analysis revealed that vowel stress tends to "block" the regressive influence of the following nasal coda, leading to less nasality in stressed vowels compared to unstressed ones. In addition, there was a positive correlation between the duration of vowels and nasal codas in L1 English, but a weak negative correlation in L2 English speech.

Keywords: Nasalization, Stress, L2, Large-scale

1. INTRODUCTION

Nasals and vowel nasalization have been extensively studied, mostly based on small amounts of articulatory, aerodynamic, acoustic, or perceptual data collected through controlled stimuli in laboratory settings. In recent years, phonetics research has started to take advantage of large speech corpora, however, very few studies have attempted to investigate nasals and vowel nasalization from this perspective, largely due to the lack of techniques for automatic measurement of nasality in the speech signal. In this paper, we developed a new method for measuring the degree of nasality automatically. With the new method, together with the widely used forced alignment technique, we conducted a study of nasal coarticulation in English produced by native English and Mandarin speakers, based on two large datasets of hundreds of hours of speech.

In speech production, a nasal consonant is made when the velum is lowered and at the same time there is a closure in the oral cavity, forcing air through the nasal passages. Vowel nasalization is the production of a vowel while the velum is lowered, so that the nasal cavities are coupled into the vocal-tract resonance system. Vowels are necessarily nasalized to some extent preceding or following a nasal consonant due to coarticulation, because it takes time to lower or raise the velum. Many acoustic parameters have been found to be related to nasalization, including a reduction in the amplitude of the first formant (A1) [14]; the relationship between A1 and the amplitude of the first harmonic (H1) [16]; nasal poles, one below the first formant (P0) and the other above the first formant (P1) [21]; the difference between A1 and P0, and the difference between A1 and P1 [5]; nasal pole-zero pairs in the vicinity of the first formant [10, 13]; and a low-frequency center of gravity [2]. As this large number of relevant acoustic parameters suggests, the acoustic consequences of nasalization are very complex. Furthermore, the degree of vowel nasalization as a coarticulatory process has also been found to be related to a wide range of factors such as vowel height [1, 3, 12, 24], phonetic context [4], non-segmental factors such as stress, prosody, and speaking rate [6, 20], as well as speaker and language characteristics [7, 11, 29]. It is, therefore, desirable to investigate nasalization using large speech corpora, with an automatic measure of nasality.

In this study, we are interested in two aspects of nasal coarticulation: the effect of lexical stress and the difference between L1 and L2 English. With regard to how stress affects vowel nasalization, there were contradictory results in the literature. Some claimed that stressed vowels appeared more nasalized by coarticulation than non-stress vowels [19, 23], while others claimed that vowels resist nasalization under stress or focus [6, 28]. Also, languages differ in terms of not only the phonemic or allophonic status of vowel nasality [9], but also how vowel nasalization is realized [7, 26]. Some studies have suggested that coarticulatory vowel nasalization is under speaker control, and fine-tuned in language-specific ways [17]. We believe that analyzing large speech corpora may help to answer these questions.

2. DATA AND SEGMENTATION

The data used in this study consisted of read sentences in English and were collected through an English-learning mobile app. Two sets of speech data were used, one from native American English

speakers (L1) and the other from Mandarin Chinese speakers (L2). The L1 dataset contained approximately half a million utterances and 500 hours of speech. The L2 dataset contained approximately 550K utterances and 620 hours of speech. All L2 speakers were from northern Mandarin dialect regions, in which there is a contrast between alveolar and velar nasal codas.

Phonetic segmentation was done automatically through forced alignment using the HTK Toolkit [15]. To achieve better segmentation accuracy, we trained GMM-HMM acoustic models of both phones and phone boundaries, following the method in [27]. The phone models were standard 3-state HMMs. The phone boundary models were a special 1-state HMM, in which the state cannot repeat itself. Forced alignment on the L1 and L2 data was performed separately, with acoustic models trained on the two datasets respectively.

The CMU pronouncing dictionary was used for forced alignment [8]. In the dictionary a stress marker is placed after every vowel: "1" for primary stress, "2" for secondary stress, and "0" for no stress (reduced vowels). To avoid the complexity arising from syllabification and secondary stress, we restricted our analysis to the word-final alveolar nasal /n/ and the preceding vowel, either in primary stress (hereafter V1N) or no stress (hereafter V0N). Based on the results of forced alignment, we excluded tokens in the data whose duration was not in a typical range (likely due to speech errors or alignment errors). For vowels, both V1 and V0, the duration range was set to between 60 and 300 msec; for the nasal coda, the duration range was between 40 and 200 msec. The total number of tokens used in our analysis is listed in Table 1.

Table 1: Total number of tokens for analysis.

	L1 English	L2 English
V0N	29,286	38,037
V1N	146,410	131,129

3. A METHOD FOR AUTOMATIC MEASUREMENT OF NASALITY

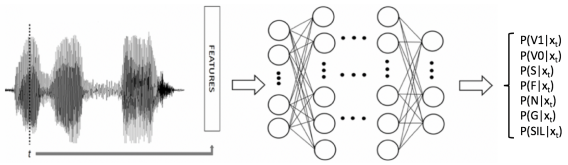
The method we proposed for automatic measurement of nasality was based on deep neural network models of broad phonetic classes. Following [25], we mapped English phonemes into six broad phonetic classes plus a class for silence, as listed in Table 2. In order to measure nasality, we used the softmax function as the output layer of the neural network, which generates a vector representing the probability of each and every broad class. The probability of the nasal class was then used to measure nasality in the speech signal.

Table 2: Broad phonetic classes.

Class	Description: phonemes
V1	Stressed vowels: 1 2
V0	Non-stress vowels: 0
N	Nasals: /M N NG/
S	Stops and affricates: /B CH D G JH K P T/
F	Fricatives: /DH F HH S SH TH V Z ZH/
G	Glides and liquids: /L R W Y/
SIL	Silence: –

We trained a Kaldi [18, 22] TDNN (nnet3) model of the seven classes (including silence) with 80% of the L1 dataset, i.e., 400 hours of L1 English speech, and used the model to compute softmax values for a given time point in the speech signal. The procedure is illustrated in Figure 1.

Figure 1: The procedure of computing broad-class probabilities.



To evaluate the performance of the model, we extracted a softmax vector at the phone center for every phone in the entire L1 dataset (the phone center was determined by the forced alignment process described above), and selected the class with the highest probability as the recognition result. The recognition accuracy is listed in Table 3, for test and training data respectively. Among the six phonetic broad classes, the overall accuracy was about 74%. Most of the errors were between similar classes, for example, V0/V1 and V1/G. If we group the six classes into nasal (N) and non-nasal (V1, V0, S, F, G), the accuracy was better than 96%.

Table 3: Broad class recognition accuracy.

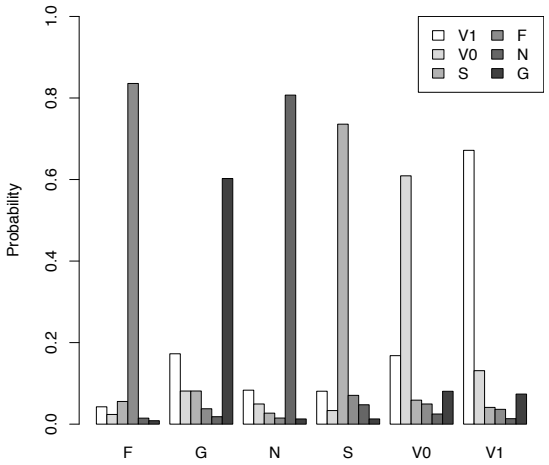
	6 classes	nasal vs. non-nasal
Test set	74.1%	96.5%
Training set	74.7%	96.6%

From Table 3 we can also see that there is little difference between the training and test sets in terms of recognition results. Therefore, we combined the

two sets in the following analysis.

Figure 2 shows the mean probabilities of the six broad phonetic classes, computed from all phones in the entire L1 dataset. Note that we obtained the probabilities for the center of the phones. We can see that for each broad phonetic class in the data (determined by the CMU dictionary, as the gold standard), the probability of the class of the gold standard is much higher than the probabilities of the other classes.

Figure 2: Probabilities of broad phonetic classes for all the phoneme tokens in the L1 dataset.



4. RESULTS

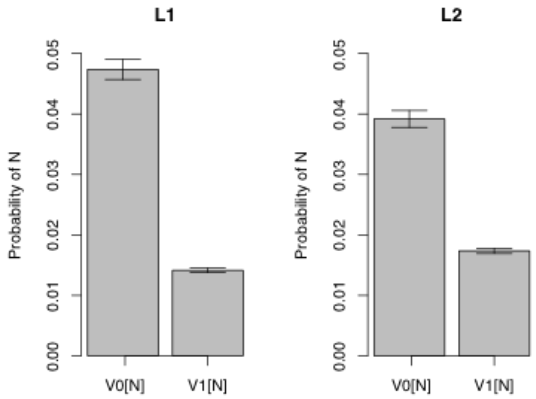
4.1. The effect of stress on vowel nasalization

Using the proposed method, we extracted a softmax vector at the center of the vowel for all the selected V1N and V0N tokens described in Section 2. The probability of the nasal dimension in the softmax vector was used to measure the degree of nasality in the vowel. Figure 3 compares the difference between V1N and V0N on the degree of nasality at the center of the vowel.

Clearly, for both L1 and L2 English the nasality is lower for V1N. That is, when preceding a nasal coda, stressed vowels bear less nasality than non-stress vowels. There are at least two possible explanations for this result. The first explanation is that stress resists or blocks nasalization by coarticulation. The second one is that because stressed vowels are longer than non-stress vowels, the center of a stressed vowel is further away from the following nasal coda and therefore less influenced by the nasal than a non-stress vowel.

To test the second hypothesis, we extracted softmax vectors at six different time points in every token: the boundary between the vowel and the nasal coda (B); 10, 20, 30 msec from the boundary into the vowel (B-1, B-2, B-3); and 10, 20 msec from the boundary into the nasal coda (B+1, B+2). The results are shown in Figure 4, for L1 and L2 respectively.

Figure 3: Nasal probability at vowel center.



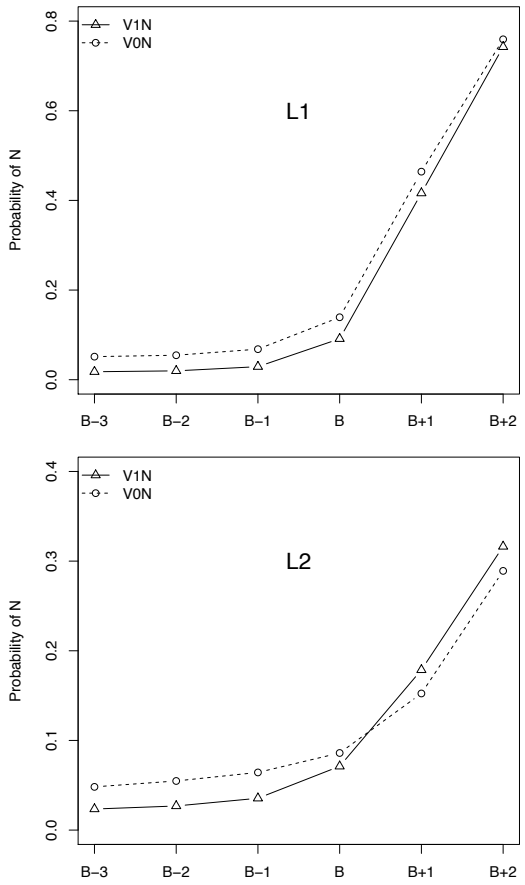
By comparing the three time points in the vowel, B-3, B-2, and B-1, we can see that when preceding a nasal coda stressed vowels are less nasalized than non-stress vowels, and this is true for both L1 and L2 English. As these points were determined by the distance in time from the nasal coda, regardless of the vowel duration, we can conclude that the second hypothesis is not supported by our data. Our data demonstrate that lexical stress resists or blocks the influence of the following nasal coda, so stressed vowels are less nasalized than non-stress vowels.

Figure 4 also shows a difference between L1 and L2 English on the nasal coda. It appeared that the nasal coda had less nasality in stressed syllables (compared to non-stress syllables) in L1 English, but more nasality in stressed syllables in L2 English.

4.2. The duration of vowels and nasal codas

With regard to the duration of vowels and nasal codas, our data show an interesting difference between L1 and L2 English. Figure 5 contains smooth scatterplots of the vowel and nasal coda duration in V1N, for L1 and L2 respectively. In L1 there was a positive correlation between the vowel and nasal coda duration ($r = 0.34$, $p < 0.001$) whereas in L2 the correlation was negative, although very weak ($r = -0.06$, $p < 0.001$).

Figure 4: Nasality at six time points.



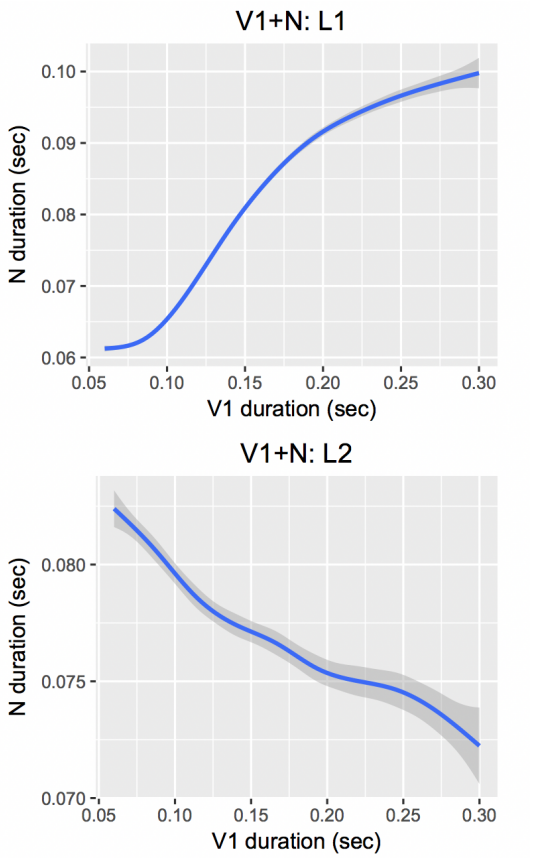
The negative correlation between the duration of stressed vowels and nasal codas in L2 English might be a result of negative transfer of the rhythm of Mandarin as a syllable-timed language. However, additional research is needed to draw a conclusion.

5. DISCUSSION AND CONCLUSIONS

We developed a new method for automatic measurement of nasality in large speech corpora, based on deep neural networks and broad phonetic classes. By grouping phonemes into broad phonetic classes, we trained acoustic models to represent the acoustic property of each broad class instead of individual phonemes, e.g., the common property (nasality) of all nasal sounds.

We used the softmax function as the output layer of the neural network, and the probability of the nasal class to measure nasality in the speech signal. The nasal probability in nasalized vowels were relatively small, as we can see from the results. This

Figure 5: Duration of V1 and N in V1N.



does not affect our analysis because we only compared vowels in different contexts. However, if the goal is to detect nasality in speech, we need to calibrate/normalize the measure using non-nasalized vowels (vowels not adjacent to a nasal consonant) as a reference.

We trained only one model using L1 speech, and the vowel classes were trained using all the vowel tokens, including those preceding or following a nasal consonant. Further effort is needed to test how models trained on L2 or only on non-nasalized vowels will affect the nasality measure.

In summary, our analysis of large speech corpora demonstrated that in both L1 and L2 English lexical stress blocks vowel nasalization, leading to less nasality in stressed vowels. We found a positive correlation between the duration of vowels and nasal codas in L1 English but a weak negative correlation in L2 English.

6. REFERENCES

[1] Abramson, A., Nye, P., Henderson, J., Marshall, C. 1981. Vowel height and the perception of consonant nasality. *J Acoust. Soc. Am.* 70, 329–339.

[2] Beddor, P. 1982. *Phonological and phonetic effects of nasalization on vowel height*. University of Minnesota: PhD thesis.

[3] Bell-Berti, F. 1993. Understanding velic motor control: Studies of segmental context. In: Huffman, M., R.A., K., (eds), *Nasals, Nasalization, and the Velum*. New York: Academic Press 63–85.

[4] Busà, M. 2007. Coarticulatory nasalization and phonological developments. In: Solé, B. P., M.J., M., O., (eds), *Experimental Approaches to Phonology*. Oxford: Oxford University Press 155–174.

[5] Chen, M. 1997. Acoustic correlates of english and french nasalized vowels. *J Acoust. Soc. Am.* 102, 2360–2370.

[6] Cho, T., Kim, D., Kim, S. 2017. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in english. *J Phon.* 64, 71–89.

[7] Clumeck, H. 1976. Patterns of soft palate movements in six languages. *J Phon.* 4, 337–351.

[8] The cmu pronouncing dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[9] Cohn, A. 1990. Phonetic and phonological rules of nasalization. *UCLA WPP* 76, 1–224.

[10] Dang, J., Honda, K., Suzuki, H. 1994. Morphological and acoustical analysis of the nasal and the paranasal cavities. . *J. Acoust. Soc. Am* 96, 2088–2100.

[11] Ha, S., Kuehn, D. 2006. Temporal characteristics of nasalization in children and adult speakers of american english and korean during production of three vowel contexts. *J. Acoust. Soc. Am* 120, 1622–1630.

[12] Hajek, J., Maeda, S. 2000. Investigating universals of sound change: the effect of vowel height and duration on the development of distinctive nasalization. In: Broe, M., J., P., (eds), *Papers in Laboratory Phonology V*. Oxford: Cambridge University Press 52–69.

[13] Hawkins, S., Stevens, K. 1985. Acoustic and perceptual correlates of the nonnasal-nasal distinction for vowels. *J. Acoust. Soc. Am.* 77, 1560–1675.

[14] House, A., Stevens, K. 1956. Analog studies of the nasalization of vowels. *J. of Speech and Hearing Disorders* 21, 218–232.

[15] The htk toolkit. http://htk.eng.cam.ac.uk.

[16] Huffman, M. 1990. Implementation of nasal: Timing and articulatory landmarks. *UCLA WPP* 75, 112–143.

[17] Jang, J., Kim, S., Cho, T. 2018. Focus and boundary effects on coarticulatory vowel nasalization in korean with implications for cross-linguistic similarities and differences. *J. Acoust. Soc. Am.* 144, EL33–EL39.

[18] The kaldi speech recognition toolkit. http://kaldi-asr.org.

[19] Krakow, R. 1989. *The Articulatory Organization of Syllables: A Kinematic Analysis of Labial and Velar Gestures*. Yale: PhD thesis.

[20] Krakow, R. 1993. Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate. In: Huffman, M., R.A., K., (eds), *Nasals, Nasalization, and the Velum*. New York: Academic Press 87–113.

[21] Maeda, S. 1982. The role of the sinus cavities in the production of nasal vowels. *Proc. ICASSP 1982* Paris. 911–914.

[22] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K. Dec. 2011. The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

[23] Schourup, L. 1973. A cross-language study of vowel nasalization. *OSU WPL* 15, 190–221.

[24] Young, L., Zajac, D., Mayo, R., Hooper, C. 2001. Effects of vowel height and vocal intensity on anticipatory nasal airflow in individuals with normal speech. *J. Speech Lang. Hear. Res.* 44, 52–60.

[25] Yuan, J., Liberman, M. 2010. Robust speaking rate estimation using broad phonetic class recognition. *Proc. ICASSP 2010* Dallas. 4222–4225.

[26] Yuan, J., Liberman, M. 2011. Automatic measurement and comparison of vowel nasalization across languages. *Proc. 17th ICPhS* Hong Kong. 2244–2247.

[27] Yuan, J., Ryant, N., Liberman, M., Stolcke, A., Mitra, V., Wang, W. 2013. Automatic phonetic segmentation using boundary models. *Proc. Interspeech 2013* Lyon. 2306–2310.

[28] Zajac, D., Mayo, R., Kataoka, R. 1998. Nasal coarticulation in normal speakers: A re-examination of the effects of gender. *J. Speech Lang. Hear. Res.* 41, 503–510.

[29] Zellou, G. 2017. Individual differences in the production of nasal coarticulation and perceptual compensation. *J. Phon.* 61, 13–29.