Volume 185, Nos. 1–3, 10th March 2009
ISSN 0379-0738
ELSEVIER
FSI
FORENSIC SCIENCE INTERNATIONAL
An international journal dedicated to the applications of medicine and science in the administration of justice
Editor-in-Chief
P. Saukko
Associate Editors
A. Carracedo
C. Cattaneo
O.H. Drummer
M. Hall
P. Margot
G. Willems

Forensic Science International 185 (2009) 21–28

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Speaker-dependent characteristics of the nasals

Kanae Amino [*], Takayuki Arai

Faculty of Science and Technology, Department of Electrical and Electronics Engineering, Sophia University, Tokyo, Japan

ARTICLE INFO

Article history:
Received 3 June 2008
Received in revised form 27 November 2008
Accepted 28 November 2008
Available online 21 January 2009

Keywords:
Speaker identification
Inter-speaker variations
Nasals
Velic action
Energy contours

ABSTRACT

Investigation on human speaker identification enables us to know the indexical cues to speakers, and it may consequently lead to the effective acoustical parameters that can be used for forensic speaker recognition. It is known that speaker individuality interacts with the phonological or linguistic information contained in speech signals. As proof, the accuracy of perceptual speaker identification (PSI) performances depends on what types of sounds are presented to the listeners. In a series of our previous experiments, we have been investigating the effective sounds for PSI, and the stimuli containing a nasal were found to be the ones. In this present study, we conducted another PSI experiment in order to examine the reproducibility of the nasal effectiveness, and to see the effects of the following vowels. Coronal nasals were shown to be effective despite the different speaker set or the following vowels, and the stimuli containing a nasal were significantly better than those without it. In the second part of this paper, we introduce the results of the acoustical analysis of the stimuli. The contours of the energy transitions showed variations in shape among speakers for all three types of the analysis targets; nasals, stops, and fricatives, although the inter-speaker difference in the energy slopes for the consonant articulation was significant especially in nasal sounds. We also examined the effects of the sampling frequencies and the speech codecs, and found that the speaker-dependent shapes of these energy contours were maintained as long as the speech materials were uncompressed. The contours of the nasals appeared to be stable within a speaker, compared to other types of sounds.

1. Introduction

Human beings are able to reliably identify familiar speakers only by speech sounds. In the viewpoint of the speech chain, the speaker emits various kinds of information conveyed by the utterance, and the listeners perceive and recognise them in order for a communication to be successful [1,2]. The information carried by speech sounds includes not only the linguistic information, but also the information about the speaker. The former is the contents or the message of the utterances, and the latter covers the physiological and behavioural traits of the speakers.

Understanding the production and perception of the speaker individuality is important for speaker identification in forensic fields. Specifically, perceptual speaker identification (PSI) is closely related to the judgments of the adequacy of the earwitnesses' testimonies. A research approach is proposed where PSI experiments are conducted in order to investigate speaker individualities [3], and this is often found in the studies for forensic and security purposes. Some of them have tried to find out the factors that affect PSI performances [4–8], and others have attempted to find the acoustical correlates to those factors [9,10].

The factors such as speakers' health states, the degree of familiarity between the speakers and listeners, the quality of the recordings, and non-contemporary speech materials are all known to degrade the identification performances [4,11–14]. Voice disguise is one of the most important issues in forensic speaker recognition, although it is not an issue for wiretapping. Zhang and Tan [15] examined the effects of the ten types of voice disguise, including objects in mouth and pinched nostrils, and found that masking on mouth and whisper were most effective types. Hirson and Duckworth [16] reported the phonation type, creak in this study, affected the perception, though the spectra of /s/ were relatively stable and robust against creakiness within one speaker.

Limitations of the human memory, how long one can remember the voice of a given speaker, and how many speakers one can remember, and the cognitive processing for speaker identification tasks, either for matching or naming, are also important factors in PSI [14,17]. The duration of the speech data needed for accurate speaker identification is not clarified yet [14,18]. Pollack et al. [19] reported that PSI performances were improved with increasing utterance duration but this occurs only for brief speech, up to 1200 ms length.

Other studies also suggested that phonemic variations are more important for identification accuracy rather than the utterance

[*] Corresponding author at: 7-1 Kioi-cho, Chiyoda-ku, Tokyo 102-8554, Japan. Tel.: +81 3 3238 3417.
E-mail addresses: amino-k@sophia.ac.jp (K. Amino), arai@sophia.ac.jp (T. Arai).

0379-0738/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2008.11.018

Table 1
Summary of our previous research; experimental conditions.

Experiments	No. of Speakers[a]	No. of Listeners[b]	Stimuli	Effective syllables
Amino [25]	3 (F)	14 (familiar)	CV syllables (in isolation)[c]	Syllables containing nasals
Amino [26]	3 (M), 3 (F)	18 (familiar)	CV syllables (excerpted)	Syllables containing nasals
Amino et al. [27]	10 (M)	5 (familiar)	CV syllables	Syllables containing nasals
Amino et al. [28]	8 (M)	8 (familiar)	Monosyllables of various structures	Syllables containing nasals
Amino and Arai [29]	4 (M)	16 (unfamiliar)	CV syllables	Syllables containing nasals

[a] Speakers' sex is shown in the brackets; M stands for male speakers, and F for female speakers.
[b] Familiarity to the speakers is shown in the brackets.
[c] CV stands for a consonant–vowel sequence.

Table 2
List of the stimulus monosyllables; three tokens uttered by four speakers were used.

Consonant	/a/	/e/	/i/	/o/	/ɯ/
None	/a/	/e/	/i/	/o/	/ɯ/
Stops /t/, /d/	/ta/, /da/	/te/, /de/	–	/to/, /do/	–
Tap or flap /ɾ/	/ɾa/	/ɾe/	/ɾi/	/ɾo/	/ɾɯ/
Fricatives /s/, /z/, /ʃ/	/sa/, /za/, /ʃa/	/se/, /ze/	/ʃi/	/so/, /zo/, /ʃo/	/sɯ/, /ʃɯ/
Affricates /t͡ʃ/, /t͡s/, /d͡ʒ/, /d͡z/	–	–	/t͡ʃi/, /d͡ʒi/	–	/t͡sɯ/, /d͡zɯ/
Nasals /m/, /n/, /ɲ/	/ma/, /na/, /ɲa/	/me/, /ne/	/mi/, /ni/	/mo/, /no/, /ɲo/	/mɯ/, /nɯ/, /ɲɯ/
Approximants /j/, /w/	/ja/, /wa/	–	–	/jo/	/jɯ/

duration [4,5,17,20]. In addition, differential effects of the phonemes in the availability for identifying speakers are pointed out [19,20]. If we could find the sounds that indicate speaker individuality more than other sounds do, we can use them in the utterances for text-prompt speaker identification, or we can efficiently identify speakers by focusing on those sounds.

It is known that listeners can identify speakers more accurately when vowels and voiced consonants are presented to them [9,19–22]. Specifically, the availability of liquids was reported in English [23] and Chinese [24]. In a series of our previous experiments [25–29], nasals were consistently effective for PSI, despite different sets of speakers and listeners. Summary on test conditions of these works are shown in Table 1. Correspondences between the spectral properties of the stimuli and PSI accuracy were also observed [27]. Thirtieth-order FFT cepstra were used to calculate inter- and intra-speaker spectral distances. The ratios of inter-speaker distances to intra-speaker distances were greater in nasal sounds than in oral sounds, and this means that nasals have greater inter-speaker variations and smaller intra-speaker variations. Effectiveness of nasals is also reported in automatic speaker recognition [30–32].

The fact that listeners can identify a speaker means that there are acoustical correlates contained in speech sounds, and investigating those acoustical cues can contribute to automatic forensic speaker recognition, too. Moreover, the fact that the differences exist among phonemes in the effectiveness for PSI means that speaker-dependent characteristics interact with phonological information of speech [33], thus acoustical properties important for speaker individuality may, or may not, lie in those acoustical cues that are crucial for phoneme distinctions. In this present study, we conducted a PSI experiment and inspected the acoustical properties of the stimuli on the basis of the experimental results.

In the experiment, we focused on the effect of the consonants in monosyllabic stimuli and the vowels which followed the consonants (CV). In the previous experiments, we had only one vowel /a/ in the stimuli in order to make the experiments simple [25–29]. However, the syllable onset consonants inevitably go under the process of co-articulation, which we must consider in experimental conditions [34]. Another aim of this study was to see whether the reproducibility of nasals appears again with a different set of speakers. In the second part of this paper, we conducted an acoustical analysis of the stimuli used in the experiment in order to find a speaker-dependent feature. The speech materials involved in forensics often suffer from a lack of duration and a low quality of recording [14]. In this study we will see the basic behaviours of the monosyllables as a first step. In the acoustical analyses, we will also see the effects of the different types of speech samples and codecs.

2. PSI experiment

2.1. Speech materials

Speech data of four male speakers were selected from JEIDA (Japanese Electronic Industry Development Association) speech corpus to be used in the experiment [35]. All the data were digitised at the sampling frequency of 48 kHz with 16-bit resolution. Among 110 entries of Japanese monosyllables, 48 syllables shown in Table 2 were selected in accordance with our previous experiments [27–29]. Information on the four male speakers is shown in Table 3. They were all native speakers of Tokyo Japanese. Their recordings were held in a quiet room, thus the speech data contained little background noise.

Three tokens for each speaker and for each syllable were used in the experiment. Before the experiment two naive listeners who have never had phonetic training and whose mother language is Tokyo Japanese listened to all the stimulus monosyllables, and confirmed that all of them sounded as the ones intended by the speakers.

2.2. Participants and procedures

Fifteen (eight male and seven female) volunteers participated in the listening experiment. Their mean age was 23.4 years at the

Table 3
Speaker ensemble of four male speakers; average fundamental frequencies for the four speakers; average of all 144 monosyllables, analysed for the stable vowel parts, manually.

Speakers	Age	Height [cm]	Average $F0$ [Hz]	S.D. of $F0$ [Hz]
Speaker #1	In 20s	181	148.87	6.69
Speaker #2	In 20s	171	126.97	3.91
Speaker #3	In 30s	169	164.70	6.47
Speaker #4	In 40s	164	121.48	3.86

time of the experiment, and none of them had known hearing problems. No one had heard the speakers' utterances before.

All the experiments were conducted in a sound-treated room and all the speech data were played on a computer. The stimuli were presented to the listeners through headphones (SONY MDR-Z 700). First, the participants listened to each speaker's sample words in order to get familiarised with them. They listened to the words as many times as they wanted. The sample words were the following three: /hoɾʲɯː/(保留, suspension), /kaigʲoː/(改行, creating a new line), and /heɴkaɴ/ (変換, conversion). These words were selected from the JEIDA corpus again, on the basis that they do not contain any of the stimulus syllables.

Next they practised the experimental task using the sample words described above. The task was designed and conducted with Praat MFC (Multiple Forced Choice) experiment programme [36]. Feedback was given after each trial in the practice phase. We repeated familiarisation and practice until the listeners could tell the speakers with more than 90% accuracy. This learning session took them about 15 min on average.

Then we moved on to the test sessions, where no feedbacks were given. The listeners were no longer allowed to listen to the sample utterances, replay the stimuli, nor change their response once answered. The total number of the stimuli was 576, and the listeners took breaks after every 192 trials.

3. Results and discussion

The results of the experiment are evaluated by the percent correct speaker identification. Fig. 1(a and b) indicates average PSI results for vowels and consonants, respectively. All the consonants and the vowels gained the identification scores higher than the chance level, i.e. 25% correct.

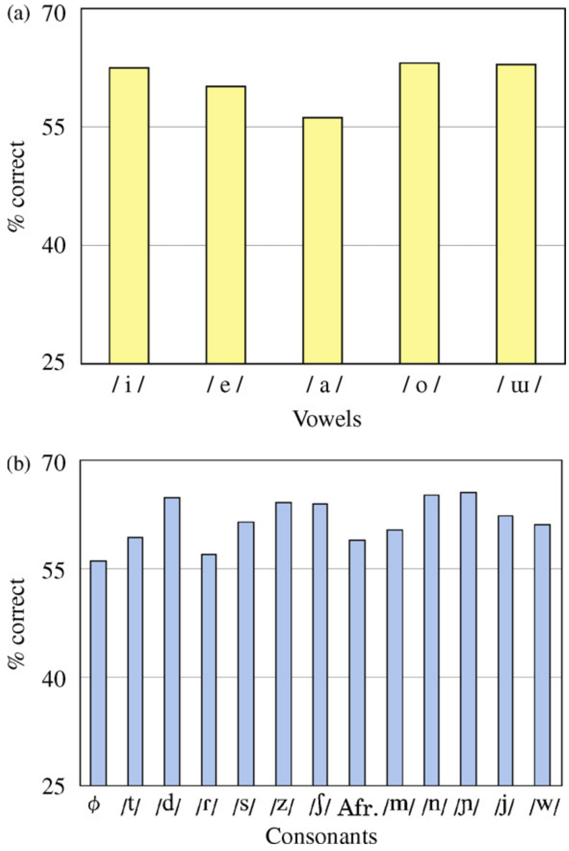

Fig. 1. Results of perceptual speaker identification according to (a) vowels and (b) consonants.

When we focus on the effects of the vowels in Fig. 1(a), back vowels /o/, /ɯ/ and /a/ were more effective for PSI than the front vowels /i/ and /e/. The difference was again significant in Mann–Whitney *U*-test ($p = 0.003$).

Among the consonants, in Fig. 1(b), coronal nasals /n/ and /ɲ/ gained relatively higher identification scores than others. Coronal sounds are articulated with the front part of the tongue, and include dental, alveolar and post-alveolar consonants. The voiced coronal stop and fricative /d/ and /z/ also obtained higher scores. Again, the asymmetry between coronal and labial nasals was observed. These tendencies are consistent with those seen in our previous experiments, despite the different sets of speakers and listeners. The difference among consonants showed a significant tendency in one-way ANOVA ($F(11, 564) = 1.75$, $p = 0.059$). Specifically, the difference between the nasals and non-nasals was significant in Mann–Whitney *U*-test ($p = 0.045$).

As Bricker and Pruzansky suggested, not only the selection of the speakers but also that of the speaker ensembles are important in PSI experiments, that is, the results would be influenced by the speaker ensembles, besides the speakers themselves. This means that it is also important to take into account to whom the speakers are compared [17]. Table 4 shows the confusions among the speakers. We can see that certain pairs of speakers, for example, speakers #2 and #4, were more often confused than other pairs like speakers #2 and #3. Different speaker sets may yield different PSI performances. In this sense, the differences among the phonemes and the effectiveness of the coronal nasals in PSI have been repeatedly confirmed with different sets of speakers, and it is worth analysing these effective sounds for seeking speaker-specific acoustical features.

4. Acoustical analyses

4.1. Methodology

The targets of the acoustical analysis are the stimulus syllables containing one of the following six consonants; /m/, /n/, /t/, /d/, /s/, and /z/. We used the subset of the stimuli used in the PSI experiment. As Table 2 shows, there were five syllables for the two nasals, four for the two fricatives and three for the two stops, thus we had 24 syllables for the analysis. All three tokens for each were analysed.

The analysis parameter was the transitions of energy across the time. This parameter was selected because it captures the abrupt temporal changes well [37]. Also, this parameter is suitable for the observation of the articulatory idiosyncrasies. First, all the target syllables were down-sampled from 48 kHz to 16 kHz. In order to see the effects of the sampling frequencies and the speech codecs, we also down-sampled these uncompressed linear PCM materials to 8 kHz as well as we created the materials of the ACELP codec (ITU-T G.729), thus three types of speech materials were submitted to the acoustical analyses.

Then we calculated the energy for each syllables by frames of 30 ms length with a shift of 10 ms. The energy contours for each syllable were normalised by the maximal value, and the contours for the stop consonants, fricatives and nasals are plotted across time. The figures for the uncompressed speech data (linear PCM) with the

Table 4
Confusion matrix among perceived and actual speakers.

	% Response	Perceived Speakers			
		Speaker #1	Speaker #2	Speaker #3	Speaker #4
Actual speakers	Speaker #1	63.4	13.4	11.9	10.7
	Speaker #2	25.3	46.5	6.9	32.3
	Speaker #3	8.2	0.8	79.4	0.9
	Speaker #4	3.1	39.3	1.8	56.1

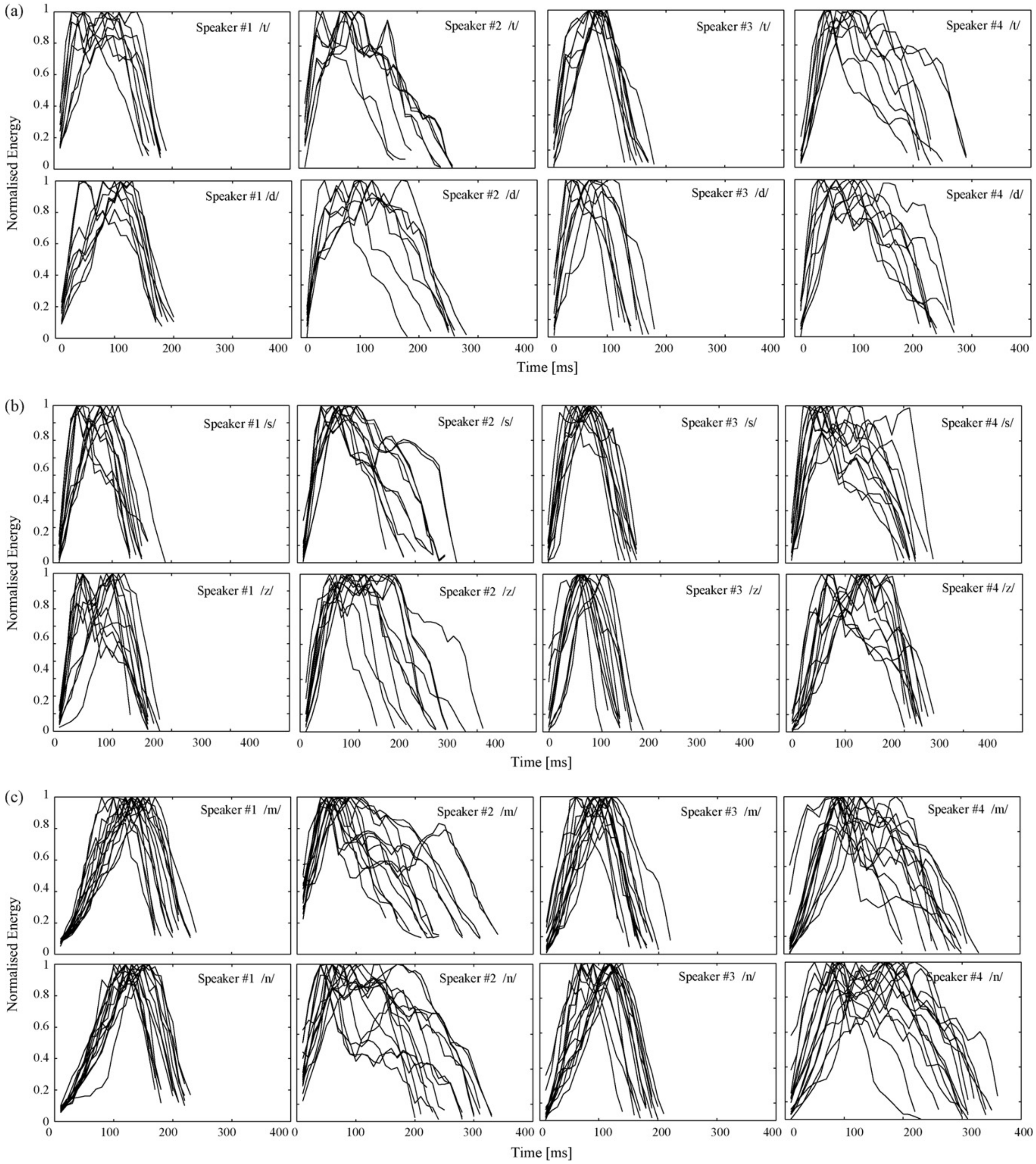

Fig. 2. Energy contours for the syllables containing (a) a stop, (b) a fricative, and (c) a nasal consonant; the data are uncompressed linear PCM of 16 kHz sampling rate; upper panels show the contours for voiceless consonants (bilabial for the nasal), and lower panels voiced (alveolar for the nasal); from left to right Speakers #1 to #4.

16 kHz sampling frequency, 8 kHz sampling frequency, and the compressed materials (G.729) are shown in Figs. 2–4, respectively.

In Fig. 2(a–c), we can see that the contours have speaker-dependent shapes. The upper panels in Fig. 2(a and b) show the contours for voiceless stop /t/ and voiceless fricative /s/, the lower ones the voiced stop /d/ and voiced fricative /z/, respectively. We find that the contours for voiceless sounds are more uniform than voiced sounds. In Japanese, the sentence-initial or word-initial /z/ that follows a pause may be realised as an affricate /dz/, and the timing of the consonant–vowel transitions in /d/ or /dz/ vary due to the pre-voicing during the closure for /d/. As for the nasals in Fig. 2(c), the durations of the nasal murmur may also vary among utterances, and this is why the timing when the consonant–vowel transition commences is not constant within a speaker. However, in these figures we notice that the left slopes of the contours are relatively consistent in a given speaker.

In comparison to Figs. 3 and 4, the contours in Fig. 2 seem to have speaker-dependent shapes. Fig. 3 also shows that the shapes of the energy contours are relatively stable within a speaker, and differ among speakers. We can say that the effect of down-

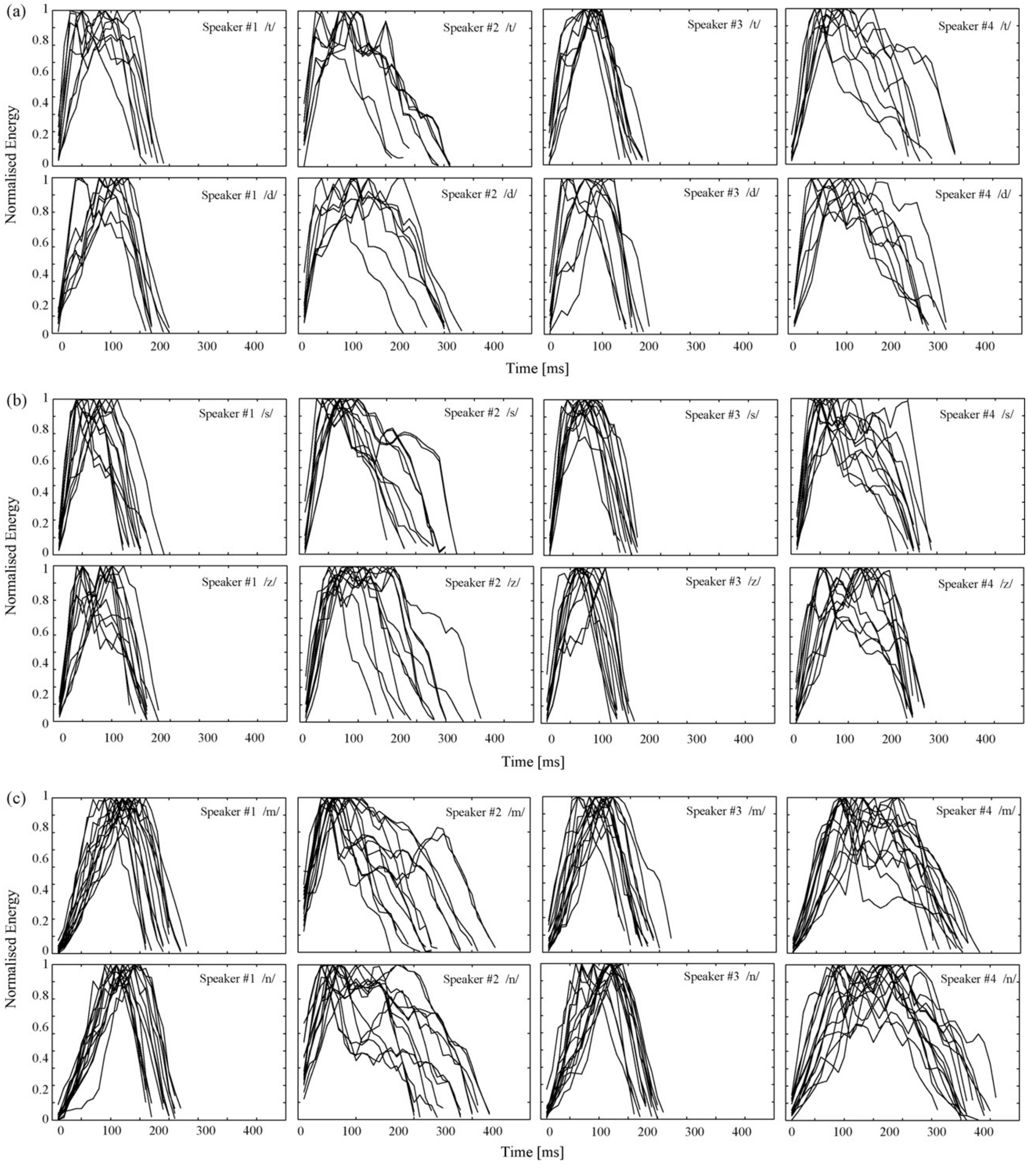

Fig. 3. Energy contours for the syllables containing (a) a stop, (b) a fricative, and (c) a nasal consonant; the data are uncompressed linear PCM of 8 kHz sampling rate; upper panels show the contours for voiceless consonants (bilabial for the nasal), and lower panels voiced (alveolar for the nasal); from left to right Speakers #1 to #4.

sampling was not great. On the other hand, the effect of the codecs is obvious, when we compare the contours in Figs. 2 and 3 and those in Fig. 4. Speaker-dependency of the energy contours was lost, when the speech materials got compressed.

In the three figures above, the left sides of the contours indicate consonant articulations, and the slopes directly reflect the energy transitions in consonant–vowel shifts. We calculated the linear approximations for the left side slopes in linear PCM materials with both 16 kHz and 8 kHz sampling frequencies. Slope calculations were performed on the basis of the following two criteria; calculation includes (1) the intervals where the normalised energy values are greater than 0.1, and (2) the intervals of monotonous increase. In the slope analyses, six more speakers were added to our data, apart from the four speakers whose materials were used in the perception test, in order to make the data more convincing.

The mean slope values are shown in Table 5. With the materials of 16 kHz sampling frequency, one-way ANOVA (Analysis of Variance) showed a significant difference among speakers in nasals (F (9, 299) = 21.67, $p < .0001$). Inter-speaker differences were less significant in stops (F (9, 179) = 2.07, p = 0.03) or in fricatives (F (9,

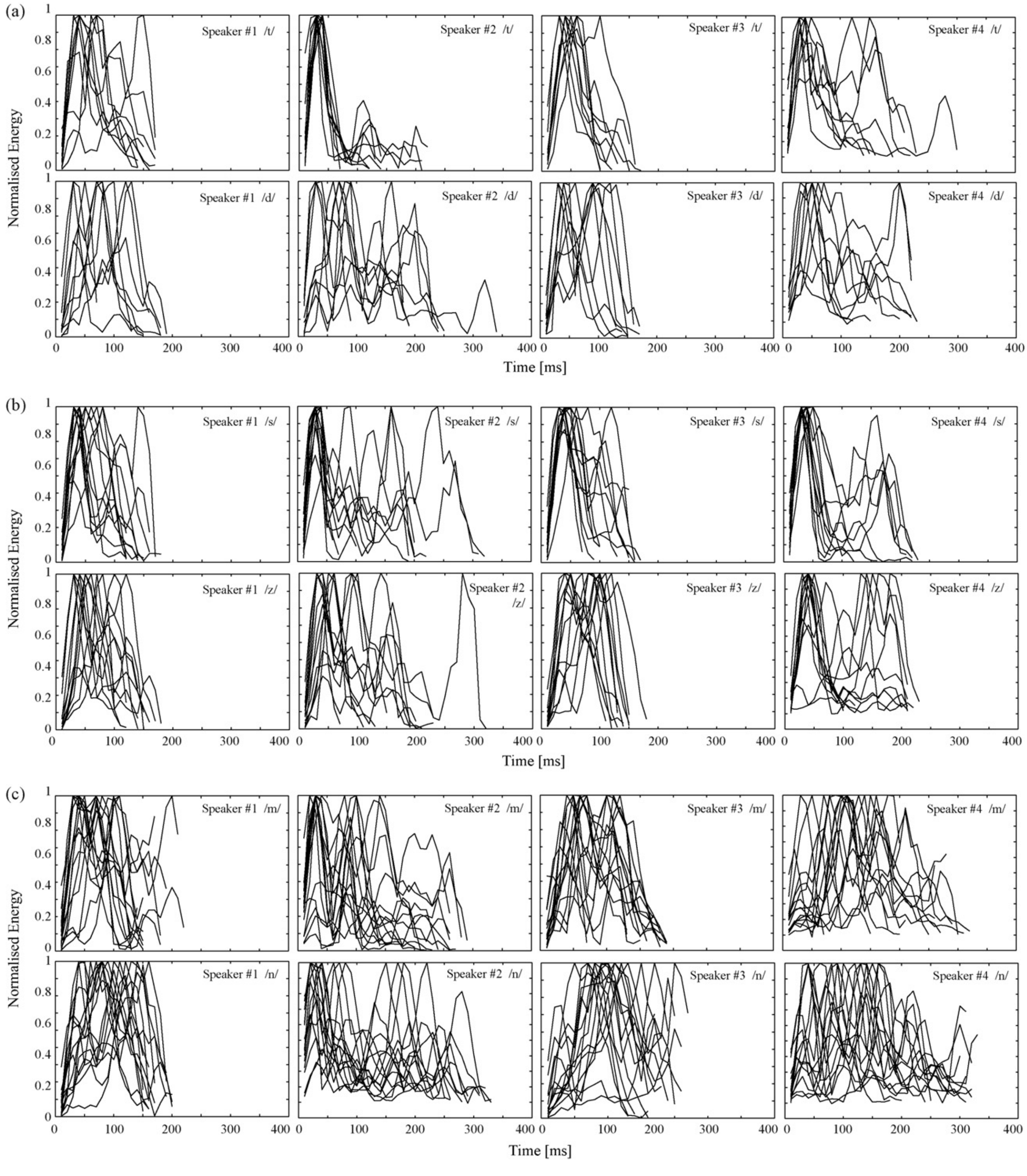

Fig. 4. Energy contours for the syllables containing (a) a stop, (b) a fricative, and (c) a nasal consonant; the data are in ITU-T G.729 codec; upper panels show the contours for voiceless consonants (bilabial for the nasal), and lower panels voiced (alveolar for the nasal); from left to right Speakers #1 to #4.

239) = 2.74, $p < 0.01$). The results of ANOVA with the materials sampled at 8 kHz showed significant differences among the speakers in all types of sounds ($p < 0.01$), but Tukey's post hoc tests showed that the slope values of the nasals significantly differentiated the most speaker-pairs.

5. General discussion

In this study, we conducted a PSI experiment where differential effects of the stimuli were examined. The results showed that the coronal nasals, /n/ and /ɲ/, are effective for identifying the speakers. This tendency has been consistently observed in our previous experiments despite different speaker sets [25–29]. The nasal-oral difference was significant, too. Also in previous works, nasals had greater inter-speaker spectral distances [32] compared to non-nasal sounds [27]. The nasal consonants have speaker-dependent resonant properties, because the resonant cavities involved in their articulation, nasal cavity, velopharyngeal cavity, and paranasal sinuses, have morphological variations among individuals [38]. PSI performances were significantly better when the back vowels, /o/,

Table 5
Mean slope values for the left sides of the energy contours; the number of samples *N* was 9 for the stops, 12 for the fricatives and 16 for the nasals. Slopes were calculated in the intervals between 0.1 and 1.0 (normalized maximum), and only in the intervals of monotonous increase.

Speakers		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Uncompressed linear PCM (16 kHz)											
	/t/	0.187	0.194	0.110	0.154	0.117	0.085	0.106	0.132	0.186	0.210
	/d/	0.126	0.144	0.171	0.147	0.117	0.139	0.126	0.167	0.130	0.139
	/s/	0.184	0.188	0.170	0.179	0.144	0.112	0.108	0.148	0.171	0.175
	/z/	0.149	0.146	0.139	0.129	0.127	0.117	0.125	0.129	0.198	0.145
	/m/	0.088	0.139	0.110	0.104	0.071	0.086	0.069	0.075	0.114	0.108
	/n/	0.084	0.133	0.098	0.100	0.064	0.068	0.060	0.078	0.111	0.096
Uncompressed linear PCM (8 kHz)											
	/t/	0.214	0.201	0.144	0.167	0.143	0.080	0.104	0.137	0.179	0.210
	/d/	0.124	0.136	0.169	0.141	0.126	0.144	0.126	0.168	0.125	0.126
	/s/	0.187	0.176	0.180	0.199	0.130	0.101	0.117	0.146	0.183	0.174
	/z/	0.179	0.159	0.155	0.151	0.127	0.119	0.129	0.128	0.199	0.151
	/m/	0.093	0.139	0.120	0.096	0.070	0.086	0.068	0.074	0.113	0.107
	/n/	0.087	0.132	0.107	0.091	0.062	0.067	0.058	0.076	0.121	0.097

/ɯ/ and /a/, followed the consonants in the stimulus syllables, compared to the syllables with the front vowels. The reason for this was not examined in this study, but the back vowels have low F2 values and this may affect the perception.

In the acoustical analysis, we used a parameter of energy transitions, and drew the energy contours for the stops, the fricatives and the nasals, for three types of speech materials; uncompressed linear PCM with 16 kHz and 8 kHz of sampling frequencies, and compressed ACELP (G.729) data. The energy contours showed speaker-dependent curve shapes. The left sides of the contours, which reflect the consonant articulation, showed significant inter-speaker variations, especially in nasal sounds. The effect of the sampling frequencies was not great, and the shapes of the energy contours showed a speaker-dependency as long as the speech data were not compressed. However, in compressed (G.729) data, these speaker-characteristics were no longer observed, thus the parameters of the energy contours should be used only for the uncompressed speech materials.

When produced at the sample place, the only difference between the articulatory gesture of a stop and that of a nasal lies in whether it involves the raising or lowering gesture of the velum. By lowering the velum, nasals have another pathway, the nasal tract. Any aspects of speech production are influenced by speaker's physiological idiosyncrasies and their habits in articulation. The control of the velar movements is not easy, especially in a brief interval, although the movement itself is voluntary [39]. In this study, the analysis targets were monosyllables, thus the velic action should be completed in relatively short durations. We can see in Figs. 2 and 3 that the energy transitions from the syllable onset to the nucleus peak occurred in durations of 50–100 ms. It is known that the control of the intentional movements may take 70–100 ms [40]. This means that the velar movements may have occurred out of control by the brain for some speakers.

Nasals show speaker-specific characteristics, because the timing of the velic action in nasal articulation is consistent within a speaker, and varies among speakers. The resonance properties of the nasals also show speaker-dependency, as the shapes of the resonant cavities have morphological individualities [27,38]. Moreover, these two features, the velic control and the resonant cavities, cannot be changed at the speaker's will. Ideal features for forensic speaker identification must have greater inter-speaker variations and small intra-speaker variations, and should not be controlled by speakers [23]. Nasals seem to satisfy all these three criteria, although the slope value of the energy contours alone is not a strong parameter enough to identify a single speaker out of a large speaker set, and has to be used in combination with other parameters.

In our next step, we will find more elegant ways to discriminate among speakers by using the information on the velic control in order to incorporate this into the actual forensic speaker identification system. The effectiveness of the back vowels should be explained, too. We will also investigate whether the speaker-dependent characteristics are found in intervocalic and postvocalic positions, in words and sentences, and whether they are language universal.

Acknowledgement

This work was supported by Sophia University Open Research Centre from MEXT.

References

[1] P.B. Denes, E.N. Pinson, The Speech Chain, 2nd ed., W H Freeman & Co., San Francisco, 1993.
[2] D.C. Tanner, M.C. Tanner, Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection, Lawyers & Judges Pub Co., Tucson, 2004.
[3] D. O'Shaughnessy, Speech Communications—Human and Machine, 2nd ed., Addison-Wesley Publishing Company, New York, 2000.
[4] A.D. Yarmey, A.L. Yarmey, M. Yarmey, L. Parliament, Commonsense beliefs and the identification of familiar voices, Appl. Cogn. Psychol. 15 (2001) 283–299.
[5] T.L. Orchard, A.D. Yarmey, The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification, Appl. Cog. Psychol. 9 (1995) 249–260.
[6] A.D. Yarmey, E. Matthys, Voice identification of an abductor, Appl. Cog. Psychol. 6 (1992) 367–377.
[7] B.R. Clifford, Voice identification by human listeners—on earwitness reliability, Law Human Behav. 4 (1980) 373–394.
[8] T. Arai, M. Takahashi, N. Kanedera, Y. Takano, Y. Murahara, On the important modulation-frequency bands of speech for human speaker recognition, Proc. ICSLP 3 (2000) 774–777.
[9] T. Kitamura, M. Akagi, Speaker individualities in speech spectral envelopes, J. Acoust. Soc. Jpn. (E) 16 (1995) 283–289.
[10] R.E. Remez, J.M. Fellowes, D.S. Nagel, On the perception of similarity among talkers, J. Acoust. Soc. Am. 122 (2007) 3688–3696.
[11] E. Doherty, H. Hollien, Multiple factor speaker identification of normal and distorted speech, J. Phonetics 6 (1978) 1–8.
[12] A. Schmidt-Nielsen, K.R. Stern, Recognition of previously unfamiliar speakers as a function of narrow-band processing and speaker selection, J. Acoust. Soc. Am. 79 (1986) 1174–1177.
[13] A. Alexander, F. Botti, D. Dessimoz, A. Drygajlo, The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications, Forensic Sci. Int. 146 (2004) 95–99.
[14] H. Hollien, Forensic Voice Identification, Academic Press, San Diego, 2002.
[15] C. Zhang, T. Tan, Voice disguise and automatic speaker recognition, Forensic Sci. Int. 175 (2008) 118–122.
[16] A. Hirson, M. Duckworth, Glottal fry and voice disguise: a case study in forensic phonetics, Biomed. Eng. 15 (1993) 193–200.
[17] P. Bricker, S. Pruzansky, Speaker recognition, in: N. Lass (Ed.), Experimental Phonetics, Academic Press, London, 1976, pp. 295–326.
[18] S. Cook, J. Wilding, Earwitness testimony: never mind the variety, hear the length, Appl. Cog. Psychol. 11 (1997) 95–111.
[19] I. Pollack, J.M. Pickett, W.H. Sumby, On the identification of speakers by voice, J. Acoust. Soc. Am. 26 (1954) 403–406.

[20] P. Bricker, S. Pruzansky, Effects of stimulus content and duration on talker identification, J. Acoust. Soc. Am. 40 (1966) 1441–1450.

[21] G. Ramishvili, Automatic voice recognition, Eng. Cybernetics 5 (1966) 84–90.

[22] K. Stevens, C. Williams, J. Carbonell, B. Woods, Speaker authentication and identification: a comparison of spectrographic and auditory presentations of speech material, J. Acoust. Soc. Am. 44 (1968) 1596–1607.

[23] F. Nolan, The Phonetic Basis of Speaker Recognition, Cambridge Studies in Speech Science and Communication, Cambridge University Press, Cambridge, 1983.

[24] C. Zhang, J. van de Weijer, J. Cui, Intra- and inter-speaker variations of formant pattern for lateral syllables in standard Chinese, Forensic Sci. Int. 158 (2006) 117–124.

[25] K. Amino, The characteristics of the Japanese phonemes in speaker identification, Proc. Sophia Univ. Linguistic Soc. 18 (2003) 32–43 (in Japanese).

[26] K. Amino, Properties of the Japanese phonemes in aural speaker identification, IEICE Tech. Rep. 104 (2004) 49–54 (in Japanese).

[27] K. Amino, T. Sugawara, T. Arai, Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties, Acoust. Sci. Tech. 27 (2006) 233–235.

[28] K. Amino, T. Sugawara, T. Arai, Effects of the syllable structure on perceptual speaker identification, IEICE Tech. Rep. 105 (2006) 109–114.

[29] K. Amino, T. Arai, Effects of stimulus contents and speaker familiarity on perceptual speaker identification, Acoust. Sci. Tech. 28 (2007) 128–130.

[30] S. Nakagawa, T. Sakai, Feature analyses of Japanese phonetic spectra and considerations on speech recognition and speaker identification, J. Acoust. Soc. Jpn. 35 (1979) 111–117 (in Japanese).

[31] L.S. Su, K.P. Li, K.S. Fu, Identification of speakers by use of nasal co-articulation, J. Acoust. Soc. Am. 56 (1972) 1876–1882.

[32] J.W. Glenn, N. Kleiner, Speaker identification based on nasal phonation, J. Acoust. Soc. Am. 43 (1967) 368–372.

[33] L. Nygaard, Perceptual integration of linguistic and nonlinguistic properties of speech, in: D. Pisoni, R. Remez (Eds.), The Handbook of Speech Perception, Blackwell Publishing, Oxford, 2005, pp. 390–413.

[34] P. Ladefoged, A Course in Phonetics, 4th ed., Heinle & Heinle Publishers, Boston, 2000.

[35] JEIDA Japanese Common Speech Data Corpus, http://www.sunrisemusic.co.jp/dataBase/fl/voicebase01_fl.html.

[36] P. Boersma, D. Weenink, Praat: doing phonetics by computer, Ver.4.5.14, Retrieved from http://www.praat.org/, Computer programme, 2005.

[37] T. Pruthi, C.Y. Epsy-Wilson, Acoustic parameters for automatic detection of nasal manner, Sp. Comm. 43 (2004) 225–239.

[38] J. Dang, K. Honda, Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation, J. Acoust. Soc. Am. 100 (1996) 3374–3383.

[39] G.K. Tortora, S.R. Grabowski, Introduction to the Human Body—The Essentials of Anatomy and Physiology, 5th Ed., Japanese Translation, Maruzen Co. Ltd., Tokyo, 2002.

[40] J. Hollerbach, Computers, brains and the control of movement, Trends Neurosci. 5 (1982) 189–192.