Nasality in Speech and Its Contribution to Speaker Individuality

Kanae Amino\(^1\), Hisanori Makinae\(^1\), Tatsuya Kitamura\(^2\)

\(^1\) National Research Institute of Police Science, Japan
amino@nrips.go.jp, makinae@nrips.go.jp, t-kitamu@konan-u.ac.jp

\(^2\) Konan University, Japan

Abstract

The term nasality refers to the timbre of the nasal phonemes. It is also used to express the quality of sound that characterises some speakers. In this paper, we propose to classify nasality in natural speech into four types: phonemic nasality, nasality in assimilation, incidental nasality in the production of voiced plosives, and nasality associated with speaker individuality. Speech sounds recorded separately for oral and nasal outputs were analysed and the four types of nasality were observed individually. In order to investigate the relationship between the nasality in running speech and the perception of speaker similarity, we conducted an experiment. The results revealed that listeners rated speaker similarity exploiting phonemic nasality when it existed in the utterance and also used speaker-related nasality regardless of the existence of phonemic nasals.

Index Terms: nasality, speaker individuality, oro-nasal speech, perception of speaker similarity

1. Introduction

Previous studies on automatic speaker identification report the effectiveness of focusing on nasal phonemes as speaker indexing cues \([1, 2]\). It is also true that the stimuli containing nasal phonemes are more effective for perceptual speaker identification than those without them \([3]\). These reports imply that nasal sounds convey more speaker-related information than non-nasal sounds. On the other hand, people can well identify familiar speakers by speech even without any nasal phonemes \([3]\). Rose \([1]\) points out that some people “characteristically speak with a nasal-twang,” and we can predict that these people always speak in a nasal voice. Considering these facts, there are at least two types of nasality, phonemic and speaker-individual. The classification of the types of nasality has not yet been seriously discussed in speech science. The present paper re-organises the types of nasality, and shows that different types of nasality can be observed separately in speech signals by using an oro-nasal recording mask. We also show how people exploit speaker-related information in the nasality when they perceive speaker individualities.

We first look at the different types of nasality existing in speech sounds. We propose to classify nasality into the four types shown in Table 1. In phonetics, the term nasal is defined as the class of sounds produced with the lowered velum, thus accompanying an airflow through the nasal tract. The timbre of the nasal sounds is called nasality, and in phonology, nasality is associated with the phonological feature \([\text{+nasal}]\). Nasality in phonemic nasals \((N_{ph})\) thus appears in the nasal consonants (e.g., /m/, /n/ in many languages including Japanese and English) and nasalised vowels (e.g., /â/ in French or /u/ in Portuguese). In natural speech, it often occurs that the feature \([\text{+nasal}]\) of a nasal phoneme is spread to neighbouring segments due to perseverative and/or anticipatory assimilations. This phenomenon, so-called nasalisation, is contextual and can be explained by the physiological constraints of the human vocal tract, by which muscular control of the velum is slower than the movements of other articulators, such as the tongue. Contextual nasality \((N_{c})\) can be observed in the segments adjacent to phonemic nasals (e.g., [â] in /ma/ in Japanese).

Other types of nasality do not necessarily involve lowering of the velum. Dang \([4]\) found that voiced plosives followed by a vowel, i.e., phonologically \([-\text{nasal}]\) sound sequences, have some nasality that originates from the vibration of the velum. This is not a result of an active velic movement but of sound pressure transmission from the oral cavity. This type of nasality \((N_{vq})\) can be seen in voiced plosives, particularly in bilabial and alveolar ones (e.g., [ba], [da]). Finally, nasality as a feature of voice quality \((N_{vq})\) characterises a speaker’s individuality. Abercrombie \([5]\) defines voice quality as “a quasi-permanent quality running through all the sound that issues from his [speaker’s] mouth,” and so \(N_{vq}\) is inevitably conveyed in all speech sounds uttered by a speaker.

In this paper, we aim to (1) show that these four types of nasality exist in running speech, (2) investigate their relationships, and (3) examine how nasality contributes to the perception of speaker individuality. Section 2 illustrates how we observed the four types of nasality in speech samples recorded separately for the oral and nasal outputs. Section 3 describes the procedures and results of the perceptual experiment where listeners rated the similarity among speakers. Finally, in Section 4, we discuss the position of the nasality in speaker individuality by looking at the correspondences between acoustic and perceived speaker similarity.

<table>
<thead>
<tr>
<th>Types of Nasality</th>
<th>Exists where?</th>
<th>Phonetic or Phonological?</th>
<th>States of the Velum</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N_{ph})</td>
<td>Nasality in phonemic nasals</td>
<td>Nasal consonants and nasalised vowels</td>
<td>Phonological, [+nasal]</td>
</tr>
<tr>
<td>(N_{c})</td>
<td>Nasality in contextual nasalisation</td>
<td>Segments adjacent to the phonemic nasals</td>
<td>Phonetic/phonological, [+nasal] (spread)</td>
</tr>
<tr>
<td>(N_{vq})</td>
<td>Nasality due to the vibration of the velum</td>
<td>Voiced plosives</td>
<td>Phonetic</td>
</tr>
<tr>
<td>(N_{vq})</td>
<td>Nasality as a feature of voice quality</td>
<td>Through all the sounds</td>
<td>Phonetic</td>
</tr>
</tbody>
</table>
2. Analysis of Oro-Nasal Speech

2.1. Recording of the speech materials

Sixteen male speakers participated in the recording sessions. They were in their 20s or 30s (mean age of 29.8 years), and all spoke Japanese as their native language. None had previously suffered from any nasal diseases such as empyema, adenoid or septal deviation.

The recordings were conducted in an anechoic room at NRIPS. Speakers read short sentences and syllables written on the cards three times each in two recording sessions. In the first recording session, speech samples were recorded monaurally by using a microphone (SONY ECM-23F5) and PCM recorder (Marantz PMD671) at a sampling frequency of 48 kHz with 16-bit resolution. Henceforth, we call these samples M-speech. In the second recording session, oral and nasal speech signals (hereafter, O- and N-signals) were simultaneously recorded by using the oro-nasal mask (Glottal Enterprises OroNasal Speech Tutor Mask MA-1L) shown in Fig 1, microphones (SONY ECM-77B), and an audio interface (Roland EDIROL UA-5). The speech signals were sampled at 48 kHz with 16-bit resolution and saved on a laptop by using Praat [6].

2.2. Methodology

The analysis targets are the two short sentences: "倫敦でミイラを見る /rondonde mi:rao mira/ (Null-subj. see a mummy in London)," and "アラビアで油を売る /arabiade aburao mura/ (Null-subj. sell oil in Arabia)" (hereafter, London-sentence and Arabia-sentence). The London-sentence contains four phonemic nasals, whereas the Arabia-sentence contains none.

The target sentences uttered by the sixteen speakers were analysed. Before the analysis, all speech materials were resampled at 16 kHz. The power of the O- and N-signals was calculated using a 30-ms Hanning window with 10-ms shifts. In order to see the relative movements of the two signals, the log power ratio $R$ [dB] was obtained by the following equation.

$$ R = 10 \times \log_{10}(P_o/P_n), \quad (1) $$

where $P_o$ and $P_n$ are the powers of the N- and O-signals, respectively.

2.3. Results and Discussion

Examples of waveforms and spectrograms of the O- and N-signals uttered by two speakers are indicated in Fig. 2 along with the power-ratio contours. On observing the waveforms of the N-signals for the London-sentence, we notice from Figs.

Figure 1: The oro-nasal mask that separates oral and nasal airflow; the output side (left) and the input (worn) side (right).

Figure 2: The waveforms and spectrograms of the O-signal (above), and the N-signal (middle), and the N/O power-ratio contour (below); (a) and (c) represent the London-sentence uttered by two male speakers, while (b) and (d) represent the Arabia-sentence uttered by the same speakers.
2(a) and 2(c) that the amplitude becomes large in syllables with phonemic nasals. These amplitude peaks are associated with \( N_{ph} \) (for nasal phonemes) and \( N_c \) (for the whole syllables). As shown in Figs. 2(b) and 2(d), by observing the waveforms for the Arabia-sentence, which contains no phonemic nasals, we see that the N-signal are still detected, although their amplitudes are much smaller than those in the London-sentence. This N-signal can be considered as \( N_{ph} \), and possibly corresponds to the degree of nasal timbre that characterises some speakers. If this is true, \( N_{ph} \) also appears in the London-sentence, although it is masked by \( N_c \).

Finally, in the power-ratio contours, we see the peaks occurring in synchronisations with certain consonants, i.e., /\( b/\), /\( d/\), and /\( r/\). These peaks are assumed to correspond to \( N_vq \). It is difficult to see \( N_{ov} \) in the London-sentence, as the peaks for /\( d/\) overlap the peaks for \( N_{ph} \) and \( N_c \), but the peaks for /\( r/\) in the first and last syllables are recognisable in both Figs. 2(a) and 2(c). Dang [4] confirmed that the vibration of the velum occurs in bilabial and alveolar voiced plosives, /\( b/\) and /\( d/\), but not in the alveolar flap /\( r/\). The articulation of the flap is similar to that of plosives in Japanese; the only difference is that the closure duration is much shorter and the articulatory movement is much more rapid in flaps.

3. Perception Experiment

3.1. Procedure

In order to investigate the role of nasality in the perception of speaker individuality, we conducted an experiment to rate the speaker similarity. A subset of the recorded speech samples were used as stimuli. The two target sentences analysed in Section 2 (the London- and Arabia-sentences) were used. We used M-speech recorded from eight male speakers, resampled at 16 kHz. Twenty-eight listeners volunteered to participate in the experiment. All had normal hearing and their mean age was 34.6 years. They had never heard the voice of any of the speakers before.

The experiment was conducted using Praat [6] in a sound-treated room. The sentences uttered by two different speakers were presented pairwise to the listeners through a digital audio processor (ONKYO SE-U55SX) and headphones (Sennheiser HD650) at a comfortable sound pressure level. The listeners rated the similarity of the two speakers on the five-point scale shown in Table 2.

The total number of stimuli was 144, which corresponds to 36 speaker pairs (28 different- and 8 same-speaker pairs), 2 sentences, and 2 presenting orders. Two different tokens were used for the same-speaker pairs. No feedback was given and the listeners took breaks after every 36 stimuli.

3.2. Results

Speaker similarity matrices were obtained for each of the two sentences, as shown in Tables 3 and 4, respectively. The scores were averaged over the listeners. In both matrices, the similarity scores were close to 5 for the same-speaker pairs.

By comparing the two matrices, we can estimate the effects of \( N_{ph} \) and \( N_c \), although the effects of other phoneme differences are also involved. For some speaker pairs (e.g., S5-S6 and S6-S8), perceptual similarity became greater by adding \( N_{ph} \) and \( N_c \) in the London-sentence; and for other pairs (e.g., S3-S5 and S5-S8), the similarity decreased. The effect of \( N_{ov} \) cannot be directly observed; however it should be existing in conjunction with other speaker individualities such as phonation properties.

4. Correspondences between Acoustic and Perceived Speaker Individuality

4.1. Distance analysis

The correspondences between acoustic distance and perceptual speaker individualities were examined. For calculation of the acoustic distances, M-speech and the O- and N-signals uttered by the eight speakers of the perceptual experiment were used. FFT cepstra were obtained by 30-ms frames with a 10-ms shift. We used 30-dimensional coefficients excluding the zeroth coefficient. For all token pairs, the cepstral distances were calculated by applying dynamic time warping (DTW) [7]. For each pair, the distance was normalised by dividing it by the total number of frames.

Intra- and inter-speaker distances were averaged over different token pairs for each sentence and each speech signal and summarised as distance matrices. Finally, the cepstral distance matrices and perceptual speaker similarity matrices were submitted to correlation analyses, for which the intra-speaker distance and similarity were omitted from the data.

Table 2. Five-point scale for speaker similarity rating.

<table>
<thead>
<tr>
<th>Point</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dissimilar</td>
</tr>
<tr>
<td>2</td>
<td>Slightly dissimilar</td>
</tr>
<tr>
<td>3</td>
<td>Slightly similar</td>
</tr>
<tr>
<td>4</td>
<td>Similar</td>
</tr>
<tr>
<td>5</td>
<td>Same speaker</td>
</tr>
</tbody>
</table>

Table 3. Speaker similarity matrix for the London-sentence; average scores of the twenty-eight listeners.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4.88</td>
<td>2.84</td>
<td>1.52</td>
<td>2.77</td>
<td>1.66</td>
<td>1.75</td>
<td>1.88</td>
<td>2.14</td>
</tr>
<tr>
<td>S2</td>
<td>5.00</td>
<td>2.00</td>
<td>2.89</td>
<td>1.61</td>
<td>2.11</td>
<td>1.46</td>
<td>2.80</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>4.98</td>
<td>2.13</td>
<td>1.73</td>
<td>1.93</td>
<td>1.36</td>
<td>2.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>4.95</td>
<td>2.32</td>
<td>2.66</td>
<td>2.21</td>
<td>3.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>5.00</td>
<td>3.41</td>
<td>2.54</td>
<td>1.84</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>4.98</td>
<td>2.48</td>
<td>2.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>4.88</td>
<td>1.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>4.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Speaker similarity matrix for the Arabia-sentence; average scores of the twenty-eight listeners.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>4.98</td>
<td>2.55</td>
<td>1.73</td>
<td>2.14</td>
<td>1.77</td>
<td>1.63</td>
<td>2.00</td>
<td>2.02</td>
</tr>
<tr>
<td>S2</td>
<td>4.95</td>
<td>2.16</td>
<td>2.29</td>
<td>1.66</td>
<td>1.55</td>
<td>1.45</td>
<td>3.07</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>4.89</td>
<td>2.05</td>
<td>2.48</td>
<td>2.41</td>
<td>1.73</td>
<td>1.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>4.93</td>
<td>2.73</td>
<td>2.48</td>
<td>2.34</td>
<td>3.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S5</td>
<td>4.82</td>
<td>2.63</td>
<td>2.43</td>
<td>2.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>4.82</td>
<td>2.55</td>
<td>1.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>4.86</td>
<td>1.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>4.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2. Results and discussion

The distance matrices for the O- and N-signals of the London- and Arabia-sentences are shown in Tables 5 and 6, respectively, and the results of the correlation analyses are summarised in Table 7.

For the London-sentence, negative correlation was significant in all three speech signals: M-speech ($t(28) = 3.78$, $p < .001$), O-signal ($t(28) = 3.60$, $p = .001$), and N-signal ($t(28) = 2.50$, $p = .02$). For the Arabia-sentence, on the other hand, negative correlation was significant in the O-signal ($t(28) = 2.46$, $p = .02$) and showed a significant tendency in the N-signal ($t(28) = 1.82$, $p = .08$). This means that the rating of speaker similarity was conducted by using the O-signal in both sentences, while the N-signal was also used depending on its availability. The effect of the N-signal was greater for the London-sentence, where not only $N_{ph}$ but also $N_p$ and $N_c$ were available, as compared to the Arabia-sentence, where only $N_{vv}$ was available.

5. Discussion and Conclusions

By analysing speech data recorded separately for oral and nasal outputs, we succeeded in individually observing the four kinds of nasality: $N_{ph}$, $N_c$, $N_{vv}$, and $N_{vv}$. Although the size of the speaker set was small, observed tendency was consistent among speakers. The results of the perception experiment and the analyses of the speech data revealed that perceptual speaker similarity significantly correlated with inter-speaker cepstral distances of the N-signal in the [+nasal] sentence and showed a correlation tendency in the [-nasal] sentence. This implies that the rating of speaker similarity is associated with $N_{ph}$ and $N_c$ when they are available, and also on $N_{vv}$ regardless of the [+/-nasal] feature of the sentence.

The resonant properties of the nasal cavity and paramanual sinuses are speaker-dependent, because the morphology of these cavities differs greatly among individuals [8]. This speaker-dependent resonance should be reflected in speech signals as $N_{ph}$ and $N_c$ when the velum is lowered, and also as $N_{vv}$ regardless of the state of the velum. Moreover, speaker-dependent velic movements are reflected in the articulation of phonemic nasals [9], and the listeners could have used this information as part of the speaker indexing cues, too. We have not yet clarified how different the articulatory movements are among speakers. Especially the control of the velic actions may affect both the resonance property of $N_{ph}$ and the timing relation of $N_{ph}$ and $N_c$. Observing the inter-speaker differences in articulatory movements will lead us to know more about the relationship between speaker individuality and nasality.

Although the nasality was effective in the perception of speaker similarity, the effect of the O-signal was yet greater and there must be influential oral features. In future work, we should find these speaker-related acoustic features. The target sentences in this study differed only in the [+/-nasal] feature of the sound sequences they contained, and other segmental and supra-segmental properties were not controlled. We can easily predict that the effect of the fundamental frequency is considerable [10, 11]. Also, roughness of voice is reported to be an important factor in speaker individuality [12]. Testing other sentences and investigating with other experimental tasks, such as speaker identification and rating of the degree of nasality, are also some of our future tasks.

In this paper, we classified nasality into the four types. There may be more types of nasality, when we look at different aspects of speech. For example, nasality is an important factor in speech diagnosis. Hyper- and hypo-nasality should be differentiated in speech pathology from the four types of nasality proposed in this study. Also, nasality sometimes indicates the provinces of one’s origin. Some dialects are “notorious” of its strong nasality, and it might be significant in social linguistics. Furthermore, in Japanese culture, nasality is emblematic of cuteness. In this situation, nasality provides a clue to analysing the speakers’ attitudes. Re-organising the types of nasality may bring about a new trend in speech science and it may also give us a fresh vision of phonetic and phonological theories.

6. Acknowledgements

The authors would like to thank Professor Imagawa and Professor Hashi for their support in the recordings. This study was supported by Grants-in-Aid for Scientific Research KAKENHI (25280066, 25240026, 24710195, and 24810034).
7. References