Acoust. Sci. & Tech. **30**, 2 (2009)

PAPER

Effects of linguistic contents on perceptual speaker identification: Comparison of familiar and unknown speaker identifications[1]

Kanae Amino* and Takayuki Arai†

Department of Electrical and Electronics Engineering, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

(*Received 10 December 2007, Accepted for publication 29 July 2008*)

Abstract: There are several factors that affect human speaker recognition. In this study, two experiments were conducted in order to see the effects that the stimulus contents and the familiarity to the speakers give to the perception of the speakers. The results showed that: a) stimuli including a nasal were effective for accurate speaker identification; b) coronal nasals were more effective than the labial nasal, and c) the familiarity to the speakers gives a great influence on the performance. The tendencies a) and b) were observed both in familiar and unknown speaker identifications. The results of the acoustical analyses also showed that there were correspondences between the perception of the speaker identity and the cepstral distances among the speakers. The inter-speaker cepstral distances were greater in vowel intervals than in the consonant intervals; especially, notably they were greater in nasals than in orals in the consonant intervals.

Keywords: Speaker identification, Speaker individuality, Familiarity, Nasals

PACS number: 43.71.Bp [doi:10.1250/ast.30.89]

1. INTRODUCTION

It is clear that human beings have the ability to identify the speakers by speech sounds alone. When we hear an utterance, we understand the contents of the speech and, at the same time, we perceive the identity of the speaker. In daily life, everyone may have the common experiences of perceptually identifying the relatives, friends, colleagues [1], actors, and radio and television personalities [2–4].

Though many people regard this ability as natural to human beings, the questions such as where this came from or to what extent this ability is reliable are not yet fully resolved. The ability of vocalic identification of individuals is also reported in nonhuman primates. Many of the group-living primates use contact calls in order to maintain the cohesiveness of the group [5]. Rendall *et al.* [6] showed that rhesus macaques, *Macaca mulatta*, have the ability to identify the kin and the individuals. Cheney and Seyfarth [7] reported that vervet monkeys, *Cercopithecus aethiops*, also recognise individuals vocalically. The primates do not only perceive but emit individuality in vocalisations. Masataka and Fujita [8] reported that Japanese macaques, *Macaca fuscata*, and rhesus macaques learn and signal species-specific calls. The acoustic features of caller individuality in nonhuman primates are pointed out in many animals [5], including Japanese macaques [9].

Vocal recognition in these species can be considered as a part of their social behaviour. Rendall *et al.* [6] says "the capacity for vocal recognition of individuals and kin may be an important adaptation facilitating social communication at a distance," and they needed this capacity in order to "regulate intra-group social behaviours." For the most part, human communication is thought to be similar to or have foundations on the vocal exchange of contact calls in nonhuman primates [5], and the ability of individual identification may also have evolved from the capacity of these nonhuman primates.

Individuality contained in human speech is used, for example, to gauge communicative settings [10], and thus perceptual speaker recognition is important for a successful communication in daily life as well as the perception of the utterance contents. In human communication, we perceive both the linguistic contents of the utterance and the identity of the speaker, and this means that speech sounds convey

[1]This paper is a summary of our previous works; Amino *et al.* "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *AST*, **27**(4), pp. 233–235 (2006), and Amino and Arai "Effects of stimulus contents and speaker familiarity on perceptual speaker identification," *AST*, **28**(2), pp. 128–130 (2007). Additional analyses are also introduced.

*e-mail: amino-k@sophia.ac.jp

†e-mail: arai@sophia.ac.jp

both linguistic or phonological information and speaker information. Many studies have attempted to find corresponding acoustic features for these two types of information, as these features would enable us to improve the speech technologies such as automatic speech recognition and automatic speaker recognition, by extracting the phonological information and speaker information, respectively [11]. The study of the acoustic correlates that reflect individualities can also contribute to the linguistic or phonetic theory, as it leads to a clear definition of the phonemes and to a more specific modelling of the sound patterns of the language.

When we assume that the human percepts can be explained by some acoustic features, the acoustic correlates for speaker information should be measured as acoustic parameters [12]. Formant frequencies of the vowels [13–15], long-term average speech spectra [16] and average pitch frequency are said to indicate speakers' individuality [15,17,18], just to name a few. O'Shaughnessy [19] suggests that one way to find the acoustic parameters that indicate the speaker individuality is to conduct perceptual speaker identification tests where various kinds of sounds are examined. The parameter that is crucial for correct speaker identification is regarded as indicating speaker information.

Investigation of speaker identification by human perception can contribute not only to clarifying the acoustical properties for speaker identity, but also to understanding the mechanism of human perception and to the forensic sciences [20–22]. It is important to understand the properties of human perception of speaker identity when we estimate the validity of a testimony given by an ear-witness in the court. Speech can be used as evidence, for instance, in the case of kidnapping, stalking, robbery or any other crimes where speech data are available by telephone recordings or in the anti-holdup cameras [23–25].

It is pointed out that human identification of familiar speakers is not always perfect, though highly accurate [20,21,26–28]. Some factors are known to affect the accuracy of the identification.

One of these factors is the change in the laryngeal source of speech. This is the most popular factor that the speaker may alter when he or she wants to disguise a voice. Pollack *et al.* [29] found that the whispered speech greatly reduces the familiar speaker identification rates, and in order to obtain the equivalent scores of identification, the duration of the whispered speech must be three times as long as the modally phonated speech. Orchard and Yarmey [30] and Yarmey *et al.* [28] showed that the identification performance on whispered speech fell significantly also in unknown speaker identification task.

The importance of the fundamental frequency in speaker identification is reported in many studies, including the studies on automatic speaker recognition [17,18]. Hashimoto *et al.* [31] reported that the average fundamental frequency is the most effective parameter in unknown speaker identification. Kitamura and Mokhtari [32] and Kitamura and Saito [33] showed that normalisation of the pitch frequency gave rise to the significant decrease in identification performance.

On the other hand, Coleman [34] conducted a speaker identification experiment with nil-phonated speech. Instead of modal phonation, the speakers were instructed to use an electric buzzer as the laryngeal source. Though there were no laryngeal features conveyed by stimuli, the listeners could identify the five speakers more than 90 percent correct. These findings suggest that there is much information remained even if the individualities in phonation are taken away, and the resonance features of the speech sounds are enough for the identification.

Another factor on which the identification accuracy is dependent is the duration and the contents of the stimuli. As pointed out by Nygaard [10, p. 393], there is an interaction between perception of linguistic contents and that of speaker identity, although these two kinds of information are processed separately in the brain. Pollack *et al.* [29] showed that the identification rates of the familiar speakers would increase as the function of the stimulus duration, but it gets saturated at around 1.2 s. In Bricker and Pruzansky [35] and Roebuck and Wilding [36], the relationship between the duration and the phonemic variation of the stimuli was examined. Both studies concluded that the identification rate increased with duration only if the longer stimuli contained more phonological variation.

Most of the studies that examined the differences among the phonemes in the effectiveness for identifying the speakers report that the vowels and the voiced sonorants obtained higher scores than the voiceless consonants or the obstruents [35,37–41]. Specifically, the identification performance was significantly better when the nasal sounds were presented than when the stimuli contained only oral sounds [37]. The tendency that voiced sonorants are more effective for identifying the speakers is also seen in automatic speaker recognition [18,42,43].

Familiarity with the speakers also affects the identification performances. Van Lacker *et al.* [3,4] showed that identification of familiar speakers and that of unknown speakers go through different processes, i.e. the former is rather like pattern recognition, while the latter is more like feature analysis. Hashimoto *et al.* [31] claimed that the contribution of the acoustic parameters, such as spectral information, the fundamental frequency, and tempo information, would differ according to whether the listeners are identifying familiar speakers or unknown speakers. Correlation between the subjective estimation of the familiarity to the speakers and the accuracy of speaker

identification was reported in Schmidt-Nielsen and Stern [1,44].

This present study integrates the two perceptual speaker identification tests conducted by the authors previously, in order to investigate the differential effects of the stimulus contents on the identification accuracy and to compare the identification performances between the identifications of familiar and unknown speakers [45,46]. We will also look into the interaction between the stimulus contents and the listeners' familiarity to the speakers.

In the first experiment [45], familiar speakers were identified, and in the second experiment same speakers were identified by the listeners who were previously unfamiliar with them [46]. In the analysis of the results, we focused on the effects of the stimulus contents, and found that the stimuli containing a nasal sound were more effective for speaker identification than the stimuli without a nasal. After the perception tests, we inspected the spectral properties of the stimuli used in both experiments in order to explain the differences in the perception tests. The results showed that the cepstral distances among the speakers were greater in nasal sounds than in oral sounds, and also the nasal sounds had longer intervals that listeners may exploit for speaker identification than oral sounds did.

2. EXPERIMENTS

2.1. Expeiment 1: Identification of Familiar Speakers

2.1.1. Participants and materials

Fifteen male students at Sophia University participated in this experiment. They had lived in the same dormitory for more than four years, thus they had known to each other very well. Ten of them served as the speaker, and the rest as the listeners. They were all native speakers of Japanese without any strong accents or speech disorder. None of them had known hearing impairments.

In order to examine the effects of the phonemic variations on speaker identification accuracy, a variety of stimuli should be used in the test. At the same time, the test time should not be too long in consideration of the burden for the listeners. We selected nine consonants of Japanese that are articulated in the coronal region, /d/ /t/ /z/ /s/ /j/ /ɾ/ /m/ /n/ and /ɲ/. These consonants were combined with the vowel /a/ to be put into non-sense words for the recordings. We used only one vowel /a/ to make the experiment simple.

The recording sessions were held in a soundproof room. The speakers uttered the non-sense words, 'aCaCaCa,' where 'a' stands for the vowel /a/ and 'C' stands for one of the nine consonants described above. These non-sense words were embedded in the carrier sentence: /'aCaCaCa' to: o ʃiʒi ʃimasɯ̥/ (I support the 'aCaCaCa' political party). We assumed the word 'aCaCaCa' to be the name of a fictional political party, because the suffix "–to: (-party)" forms compound words that are uttered with relatively flat pitch contour after the third mora [47,48]. In this experiment, the last, or the fourth, morae of the names of the parties were excerpted manually, and the excerpted monosyllables were used as the stimuli. All the speech data were adjusted in amplitude so that their maximal values become 90 percent of the full range.

Ten repetitions for each monosyllable were recorded onto the digital audio tape, at the sampling frequency of 48 kHz with 16 bit resolution, and five tokens that were uttered most clearly were used in the experiment.

Sample sentences of each speaker were also recorded and were used in the practice sessions. These were different from the sentences used in the experiments. The speakers uttered a sentence, /hoɴ.dʒi.tsɯ wa sei.teɴ na.ɾi/ ("It is fine today") for five times, and two of them were selected to be used in the experiment.

2.1.2. Procedure

The experiment was conducted in the same soundproof room as the recordings. The stimuli were presented on a computer and the participants listened to the following stimulus monosyllables through headphones: /da/ /ta/ /sa/ /za/ /ja/ /ɾa/ /ma/ /na/ and /ɲa/. The total number of the stimuli was 450 (ten speakers, nine syllables and five tokens) and they were presented in a random order using a Praat Multiple Forced Choice experiment programme [49]. The stimuli were played automatically, once the listeners answered the previous trial. They answered a speaker to whom they thought the stimulus belonged to. We did not create a "replay button," so the listeners listened to one stimulus for only once, and the experiment was conducted at self-pace.

Before the test began, the listeners were informed of the names of the candidate speakers, and listened to each speaker's sample sentences shown above, which were different from the sentences recorded for test stimuli. Then they practised with sample sentences for several times. They were instructed to answer the name of the speaker for each trial in the way of multiple forced choice. Once the experiment started, the listeners were not allowed to listen to the sample speech any more.

2.1.3. Results

Percentages of the correct identification for each stimulus are shown in Table 1. The number of evaluation for each stimulus was 250 (ten speakers, five tokens and five listeners). The main effect of the stimulus contents was not significant in ANOVA, an analysis of variance ($p = 0.11$), except that the nasals /na/ and /ɲa/ were significantly better than the stops /ɾa/ and /ta/ ($p < 0.05$).

The nasal sounds, /na/, /ɲa/ and /ma/, were also significantly better than other oral sounds in the t-test ($p < 0.01$). Voiced consonants gained higher scores than

Table 1 Percent correct for each stimulus in identification of familiar speakers (the first experiment). $N = 250$.

Stimulus	Percent correct (%)
/na/	86.0
/ɲa/	85.6
/ma/ /za/	80.8
/sa/	78.8
/ja/	78.4
/da/	78.0
/ɾa/	74.4
/ta/	73.6

the voiceless counterparts, although the difference was not significant ($p = 0.36$). As to the manners of articulation, fricatives (/sa/ and /za/) followed nasals, and oral stops (/da/, /ɾa/ and /ta/) ranked the lowest.

2.2. Experiment 2: Identification of Previously Unknown Speakers

2.2.1. Participants and materials

Out of the ten speakers participated in the first experiment, four speakers were selected and their speech data were used again in the second experiment. The selection of the speakers was based on the resemblance of the average fundamental frequency, taking into account the report that the fundamental frequency has large effects on unknown speaker identification [26,32,33], and also that our aim was to see the articulatory properties of the speakers rather than the phonation properties. The average fundamental frequency of all the stimuli of the four speakers was 109.9 Hz ($N = 160$, i.e. 40 utterances for each speaker, $S.D. = 7.7$) with the range of 102.4 Hz to 118.5 Hz.

Sixteen university students who had never known any of the speakers served as the listeners. They were all native speakers of Japanese and had normal hearing.

The stimuli used in this experiment were identical to those used in the first experiment, but only of the four speakers described above. Nine monosyllables excerpted from carrier sentences were again presented to the listeners in a random order through headphones. The number of the tokens for each speaker was also the same, i.e. five tokens for each monosyllable.

2.2.2. Procedure

All the test sessions were conducted in the same soundproof room as the first experiment. The listeners went through the familiarisation session first. They listened to the sample sentences of the four speakers, /hoɴ.dʒi.tsɯ wa sei.teɴ na.ɾi/ ("It is fine today"). The listeners could listen to these sample sentences as many times as they wanted, but these sentences were always presented as one set of the four speakers, i.e., the listeners were not allowed to listen to the utterance of a particular speaker for many times. The speakers were introduced using speaker IDs, from number 1 to number 4, not by their names.

After the listeners showed some confidence, practice sessions were carried out using these sample sentences. Sentence uttered by one of the speakers was presented, and the listeners answered the speaker ID. Feedback was given after each trial. After this we moved on to the second practice session, where they identified the speakers by sample monosyllables, which were different from those used in the test session. These monosyllables consisted of the consonants that are different from those used in the test session and the vowel /a/. Feedback was given after every trial in the practice session.

We repeated familiarisation and practice sessions until the listeners achieved more than 90 percent correct identification in both practice sessions. It took the listeners 10 to 20 minutes before they reached the desired accuracy.

In the test session that was conducted immediately after the practices, the listeners answered the speaker ID for each stimulus. The stimuli were presented to the listeners only once, using the same Praat programme as in Experiment 1, and no feedback was given in this session. The total number of the stimuli was 180 (four speakers, nine syllables and five tokens). The listeners were not allowed to go back to listen to the sample sentences during the test session. They took break after every 90 trials.

2.2.3. Results

The percentages of the correct speaker identification are shown in Table 2. The number of evaluation for each stimulus is 320 (four speakers, five tokens and sixteen listeners). All the stimuli gained higher scores than the chance level (25%).

As was in the first experiment, the nasals /na/ and /ɲa/ ranked the highest, and fricatives and oral stops followed. The nasal /ma/ did not obtain as high score as in the first experiment, on the other hand, /ja/ ranked higher. In the analysis of variance, the effect of the stimuli was significant ($p < 0.01$). There were significant differences between the nasal /na/ and the oral stops /da/, /ta/ and /ɾa/.

Table 2 Percent correct for each stimulus in identification of previously unknown speakers (the second experiment). $N = 320$.

Stimulus	Percent correct (%)
/na/	59.0
/ɲa/	53.67
/ja/	52.0
/sa/	50.0
/ma/	49.67
/za/	46.67
/da/	46.33
/ta/	45.67
/ɾa/	44.67

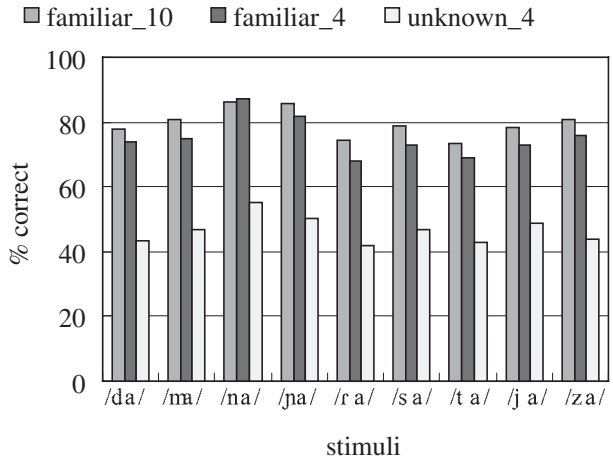

Fig. 1 Results of the two experiments: percent correct for each stimulus syllable in speaker identification tests. Familiar speakers (the first experiment) and unknown speakers (the second experiment). The middle bar shows the results of the first experiment, but only as for the four speakers whose speech materials were used in the second experiment.

2.3. Discussion for the Two Experiments

2.3.1. Effect of the stimulus contents

The results of the two experiments are shown in Fig. 1. The graph indicates the percent correct for each stimulus. The left bar shows the results for the first experiment (ten speakers, $N = 250$); the central bar shows those of the first experiment again, but only the results for the 4 speakers are included whose speech materials were used in the second experiment are included (four speakers, $N = 100$); the right bar shows the results for the second experiment (four speakers, $N = 320$). Here note that the listeners are different for the central and the right bars, although the speech materials are identical.

As can be seen, the tendencies of the three bars are quite similar, although the identification scores are lower in the second experiment. The effect of the speaker familiarity was significant in ANOVA ($F\ (1, 17) = 803.4$, $p < 0.001$), with familiar speaker identification performance ($Mean = 75.2\%$, $S.E. = 0.019$) being better than unknown speaker identification ($Mean = 46.6\%$, $S.E. = 0.014$). Also the effect of the stimulus contents was significant ($F\ (8, 17) = 10.9$, $p < 0.001$). The nasal /na/ gained significantly higher score than any other stimuli, and /ɲa/ was significantly better than /da/, /ta/ and /ɾa/.

Poor performance in unknown speaker identification is an expected outcome, just as reported in previous studies [20,21]. As mentioned above, this is because familiar and unknown speaker identification tasks undergo different cognitive processes, and the process for unknown speaker identification is a more difficult one [3,4,10].

The overall tendencies as to the stimulus contents were similar for both familiar and unknown speaker identification tasks. This means that no interaction between the stimulus contents and familiarity was seen as for the results of this experiment, though difference in cognitive processing is pointed out in these two tasks. In both experiments, the nasals were more effective for speaker identification than the oral sounds, with the alveolar nasals being better than the bilabial. And when we focus on the manners of articulation, the nasals ranked the highest, then the fricatives and the plosives, or the oral stops, followed them. This ranking of the consonants coincide with the ranking in the sonority scale [50]. The more sonorous a consonant is, the more effective it is for perceptual speaker identification. Generally speaking, sonorous consonants tend to be voiced, thus these sounds contain not only articulatory properties, but also the properties of the sound source created at the vocal folds. As to the individualities contained in sound source, voiced sounds are reported to be effective for perceptual speaker identification [40], although speakers' individual differences mainly lie in the vocal tract properties [34,51,52].

Apart from the manners of articulation, there were few differences. An approximant /ja/ ranked higher in unknown speaker identification (Experiment 2) compared to the result of familiar speaker identification. On the contrary, /ma/ did not obtain as high score as in Experiment 1.

The effectiveness of the nasals in speaker identification can be explained by the uniqueness of the morphology of the resonators. It is reported that the shapes of the nasal cavity and paranasal sinuses are different among individuals [53]. Also, the shapes of these resonators cannot be altered voluntarily. Differences in the timing of the velic action may be another factor that differentiates the nasals from oral sounds [54], and this is also something that the speakers cannot intentionally or voluntarily control by themselves. This is why the acoustical properties of the nasal sounds are of relatively stable nature, and thus stably reflects speaker's individuality.

As to the places of articulation, alveolar consonants were better in the scores than bilabials. This tendency is consistent with the results in previous experiments [55]. Variations in the nasal production are reported in Fujimura [56]. He suggested that labial /m/ has greater intra-speaker variations compared to coronal /n/. Data in Su *et al.* [43] also support this claim.

2.3.2. Speaker factor: The scores for each speaker

Speaker identification scores as for each speaker are shown in Figs. 2 and 3. Figure 2 shows the identification performances for each speaker in Experiment 1, and Fig. 3 shows the performance differences for speakers #1 to #4 in Experiments 1 and 2. The effect of the speaker factor was significant in ANOVA in both experiments ($p < 0.001$). Matsui *et al.* [39] suggests that the sounds effective for

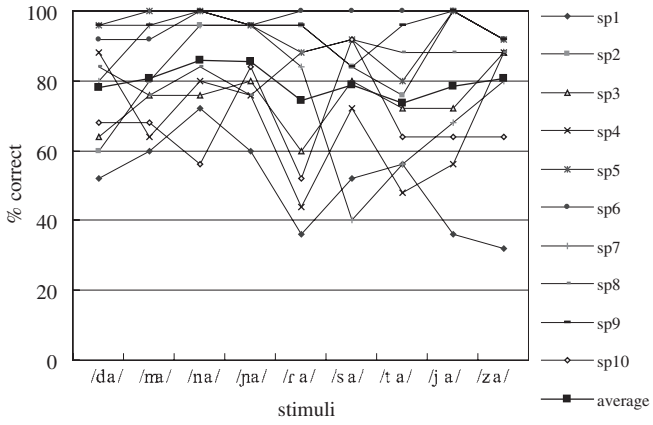

Fig. 2 Results of the first experiment: percent correct for each stimulus and for each speaker.

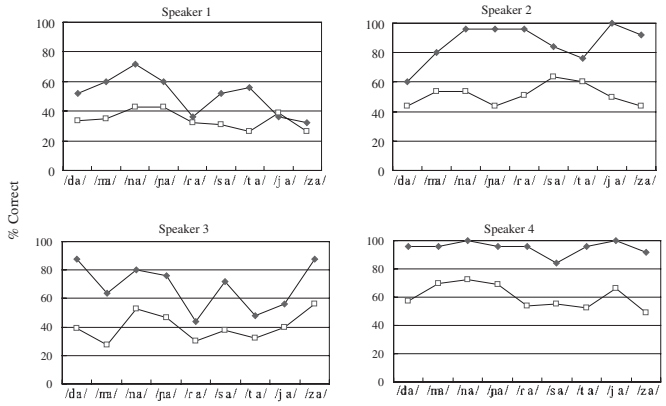

Fig. 3 Identification results of each speaker in Experiments 1 (diamonds) and Experiment 2 (outline squares): (a) Speaker 1, (b) Speaker 2, (c) Speaker 3, and (d) Speaker 4.

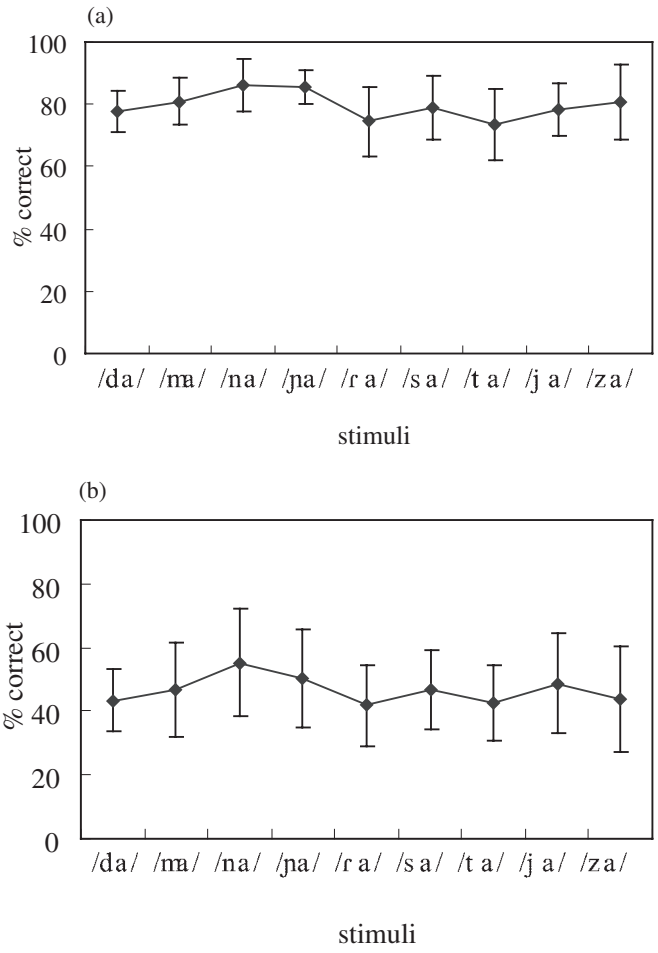

Fig. 4 Identification performances among the listeners, average percent correct identifications and the standard deviations of each listener according to the stimuli: (a) Experiment 1; (b) Experiment 2.

identifying the speakers are different for each speaker. In this study, too, the differences were seen among speakers. Nasals were not necessarily the most effective sounds for all of the speakers, for example for Speaker #3, but we can say that the nasals were relatively effective for most of the speakers.

As for the differences in the consonant rankings of Experiments 1 and 2, especially for the difference between the central and the right bars in Fig. 1, Bricker and Pruzansky [57] reported that the identification results can vary according not only to the speakers but also to the speaker ensembles where they are being compared to each other.

2.3.3. Listener factor

Differences in accuracy rate among listeners are shown in Fig. 4. Results of ANOVA showed that the differences in the performance among the listeners were significant.

It is pointed out that the ability to identify speakers is dependent on the individual listener [57], though the listener group of more than 12 people is said to be of typical size to obtain homogeneous data [58] (reviewed in [57]). In this study, the average identification rate of the 5 listeners in Experiment 1 (familiar speakers) was 79.6% with the range from 67.1% to 89.1%, whereas that of 16 listeners in Experiment 2 (previously unknown speakers) was 46.6% ranging from 35.0% to 65.0%. The average identification rates of each listener are shown in Fig. 5. Individual differences in the scores may also come from the differences in the strategies that the listener applies when identifying speakers or the differences in the priorities of the acoustical cues.

Though there are some differences, the graph shows that identifications were relatively more accurate when in the coronal nasals /na/ and /ɲa/ were presented as the stimuli than in the case of other non-nasal sounds. Differences in scores derive from the differences in the ability to identify people by their voices.

3. ACOUSTICAL ANALYSES

In order to inspect the relationship between the results

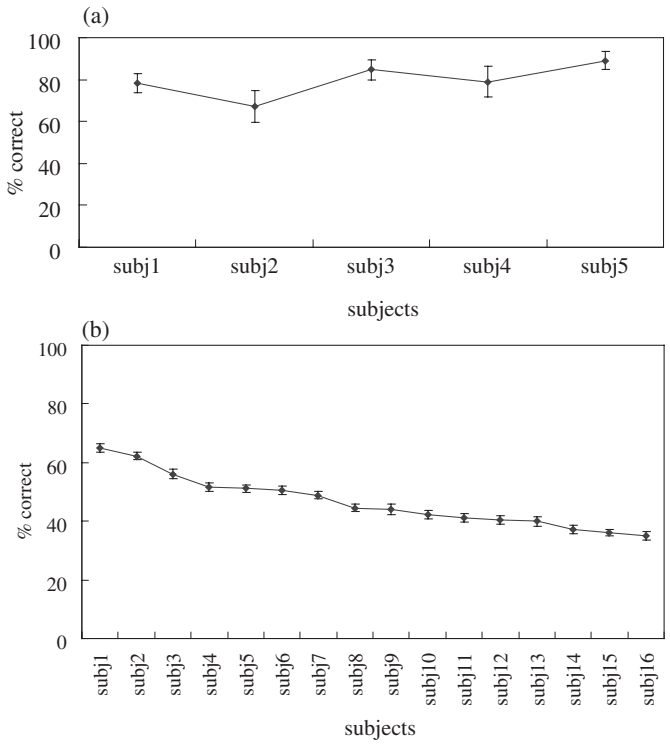

Fig. 5 Identification performances of each listener: (a) the first experiment, and (b) the second experiment.

Table 3 Excerpted intervals.

Name of interval	Description
Interval-C	Stable consonant part, /t/ and /d/ omitted
Interval-C(V)	Consonant and transition (until the second formant gets stable in the following vowel)
Interval-(C)V	Vowel part including transition (includes formant transitions)
Interval-V	Stable vowel part

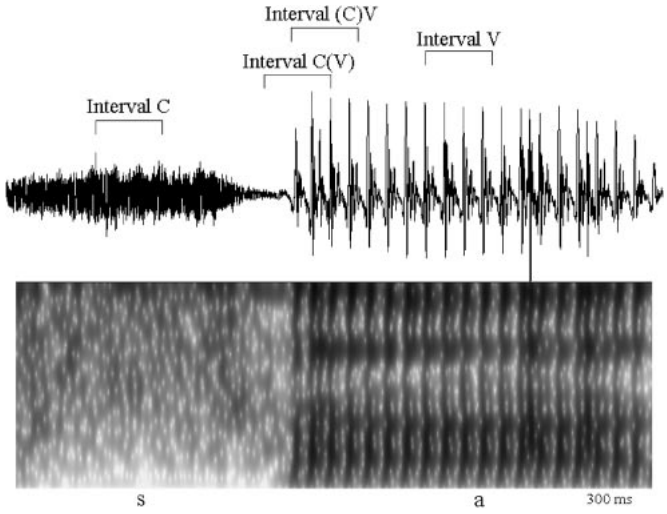

Fig. 6 Example of interval excerption: waveform and spectrogram of a sample /sa/ uttered by speaker 1.

of the perceptual speaker identification and the acoustical properties of the stimuli, two kinds of analyses were performed. In the first analysis, we aimed to see the spectral distances among the speakers in order to explain the identification differences among the stimuli. Cepstral distances were calculated and compared within and among the speakers. Here we used the ratio of intra- and inter-speaker distances as the measure. This measure is based on the concept of the *F*-ratio [11], which we use as a baseline when looking for a useful parameter in speaker recognition, thus we call this analysis "*F*-ratio analysis." In the second analysis, or "confusion analysis," our goal was to see the perceptual confusions among the speakers and to determine the interval(s) that are important for perceptual speaker identification. Following sections explain the targets and the methodologies of the analyses.

3.1. Analysis Targets

Four analysis intervals were excerpted from the following six monosyllables uttered by ten speakers: /ta/ /da/ /ma/ /na/ /sa/ and /za/. Each stimulus had five tokens for each speaker. The interval length was 30 ms, and the criteria and an example of excerption are shown in Table 3 and Fig. 6, respectively. We did not get interval C for /ta/ and /da/, as they were just silence in some utterances. Several intervals were overlapped with preceding or following interval(s) in some of the samples. As for the remaining syllables, /ja/, /ra/ and /ɲa/, we omitted from the analysis targets, since these sounds are realised as momentary or gliding sounds in Japanese, and therefore it is hard to define the boundary of the consonant and vowel. All the excerptions were conducted manually based on the waveforms and spectrograms.

In both analyses, we used 30th order FFT cepstrum as the analysis parameter. The zero-th coefficient was excluded here. Intra- and inter-speaker cepstral distances were further computed for every possible pair of five tokens of a speaker and of ten speakers, respectively. Thus we obtained 50 by 50 square matrices for each monosyllable.

3.2. Analysis Methods

3.2.1. Analysis 1: *F*-ratio analysis

In this analysis, ratios of averaged intra-speaker distances to averaged inter-speaker distances were calculated. Larger inter-speaker distance and smaller intra-speaker distance are most probably representing speaker individualities, and greater *F*-ratio values are desirable for speaker identification purposes. As mentioned above, this analysis is based on the measure called *F*-ratio [11], which is the ratio of inter-group co-variation to intra-group co-variation, and can indicate the effectiveness of an acoustic parameter in speaker recognition. We used *F*-ratio values as an acoustical measure indicating speaker characteristics.

3.2.2. Analysis 2: Confusion analysis

The purpose of the second analysis was to inspect the relationship between perceptual speaker similarities and acoustical properties of the stimuli in more detail. We analysed the perception patterns, that is, confusions among the speakers, observed in the first experiment. In order to examine the perception of speaker identity in relation to the inter- and intra-speaker cepstral distances, confusion patterns were analysed by drawing confusion matrices among the ten speakers.

We made six 10 by 10 square matrices for confusions on each of the six monosyllables and also twenty four cepstral-distance matrices, that is corresponding to four intervals for six monosyllables. Then we calculated the correlation coefficients between the confusion matrices and the distance matrices. Here we compared matrices only for the same monosyllables.

3.3. Results and Discussion

3.3.1. Analysis 1: *F*-ratio analysis

Table 4 shows the average *F*-ratio values for each interval of the stimuli. We can see that the nasal sounds obtained high values all through the four intervals. On the other hand, other oral sounds gained relatively higher values only in the vowel part of the stimuli. This means that nasal sounds have longer interval that indicates speaker individuality. We also notice that non-nasal, or oral syllables showed relatively high ratio values in intervals (C)V and V.

In fricatives, intervals C and C(V) obtained higher scores in the voiced consonants than in the voiceless counterparts. This is because voiced sounds contain source information as well as resonance information that does not accompany in the production of voiceless sounds.

3.3.2. Analysis 2: Confusion analysis

The results of the confusion analysis are shown in Table 5. As can be seen, correlations between the perception and the spectral properties were observed in all the intervals in the nasals, but only in the vowel intervals, (C)V and V, in oral sounds.

Table 4 Ratios of inter-speaker to intra-speaker cepstral distances (averaged distances among ten speakers and among five tokens of each speaker, respectively).

Stimuli		Interval-C	Interval-C(V)	Interval-(C)V	Interval-V	Avearge
Nasals	/ma/	2.35	2.26	2.22	2.30	2.28
	/na/	2.08	2.06	2.20	2.21	2.14
Fricatives	/sa/	1.45	1.54	2.05	2.24	1.82
	/za/	1.55	1.55	2.05	1.99	1.79
Stops	/ta/	N/A	1.15	2.06	2.11	1.77
	/da/	N/A	1.15	1.95	1.95	1.68

Table 5 Correlation coefficients between the perception of the speakers and the spectral distances within and among the speakers.

Stimuli		Interval-C	Interval-C(V)	Interval-(C)V	Interval-V
Nasals	/ma/	−0.81	−0.79	−0.75	−0.67
	/na/	−0.79	−0.77	−0.62	−0.63
Fricatives	/sa/	−0.38	−0.38	−0.66	−0.69
	/za/	−0.33	−0.33	−0.64	−0.58
Stops	/ta/	N/A	−0.31	−0.60	−0.57
	/da/	N/A	−0.34	−0.64	−0.63

This leads to the following two implications: first, in stimuli containing a nasal sound, listeners use all four intervals as the cue to identify speakers; and in oral sounds, both stops and fricatives, listeners tend to use only the vowel part as the speaker cue.

4. GENERAL DISCUSSION

In this study, we carried out two perceptual experiments, in order to see the effects of the stimulus contents on perceptual speaker identification. The results showed that the stimuli including a nasal sound were effective for both familiar and unknown speaker identification by listening, and the rankings of the stimuli coincided with the sonority scale of the consonants. No studies have pointed out the effectiveness of the nasals despite the difference in familiarity to the speakers.

Also, we compared the spectral properties of the stimuli by calculating intra- and inter-speaker cepstral distances. We found that the inter-speaker distances were greater in nasal sounds than in oral sounds. Furthermore, we analysed the inter-speaker distances for four intervals that temporally ranged from the onset consonant part to the stable vowel part, and we found that all the intervals correlated with the perception of the speaker identity in the stimuli containing a nasal, but not in the stimuli of only oral sounds. In spite of the significant difference in performances of familiar and unknown speaker identifications, this tendency was observed in both familiar and unknown speaker identification. Although it has been reported that the nasals are effective for speaker identification [42], the correlations between the perceptual confusions among speakers and the inter-speaker spectral distances in nasal sounds were not mentioned.

The acoustical properties of the nasal sounds are speaker-dependent, because they are produced with the resonators whose morphology differs considerably among speakers. Especially the shapes of the paranasal sinuses are known to be quite complex and speaker-specific [51,59]. In addition, the shapes of these resonators cannot be changed at speaker's will. This means that the resonance

properties of the nasals rarely change. Morphologies of the nasal cavity and other peripheral cavities cannot be measured easily, but the results of this study suggest that the morphological differences among the speakers are well reflected in the spectral properties and can be perceived by listeners.

Also, the vowel following a nasal consonant is necessarily nasalised to some extent, and nasalised vowels are predicted to contain more individuality than non-nasal vowels. This explains the results that the monosyllables containing a nasal consonant obtained high accuracy in speaker identification. Amino *et al.* [55] reported that the Japanese coda nasal /ɴ/ was also effective for perceptually identifying the speakers. This sound is categorised into the uvular sound, but Hashi *et al.* [60] reports that the place of articulation varies among speakers. We can conclude that the syllable with both an onset and a coda nasal may best reflect speaker's individuality.

As to the place of articulation, coronal nasals /n/ and /ɲ/ were better than labial /m/. This can be explained by that labial nasal has more intra-speaker variations compared to the coronal nasals [56]. Co-articulations to the following vowel are also greater in /m/ than in /n/ [43].

It is interesting to note that all the sounds that are effective for identifying individuals, i.e. nasals, vowels and coronals, are the unmarked ones in language typology and in language acquisition [61–63]. We can say that unmarked sounds bear speaker variations because they are less responsible for phoneme contrasts.

Nasal resonance can be an effective cue to speaker identification, as their properties are relatively stable compared to the fundamental frequency or other phonation parameters. On the other hand, their resonance properties are influenced seriously by head-cold [64], hay fever and other nasal diseases as the transmission through the nasal tract may be affected. This kind of problem must be solved in the future.

The final goal of this research is to understand the interaction between the phonological information and the speaker information conveyed by speech sounds, and to find out the mechanism that human beings identify individuals by speech. In this study, we found that nasal sounds contain more individuality than oral sounds. However, the problem of variations of effective sounds among speakers is pointed out in other studies [3,39], and thus the results obtained in this study may not be general. Our future task is to examine the question whether nasal sounds are effective for all the speakers and for all the listeners. Furthermore, taking into account the effects of co-articulation and other contextual factors, the combinations with other vowels than /a/ should also be examined in future experiments.

ACKNOWLEDGEMENT

This work was supported by a Grant-in-Aid for Scientific Research (A) 16203041, by a Grant-in-Aid for JSPS Fellows (17·6901), and by Sophia University Open Research Centre from MEXT.

REFERENCES

[1] A. Schmidt-Nielsen and K. Stern, "Identification of known voices as a function of familiarity and narrow-band coding," *J. Acoust. Soc. Am.*, **77**, 658–663 (1984).

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

[3] D. Van Lacker, J. Kreiman and K. Emmorey, "Familiar voice recognition: patterns and parameters part 1: recognition of backward voices," *J. Phonet.*, **13**, 19–38 (1985).

[4] D. Van Lacker and J. Kreiman, "Familiar voice recognition: patterns and parameters part 2: recognition of rate-altered voices," *J. Phonet.*, **13**, 39–52 (1985).

[5] H. Sugiura, "Vocal exchange of coo calls in Japanese macaques," in *Primate Origins of Human Cognition and Behaviour*, T. Matsuzawa, Ed. (Springer, Tokyo, 2001), pp. 135–154.

[6] D. Rendall, P. Rodman and R. Emond, "Vocal recognition of individuals and kin in free-ranging rhesus monkeys," *Anim. Behav.*, **51**, 1007–1015 (1996).

[7] D. Cheney and R. Seyfarth, "Vocal recognition in free-ranging vervet monkeys," *Anim. Behav.*, **28**, 362–367 (1980).

[8] N. Masataka and K. Fujita, "Vocal learning of Japanese and rhesus monkeys," *Behaviour*, **109**, 191–199 (1989).

[9] M. Mitani, "Voiceprint identification and its application to sociological studies of wild Japanese monkeys (*macaca fuscata yakui*)," *Primates*, **27**, 397–412 (1986).

[10] L. Nygaard, "Perceptual integration of linguistic and non-linguistic properties of speech," in *The Handbook of Speech Perception*, D. Pisoni and R. Remez, Eds. (Blackwell Publishing, Oxford, 2005), pp. 390–414.

[11] Y. Niimi and T. Sakai, Eds., *Speech Recognition* (Kyoritsu Shuppan Publishing Company, Tokyo, 1979) (in Japanese).

[12] G. Fant, *Acoustic Theory of Speech Production* (Mouton, The Hague, 1960).

[13] K. Itoh and S. Saito, "Effects of acoustical feature parameters of speech on perceptual identification of speaker," *Trans. IEICE*, **J65-A**, 101–108 (1982) (in Japanese).

[14] P. Ladefoged and D. Broadbent, "Information conveyed by vowels," *J. Acoust. Soc. Am.*, **29**, 98–104 (1957).

[15] H. Matsumoto and Y. Yamashita, "Unsupervised speaker adaptation from short utterances based on a minimized fuzzy objective function," *J. Acoust. Soc. Jpn.* (*E*), **14**, 353–361 (1993).

[16] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," *Speech Commun.*, **5**, 183–197 (1986).

[17] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion," *Speech Commun.*, **16**, 165–173 (1995).

[18] M. Sambur, "Selection of acoustic features for speaker identification," *Proc. IEEE Trans. Acoust. Speech Signal Process.*, **23**, 176–182 (1975).

[19] D. O'Shaughnessy, *Speech Communication —Human and Machine—*, 2nd ed. (Addison-Wesley Publishing Company, New York, 2000).

[20] H. Hollien, W. Majewski and E. Doherty, "Perceptual identification of voices under normal, stress, and disguise speaking conditions," *J. Phonet.*, **10**, 139–148 (1982).
[21] H. Hollien, *Forensic Voice Identification* (Academic Press, San Diego, 2002).
[22] P. Ladefoged and J. Ladefoged, "The ability of listeners to identify voices," *UCLA Work. Pap. Phonet.*, **49**, 43–89 (1980).
[23] H. Hollien, *The Acoustics of Crime: The New Science of Forensic Phonetics* (Plenum Press, New York, 1990).
[24] F. McGehee, "The reliability of the identification of the human voice," *J. Gen. Psychol.*, **17**, 249–271 (1937).
[25] F. McGehee, "An experimental study of voice recognition," *J. Gen. Psychol.*, **31**, 53–65 (1944).
[26] F. Clarke and R. Becker, "Comparison of techniques for discriminating among talkers," *J. Speech Hear. Res.*, **12**, 747–761 (1969).
[27] S. Cook and J. Wilding, "Earwitness testimony: never mind the variety, hear the length," *Appl. Cogn. Psychol.*, **11**, 95–111 (1997).
[28] A. D. Yarmey, A. L. Yarmey, M. Yarmey and L. Parliament, "Commonsense beliefs and the identification of familiar voices," *Appl. Cogn. Psychol.*, **15**, 283–299 (2001).
[29] I. Pollack, J. Pickett and W. Sumby, "On the identification of speakers by voice," *J. Acoust. Soc. Am.*, **26**, 111–117 (1954).
[30] T. Orchard and A. D. Yarmey, "The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification," *Appl. Cogn. Psychol.*, **9**, 249–260 (1995).
[31] M. Hashimoto, S. Kitagawa and N. Higuchi, "Quantitative analysis of acoustic features affecting speaker identification," *J. Acoust. Soc. Jpn.* (*J*), **54**, 169–178 (1998) (in Japanese).
[32] T. Kitamura and P. Mokhtari, "Effects of intra-speaker variation of speech sounds on perception of speaker characteristics," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 525–526 (2005) (in Japanese).
[33] T. Kitamura and T. Saito, "Effects of acoustic modifications on perception of speaker characteristics for sustained vowels," *Tech. Rep. IEICE*, **106**, 43–48 (2007) (in Japanese).
[34] R. Coleman, "Speaker identification in the absence of inter-subject differences in glottal source characteristics," *J. Acoust. Soc. Am.*, **53**, 1741–1743 (1973).
[35] P. Bricker and S. Pruzansky, "Effects of Stimulus Content and Duration on Talker Identification," *J. Acoust. Soc. Am.*, **40**, 1441–1449 (1966).
[36] R. Roebuck and J. Wilding, "Effects of vowel variety and sample length on identification of a speaker in a line-up," *Appl. Cogn. Psychol.*, **7**, 475–481 (1993).
[37] K. Amino, T. Sugawara and T. Arai, "The correspondences between the perception of the speaker individualities contained in speech sounds and their acoustic properties," *Proc. Interspeech*, pp. 2025–2028 (2005).
[38] T. Kitamura and M. Akagi, "Speaker individualities in speech spectral envelopes," *J. Acoust. Soc. Jpn.* (*E*), **16**, 283–289 (1995).
[39] T. Matsui, I. Pollack and S. Furui, "Perception of voice individuality using syllables in continuous speech," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 379–380 (1993) (in Japanese).
[40] T. Nishio, "Can we recognise people by their voices?," *Gengo-Seikatsu*, **158**, 36–42 (1964) (in Japanese).
[41] K. Stevens, C. Williams, J. Carbonell and B. Woods, "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," *J. Acoust. Soc. Am.*, **44**, 1596–1607 (1968).
[42] S. Nakagawa and T. Sakai, "Feature analyses of Japanese phonetic spectra and considerations on speech recognition and speaker identification," *J. Acoust. Soc. Jpn.* (*J*), **35**, 111–117 (1979) (in Japanese).
[43] L. S. Su, K. P. Li and K. S. Fu, "Identification of speakers by use of nasal co-articulation," *J. Acoust. Soc. Am.*, **56**, 1876–1882 (1972).
[44] A. Schmidt-Nielsen and K. Stern, "Recognition of previously unfamiliar speakers as a function of narrow-band processing and speaker selection," *J. Acoust. Soc. Am.*, **79**, 1174–1177 (1986).
[45] K. Amino, T. Sugawara and T. Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoust. Sci. & Tech.*, **27**, 233–235 (2006).
[46] K. Amino and T. Arai, "Effects of stimulus contents and speaker familiarity on perceptual speaker identification," *Acoust. Sci. & Tech.*, **28**, 128–130 (2007).
[47] T. Hirayama, *Dictionary of the Japanese Accents* (Tokyodo, Tokyo, 1960) (in Japanese).
[48] H. Kindaichi and K. Akinaga, *Dictionary of Japanese Accents*, 2nd ed. (Sanseido, Tokyo, 1981) (in Japanese).
[49] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Ver. 4.3.16 (Computer Program), URL: http://www.praat.org/.
[50] E. Selkirk, *Phonology and Syntax: The Relation between Sound and Structure* (MIT Press, Cambridge, Mass., 1984).
[51] J. Bachorowski and M. Owren, "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," *J. Acoust. Soc. Am.*, **106**, 1054–1063 (1999).
[52] T. Kitamura, K. Honda and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. & Tech.*, **26**, 16–26 (2005).
[53] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristics measurement and morphological observation," *J. Acoust. Soc. Am.*, **100**, 3374–3383 (1996).
[54] O. Engwall, V. Delvaux and T. Metens, "Interspeaker variation in the articulation of nasal vowels," *Proc. Int. Semin. Speech Production*, pp. 3–10 (2006).
[55] K. Amino, T. Sugawara and T. Arai, "Effects of the syllable structure on perceptual speaker identification," *IEICE Tech. Rep.*, **105**, 109–114 (2006).
[56] O. Fujimura, "Analysis of nasal consonants," *J. Acoust. Soc. Am.*, **34**, 1865–1875 (1962).
[57] P. Bricker and S. Pruzansky, "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, N. Lass, Ed. (Academic Press, New York, 1976), pp. 295–325.
[58] C. E. Williams, "The effects of selected factors on the aural identification of speakers," in Sect. 3 of Report ESD-TDR-65-153, *Electronics Systems Division* (Air Force Systems Command. Hanscom Field, 1964).
[59] T. Miura, *New Edition: Hearing and Speech* (*Chokaku to Onsei*) (Corona Publishing Company, Tokyo, 1980) (in Japanese).
[60] M. Hashi, A. Sugawara, T. Miura, S. Daimon, Y. Takakura and R. Hayashi, "Articulatory variability of Japanese moraic-nasal," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 411–412 (2005) (in Japanese).
[61] N. G. Clements, "Geometry of phonological features," *Phonol. Yearb.*, **2**, 225–252 (1985).
[62] A. Spencer, *Phonology* (Blackwell, Oxford, 1996).
[63] L. Whaley, *Introduction to Typology* (Sage Publications, London, 1997).
[64] R. Tull and J. Rutledge, "Analysis of cold-affected speech for inclusion in speaker recognition syste," *J. Acoust. Soc. Am.*, **99**, 2549 (1996).

Kanae Amino received her BA and MA degrees in linguistics from Sophia University, Tokyo, Japan, in 2002 and 2004, respectively. She was a JSPS Research Fellow in 2005–2006. Currently, she is a PhD student in electrical and electronics engineering at Sophia Univ. and working on the variations in speech production and the perception of the speaker individualities. Her research interests also include Japanese phonology and phonetics, dialectology and forensic sciences.

Takayuki Arai received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute (Berkeley, CA, USA). He is currently Professor of the Department of Information and Communication Sciences, Sophia Univ. In 2003–2004, he is a visiting scientist at Massachusetts Institute of Technology. His research interests include signal processing, acoustics, speech and hearing sciences, and spoken language processing.