Original Article

Speaker Identification Using Japanese Monosyllables and Contributions of Nasal Consonants and Vowels to Identification Accuracy

Kanae Amino and Takashi Osanai

National Research Institute of Police Science
6–3–1 Kashiwanoha, Kashiwa-shi, Chiba, 277–0882, Japan

(Received 22 December 2011; accepted 11 June 2012)

Previous research on speaker identification has demonstrated the effectiveness of using syllables containing a nasal consonant. In this study, we investigated the contributions of nasal consonants (/m/ and /n/) and vowels (/i/, /e/, /a/, /o/, and /u/) to identification accuracy by using them separately in speaker identification experiments. Japanese monosyllables with nasal onsets were recorded from 50 male speakers using a condenser microphone. Two recording sessions were held and thus non-contemporaneous speech data were obtained. Nasal consonants and the following vowels were excerpted from the recorded monosyllables, and 30th-order cepstral coefficients were calculated for each as acoustic features. The results revealed that the accuracy of identification using nasal consonants was not as high as that using vowels; more than six nasal tokens for a given speaker needed to be registered in order to match the score afforded by one vowel token for the same speaker. The higher vowels, /u/, /e/ and /i/, yielded significantly better identification rates than the lower vowels, /a/ and /o/, and the alveolar nasal /n/ was better than the bilabial /m/. We also conducted a factor analysis in order to clarify the effects of the attributes of speakers and speech samples on the differentiation between speakers with similar speech characteristics. Analysis was performed on frequently confused speaker pairs using 11 parameters for vowels and six parameters for nasals. The parameters were selected from various attributes related to the physiological properties of the speakers and the acoustic properties of their speech. The results showed that irregularity in phonation and the degree of vowel nasalisation were among the most influential factors. The physical size of the speakers and the average fundamental frequencies also affected the accuracy of speaker identification.

Key words: Speaker identification, Nasal consonant, Differentiation between speakers, Phonation irregularity, Degree of nasalisation

1. Introduction

Speech is a behavioural feature that can be used to identify individuals in forensics. Identification of individuals by speech (speaker identification; henceforth referred to as SPID)

* 本論文の内容は日本法科学技術学会第16回学術集会発表における奨励賞を受賞した.
can be performed by both computers (automatic SPID) and human observers. Previous research has shown that the accuracy of SPID depends on various factors\(^1\)\(^2\). In human SPID, listeners may be either earwitnesses or professional investigators. Factors that influence the ability of earwitnesses to identify a speaker include the familiarity between the speaker and listener\(^3\), stress on the listener\(^4\), and visibility of the speaker’s face\(^5\). In automatic SPID, in contrast, problematic issues include the existence of background noise and properties of the transmission channels\(^6\)\(^7\)\(^8\). Factors such as voice disguise\(^9\)\(^10\)\(^11\), non-contemporaneous speech\(^12\)\(^13\), and the duration and phonological contents of the speech\(^14\)\(^15\)\(^16\)\(^17\)\(^18\) are influential in both human and automatic SPID accuracy.

Previous research on the effects of phonological contents on SPID has shown that vowels and sonorous consonants are more effective than obstruent consonants in many languages, including English, Russian, and Chinese\(^14\)\(^15\)\(^16\)\(^17\)\(^18\). Amino and Arai\(^19\)\(^20\) investigated the differential effects among Japanese monosyllables on perceptual SPID accuracy and confirmed that this tendency is also seen in Japanese. They also found that, among sonorous consonants, nasal sounds yielded significantly better SPID scores than other non-nasal consonants, confirming the results of previous studies in other languages\(^21\)\(^22\). Furthermore, their analyses of stimuli showed that the cepstral distances among speakers were greater in nasals than in other consonants\(^20\). The availability of nasals has also been reported\(^23\)\(^24\)\(^25\) using automatic methods.

The term “nasal” refers to a class of sounds whose articulation involves the lowering of the velum (soft palate); thus these sounds are accompanied by airflow through the nasal tract. Acoustic characteristics of nasals include the resonance of the nasal cavity and paranasal sinuses. The morphology of these cavities is reported to differ greatly among individuals, resulting in acoustic differences among speakers\(^26\)\(^27\)\(^28\). In addition, the articulation of a nasal sound involves movements of the velum; these movements, or velic actions, can be used as another indexical cue to identify a speaker\(^29\). In consideration of these facts, Rose\(^30\) identified nasals as one of the strongest SPID parameters in forensics, and people who speak with a nasal twang are relatively easy to identify.

However, the above studies all investigated the effectiveness of nasal consonants together with the following vowels, rather than the effectiveness of nasal consonants alone. Unless each of the nasal consonants and vowels are excerpted from the syllables and tested separately, it remains unclear precisely to what extent nasal consonants contribute to SPID. This paper examines the contribution of nasals and vowels separately by conducting SPID experiments using nasals and vowels extracted from Japanese monosyllables. We will also focus on acoustic factors that may lead to erroneous identification.

<table>
<thead>
<tr>
<th>Table 1 Speech materials and experimental conditions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speakers</td>
</tr>
<tr>
<td>Recordings</td>
</tr>
<tr>
<td>Instruments</td>
</tr>
<tr>
<td>A/D conversion</td>
</tr>
<tr>
<td>Utterances</td>
</tr>
<tr>
<td># Tokens</td>
</tr>
<tr>
<td>Frames</td>
</tr>
<tr>
<td>Acoustic feature</td>
</tr>
</tbody>
</table>
2. SPID Experiments

2.1 Speech materials

The experimental conditions are summarised in Table 1. Monosyllables uttered by 50 male speakers were selected from the NRIPS (National Research Institute of Police Science) speech corpus. Recordings of the speech materials were conducted twice at 2- to 5-month intervals. The speech materials were recorded at 44.1 kHz sampling frequency with 16-bit resolution; all the materials were downsampled to 16 kHz before the experiments. The following 10 syllables recorded through a condenser microphone were used: /ma/, /me/, /mi/, /mo/, /mu/, /na/, /ne/, /ni/, /no/, and /nu/. Two tokens were recorded for each monosyllable in each recording session, for a total of four tokens for each speaker.

In order to examine the effects of nasals and vowels separately, we excerpted the nasal consonant and vowel portions from the recorded monosyllables. These excerpts were conducted manually on the basis of sound waveforms and spectrograms. The nasal-vowel boundary was determined by the zero-crossing point nearest to the end of the second formant transition to the following vowel.

2.2 Procedures

Automatic SPID experiments were conducted using seven segments consisting of two nasal consonants and five nucleus vowels excerpted from consonant-vowel monosyllables. We conducted two different experiments. In the first experiment, we used all of the available speech tokens obtained from the first recording session as the reference speech, and those from the second recording session as the input speech. In the second experiment we registered only one token from the first recording session (the one recorded in the first round out of the two repetitions) as the reference, and used all the tokens from the second recording session as the input. The number of registered reference tokens in the first experiment and input speech tokens in both experiments was 500 for nasals, corresponding to fifty speakers, two repetitions, and five following vowels and 200 for vowels, corresponding to fifty speakers, two repetitions, and two preceding nasals. The number of registered reference tokens in the second experiment was 50 for both nasals and vowels, and that of the input speech tokens was 500 for the nasals and 200 for the vowels. The procedure of the experiments is indicated in Fig. 1.

Following previously reported methods, we used 30th-order FFT (Fast Fourier Transform) cepstral coefficients as the acoustic feature. The cepstral coefficients were calculated every 10 ms using a 30-ms Hanning window and averaged over all frames. Next, we calculated the Euclidean distances of the cepstral coefficients between and within speakers. The referenced speech tokens were then ranked according to the cepstral distances, and the speaker of the token with the smallest distance was determined. When the speakers of the input and reference matched, we counted it as a correct identification. In the first experiment, we evaluated the results by the
rate of correct identification (percent matched) as a function of the percentage of referenced speech tokens over the total referenced speech tokens; in the second experiment, we evaluated the results as a function of the number of referenced speech tokens.

2.3 Results and discussion

Results of the two SPID experiments are shown in Fig. 2. The overall percentage of correct SPID was, as predicted, higher in the first experiment (79.3\%, when eight best candidates were referenced, averaged over seven segments) than in the second experiment (68.7\%, again, when eight best candidates were referenced, averaged over seven segments). In the first experiment, the number of reference speech tokens was ten per speaker for nasals and four per speaker for vowels. In the second experiment, in which the number of the reference speech tokens was equalized to one for both nasals and vowels, SPID rates decreased for both vowels (83.4\% to 76.7\%, when eight best candidates were referenced, averaged over five vowels) and nasals (69.1\% to 48.7\%, when eight best candidates were referenced, averaged over two nasals). As shown in Fig. 2(b), it is necessary to register more than six nasal tokens (42.3\% correct, averaged over two nasals) in order to yield the same SPID rate yielded by one vowel token (39.1\% correct, averaged over five vowels). Thus, the number of registered speech tokens influenced the SPID rate considerably.

Looking further at the results of the second experiment (Fig. 2(b)), it is clear that the SPID rate was much higher for vowels than for nasals. Among the five vowels, there was a tendency for high and front vowels, /u/, /e/ and /i/, to perform better than low back vowels, /a/ and /o/. This difference was found to be significant using Mann-Whitney’s $U$-test ($p = .025$). Vowels reported to be effective for SPID vary among experiments\textsuperscript{14,32}, perhaps due to speaker-set differences. As for nasals, the alveolar nasal /n/ was better than the bilabial /m/, corroborating the tendency identified in previous studies\textsuperscript{19,20}. Japanese has three places of articulation for oral and nasal stops: i.e., bilabial, alveolar and velar. Among these, alveolar consonants have physically the largest range of possible articulation, as the phonology of Japanese does not require phoneme contrasts in the coronal area for the stop consonants including nasals. This may lead to inter-speaker variations in articulation around the alveolar ridge. In addition, other studies have reported that bilabial nasal /m/ has greater intra-speaker variations than alveolar nasal /n/\textsuperscript{21,33}. These findings may explain the advantage of alveolar
Contribution of the nasality to speaker identification

3. Factor Analysis

3.1 Methodology

The experimental results were further subjected to factor analysis in order to investigate how and why confusions among the speakers occurred in our data. It is predicted that frequently confused speaker-pairs may share similar speech characteristics. In order to examine the effects of the physiological characteristics of the speakers and the acoustic properties of the speech on the confusion and differentiation among speakers, exploratory factor analysis was conducted. We analysed the vowels with respect to the following eleven parameters: speaker's age, height, weight, average fundamental frequency ($F_0$), jitter PPQ, shimmer APQ, local shimmer and jitter, the frequencies of the first and second formants ($F_1$ and $F_2$, respectively), and the amplitude difference between the first formant and the spectral peak caused by vowel nasalisation. The last acoustic parameter is suggested to indicate the degree of vowel nasalisation. The degree of nasalisation reflects the degree of velum lowering and this parameter is known to remain stable over time within a single speaker. The spectral peaks brought about by vowel nasalisation usually appear between the first and second formants in all vowels, and below the first formant in low and mid vowels.

Similarly, we conducted factor analysis for nasals with respect to the following seven parameters: speaker’s age, height, weight, average fundamental frequency ($F_0$), and the frequencies of the first, second and third formants ($F_1$, $F_2$ and $F_3$, respectively). The formants refer to the spectral peaks formed by the resonance of the nasal cavity.

The basic attributes of the speaker including age, height, and weight, were collected using a questionnaire distributed during the recording sessions. The acoustic properties of the speech samples were analysed by using Praat and MATLAB software. The average fundamental frequencies were calculated for the whole vowel portions using Praat’s auto-correlation method. The jitter (PPQ and local) and shimmer (APQ and local) were also obtained by Praat software. Local jitter and shimmer are the average absolute difference between consecutive periods and between amplitudes of the consecutive periods, divided by the average period and amplitude, respectively. Jitter PPQ and shimmer APQ are the average absolute difference between a period and the average of it and its four closest neighbours (five-point perturbation quotient). The vowel formants, $F_1$ and $F_2$, and nasal peaks were determined manually in the FFT spectra drawn by MATLAB. The formant frequencies of the nasals were calculated by Praat “To Formant” commands. Both vowel formants and nasal peaks were averaged values of the whole vowel portions.

We analysed the SPID results for three peripheral vowels, /i/, /a/, and /u/, and two nasals, /m/ and /n/. First, we identified the person for whom each of the 50 speakers had been most frequently mistaken for each segment. Then, we calculated the differences in the above-mentioned parameters between frequently mistaken speaker pairs. The parameters used in our analysis cover almost all of the important social and acoustic characteristics that are associated with speaker individuality; therefore, the factor solutions that gain higher communality values may explain differentiations and confusions among speakers.

The results of orthogonal (Varimax) rotation of the solution were summarised for each vowel and nasal as shown in Tables 2 and 3, respectively. Loadings less than ±0.45 were omitted in order to make interpretations easier. The analysis yielded four-, five-, and three-factor solutions for /i/, /a/, and /u/, respectively. For all three vowels, parameters loaded onto high-order factors identified fluctuations of the phonation. Two types of jitters and two types of shimmers loaded onto Factors 1 and 2 for /i/, onto Factors 1 and 3 for /a/, and both types of jitters and shimmers onto Factor 1 for /u/. Average $F_0$ together with $F_2$ loaded onto Factor 2 for the vowel /a/, although their communality
Table 2  Results of factor analysis for vowels. Orthogonally rotated component loadings for eleven parameters.

<table>
<thead>
<tr>
<th>Factors</th>
<th>/i/</th>
<th></th>
<th></th>
<th>/a/</th>
<th></th>
<th></th>
<th>/uu/</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>h^2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Age [years]</td>
<td>.53</td>
<td>.37</td>
<td>.88</td>
<td>.80</td>
<td>.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height [cm]</td>
<td>-.82</td>
<td>.10</td>
<td>.49</td>
<td>.64</td>
<td>.77</td>
<td>.66</td>
<td>-.47</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Weight [kg]</td>
<td>-.55</td>
<td>.14</td>
<td>.78</td>
<td>.61</td>
<td>.83</td>
<td>.73</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg. F0 [Hz]</td>
<td>.30</td>
<td>-.84</td>
<td>.88</td>
<td>.49</td>
<td>.53</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jitter local [%]</td>
<td>.90</td>
<td>.89</td>
<td>.93</td>
<td>.95</td>
<td>-.72</td>
<td>.55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shimmer local [%]</td>
<td>-.98</td>
<td>.99</td>
<td>.69</td>
<td>.75</td>
<td>-.82</td>
<td>.71</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shimmer APQ [%]</td>
<td>-.84</td>
<td>.82</td>
<td>.96</td>
<td>.99</td>
<td>-.80</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First formant [Hz]</td>
<td>.50</td>
<td>.26</td>
<td>032</td>
<td>-.54</td>
<td>.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second formant [Hz]</td>
<td>-.47</td>
<td>.23</td>
<td>-.80</td>
<td>.72</td>
<td>-.50</td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nasalisation [dB]</td>
<td>-.99</td>
<td>.99</td>
<td>014</td>
<td>.86</td>
<td>.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>2.04</td>
<td>2.01</td>
<td>1.62</td>
<td>1.26</td>
<td>2.00</td>
<td>1.99</td>
<td>1.80</td>
<td>1.64</td>
<td>1.36</td>
</tr>
<tr>
<td>Cumulative proportion</td>
<td>18.54</td>
<td>36.82</td>
<td>51.57</td>
<td>62.99</td>
<td>26.61</td>
<td>41.26</td>
<td>51.58</td>
<td>71.83</td>
<td>94.20</td>
</tr>
</tbody>
</table>

Table 3  Results of factor analysis for nasals. Orthogonally rotated component loadings for six parameters.

<table>
<thead>
<tr>
<th>Factors</th>
<th>/m/</th>
<th></th>
<th></th>
<th>/n/</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>h^2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Age [years]</td>
<td>.64</td>
<td>.42</td>
<td>.99</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height [cm]</td>
<td>-.78</td>
<td>.90</td>
<td>-.71</td>
<td>.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight [kg]</td>
<td>-.98</td>
<td>1.00</td>
<td>-.95</td>
<td>.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First formant [Hz]</td>
<td>-.11</td>
<td>.54</td>
<td>-.47</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second formant [Hz]</td>
<td>-.49</td>
<td>.14</td>
<td>-.87</td>
<td>.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third formant [Hz]</td>
<td>-.87</td>
<td>.54</td>
<td>-.47</td>
<td>.45</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvalue</td>
<td>1.60</td>
<td>.88</td>
<td>.62</td>
<td>1.56</td>
<td>1.25</td>
<td>1.01</td>
</tr>
<tr>
<td>Cumulative proportion</td>
<td>26.61</td>
<td>41.26</td>
<td>51.58</td>
<td>25.98</td>
<td>46.72</td>
<td>64.29</td>
</tr>
</tbody>
</table>

values were not as high as those of jitters and shimmers. Another latent variable may be the physical size of the speaker, which could affect the average F0 and formant frequencies, as body size is known to correlate with the length of the vocal folds and the size of the vocal tract. For high vowels /i/ and /u/, the degree of nasalisation loaded onto Factors 4 and 3, respectively, with the highest communality values for both vowels.

As shown in Table 3, three-factor solutions were obtained for the two nasals /m/ and /n/. The identified tendencies were almost the same for both nasals. The parameters related to the speaker’s body size loaded onto Factor 1, the speaker’s age loaded onto Factor 2, and the second and third formant frequencies loaded onto Factor 3.

To summarise, the factors responsible for confusion between speakers appear to include fluctuations in phonation, the degree of nasalisation, the physical size of the speaker, and...
the fundamental frequency, arranged according to the communalities.

4. General Discussion and Concluding Remarks

In this paper, we investigated the effects of the phonological contents of the reference speech on SPID accuracy using segments excerpted from the Japanese monosyllables. The contributions of consonants and vowels were examined separately by conducting SPID experiments using only the consonantal and vocalic portions. A traditional pattern match method was applied using the 30-th order cepstral coefficients.

In the first experiment, in which all available tokens obtained in the first recording session were registered as the reference speech, nasals showed an SPID rate similar to that of vowels. The number of registered reference speech tokens in the first experiment was 500 for nasals and 200 for vowels. However, in the second experiment, in which only one token was registered as the reference speech for both nasals and vowels, the correct SPID rates significantly decreased, especially for the nasals. We found that it was necessary to register more than six nasal tokens in order to yield the same SPID accuracy as that obtained using a single vowel token. The effectiveness of monosyllables containing a nasal consonant for SPID has been reported in previous research\textsuperscript{19-25}; however, in those studies, the nasal consonants were used together with the adjacent vowel(s). The present results suggest that individual speaker characteristics conveyed through vowel nasalisation are more useful for SPID than nasal resonance alone. In order to further clarify this point, future research should also use vowel portions excerpted from monosyllables with oral onsets.

The second half of the present paper focused on the investigation of error patterns among speakers. Based on the SPID results, data on frequently confused speaker pairs were extracted, and the value differences were calculated for several parameters related to the physiological characteristics of the speakers and the acoustic properties of the speech samples. The results of the exploratory factor analysis revealed that speakers with similar phonation properties, degree of nasalisation, physical size and fundamental frequency were frequently confused with one another.

Phonation properties such as jitter and shimmer are directly related to so-called voice quality. Recent studies report the availability of these measures in speaker recognition systems\textsuperscript{39,40}. In addition, the importance of the fundamental frequency has been pointed out in many SPID studies\textsuperscript{22,41}. As mentioned above, physical size correlates with the fundamental and resonance frequencies; taller speakers have longer vocal folds and a larger vocal tract than shorter speakers, resulting in lower fundamental and formant frequencies.

For the high vowels, /i/ and /u/, the degree of vowel nasalisation had a large influence on the confusion between speakers. However, vowel nasalisation was not found to have a strong effect for the low vowel /a/.\textsuperscript{42} Ohala explained this difference by stating that the lowering of the velum gives greater acoustic influence on the spectra of high vowels compared to those of low vowels. The degree of nasalisation has been reported to remain consistent within a single speaker over a period of several months\textsuperscript{36}. Future research on speaker individuality would benefit from acoustic analyses of vowel nasalisation. In order to use these measures effectively in forensics, it is necessary to fully understand their strengths and limitations.

5. Acknowledgement

This work was supported by a Grant-in-Aid for JSPS Fellows (22・3118) and a Grant-in-Aid for Scientific Research (C21510185) from MEXT.

6. References

1) Amino, K., Osanai, T., Kamada, T., Makinae, H., and Arai, T.: Historical and procedural overview of forensic speaker


24) Fattah, M., Ren, F., and Kuroiwa, S.:


