Is it possible to predict speaker's body size and oral cavity characteristics from speech signals? A preliminary study on Mandarin Chinese

Puyang Geng[*], Hong Guo, Qimeng Lu, Jinhua Zeng, Yan Li
Department of Forensic Audio, Image and Electronic Data Examination, Academy of Forensic Science, Shanghai, China

ABSTRACT

This paper proposes a study on whether the speaker's body size (height, weight) and oral cavity (lip protrusion LP, lip opening LO, front cavity FC) characteristics can be predicted based on the acoustic features of speech. Firstly, Pearson's correlation analysis was first conducted to examine the relationships between acoustic features and body size and oral cavity characteristics. Further, the effects of acoustic features in predicting body size and oral cavity characteristics were examined using random forest and decision tree models. The results showed that fundamental frequency statistics (i.e., mean, max, min) exhibited significant negative correlations with height, weight, and FC. Besides, good accuracies of classification in height, LP range, LO range, and FC range could be achieved based on the acoustic features. The findings in the current paper imply that acoustic features could be the potential features for identification of the speaker's body size and oral cavity characteristics. This paper will not only contribute to the researches and practices in forensic speaker profiling and but also provides foundations for the technology of automatic speaker recognition.

Keywords: Body size, oral cavity, acoustic features, correlation analysis, classification analysis

1. INTRODUCTION

Biometric features, as the information correlated with a particular person, have shown to be an important aspect of forensic science. In forensic science, the identification techniques on biometric features are significant for solving a specific type of criminal cases, such as DNA profiling, fingerprint, speaker identification, etc. Speaker profiling, as another aspect of identification technique on the voice of a particular person, refers to the revelation of the speaker's characteristics by examining the speech recordings[1]. For those cases where speech is the only accessible evidence a suspect left on the criminal scene, the techniques of speaker profiling would be a crucial approach to provide informative clues for investigation and target the criminals.

It has been proposed that various speaker profile characteristics could be discovered via speaker profiling, such as gender, age, professions, educational level, pathological feature, etc. For example, a blackmailer was profiled as working in the railway system for using a specialized term, Langsamfahrstrecke (i.e., a way to describe a low-speed section of the railway)[2]. The dialectal accent is another salient speaker profile characteristic that can reveal the speaker's regional dialectal group through accent features reflected in their speech. For example, Kulshreshtha and his colleagues conducted a study on Kahriboli (standard Hindi) and found vowel quality, word vocabulary, and tone of a speaker could contribute to the identification of dialect accent[1]. Besides, from a perspective of automatic speaker profiling, Lee et al. reported accuracies of 95%, 60%, and 40% for the classifications of the speaker's gender, age, and dialect, respectively[3]; Kalluri et al. reported the mean absolute error of age estimation based on short duration speech data is of 5.2 years for male speakers, and 5.6 years for female speakers[4].

In addition to the above speaker profile characteristics, other physiological features, such as body size and oral cavity, could also be important to identify the speaker's identity. In the following, the literature looking into the acoustic manifestations of body size and oral cavity will be reviewed.

[*] gengpy@ssfjd.cn

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 125062P
© 2022 SPIE · 0277-786X · doi: 10.1117/12.2661788

1.1 Relationship between body size and acoustic features

While some researches have been carried out to investigate the correlations between body size and acoustic features, a consensus has not been reached. Some scholars have proposed that acoustic features, such as fundamental frequency (henceforth F0) and formant frequency, are correlated with body size[5-7]. According to the source-filter theory, the thicker vocal folds typically vibrate more slowly than the thinner vocal folds, which would consequently lead to a relatively lower F0[8]. In addition, despite the dynamics of vocal folds, formant frequencies are negatively correlated with the vocal tract length, which would consequently lead to a relatively lower formant in a longer vocal tract (usually observed in taller people) than in a shorter vocal tract[8]. However, other scholars argue that F0 and formant frequencies show little association with body size within age and sex classes (i.e., between women, men, and children)[9].

The empirical studies on this topic also show divergent opinions. For instance, Evans et al. conducted a study on fifty male speakers and found body size (i.e., height and weight) was negatively correlated with F0 and formant dispersions[10]. However, Cao et al. reported that body size (i.e., height and weight) revealed no correlation with the mean and the median of F0, while significant negative correlations were found between body size and the standard deviation of F0[11]. The inconsistent findings of the above studies could result from the differences in the age range of the subjects, viz., the subjects of Evans et al. aged from 18 to 68 years and those of Cao et al. aged from 19 to 38 years.

1.2 Relationship between oral cavity and acoustic characteristics

The oral cavity, as one of the most fundamental organs for speech production, is individually different across speakers and could be a salient speaker profile characteristic for identifying a person’s identity. Surprisingly, to date, the way how oral cavity was demonstrated in speech signals has not been closely examined. Ohala proposed a hypothesis that the larger the oral cavity and the lower the frequency (i.e., F0 and formant frequency)[6]. In the following studies, Fitch[12] and Rendall et al.[9] have supported Ohala's hypothesis to some extent that the oral cavity is negatively correlated with F0 and formant frequency (i.e., the larger the oral cavity, the lower the F0 and formant frequency). However, all these studies were conducted on animals (e.g., whales, monkeys). Therefore, much uncertainty still exists about the relationship between the human oral cavity and acoustic characteristics.

1.3 The present study

In all the studies reviewed here, it seems that the conclusions on the relationship between body size and acoustic characteristics are still divergent. Besides, there has been little quantitative analysis of the relationship between the oral cavity and acoustic characteristics.

Therefore, this study set out to investigate the relationship between acoustic characteristics and body size, and oral cavity. Furthermore, the effect of acoustic features in identifying the human body size and oral cavity will be examined using random forest and decision tree models. This paper aims to make an important contribution to the field of forensic speaker profiling and automatic speaker recognition.

2. METHOD

2.1 Subjects

Thirty-three Mandarin Chinese speakers (i.e., 15 males and 18 females) were recruited for this study, all of whom were from northern China and spoke standard Mandarin. The means of age, height, and weight were 23.0 years, 176.7 cm (SD=4.7), 71.87 kg (SD=5.7), and 24.8 years, 164.7 cm (SD=4.6), and 51.3 kg (SD=3.6) for the male and female speakers, respectively. All subjects were right-handed and none had reported a history of the speech-hearing disorder.

2.2 Data collection

Body size. Height and weight were measured with the subjects’ shoes removed. The subjects were required to stand straight against the wall with their heads adjusted so that their eyes and ears were level. Height was measured using a tapeline and weight was measured using an electronic scale. All subjects were required to remove their overcoats and had no accessories on their bodies.

Oral cavity. To conduct a more comprehensive and precise study, the real-time oral cavity modulation during speech production was measured in our research. Forty-eight target sentences composed of 4-16 syllables were designed for oral cavity measurements.

Oral cavity data were recorded using NDI Wave Electro-Magnetic Articulography (EMA). As is shown in Figure 1, four five-degree-of-freedom sensors were placed on the tongue dorsum (TD), the upper and lower lips (UL and LL), and the jaw (JW) in the mid-sagittal plane to track the oral cavity data. A reference sensor (REF) was placed on the forehead to calibrate head movements and map the data to coordinate system. The movements of each sensor were tracked within the magnetic field produced by the EMA filed generator. The sampling rate was set as 200 Hz.

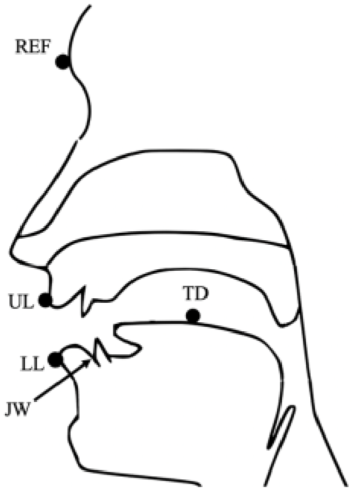

Figure 1. The placements of the sensors.

The experiment was conducted in a quiet room. A computer monitor and the OpenseSame software were used to present the target sentences. The speech was recorded by a clip microphone located 25cm away from the speaker's mouth. Each speaker was seated approximately 20cm away from the EMA magnetic field generator.

All target sentences were presented in a fixed random order and each target sentence was recorded twice for all speakers. Thus, altogether 48 (target sentences) * 2 (times) * 33 (participants) = 3168 utterances were recorded. For each sentence, the EMA Wavefront 2.0 software output oral cavity and audio data.

2.3 Measurements

Acoustic measurements. The utterances were first segmented and annotated at the syllable and phonemic levels by an automatic speech annotation software[13]. Acoustic characteristics were extracted using Praat software[14]. Fundamental frequency (F0) values were measured and manually corrected. Then, the F0 values in Hz were transformed to semitone values with 100 Hz as the reference frequency. The average value, the standard deviation, the extrema, and the range of F0 were then calculated for each target sentence. Also, the average value, the standard deviation, the extrema, and the range of intensity were calculated for each utterance.

Moreover, voice quality measures, viz., jitter, shimmer, HNR (harmonic to noise ratio), and H1-H2 (difference between first and second harmonic), were extracted from voiced segments of each target sentence using Praat. Then, the mean value of voice quality measures was calculated for each sentence.

Oral cavity measurements. To correct the speaker's head movements during the recording procedure, the oral cavity data were first calibrated and referred to the REF sensor. Then, all data were rotated to the occlusal plane (i.e., constructed by an upper incisor sensor, a molar sensor, and the REF sensor) to make the x-axis paralleled to the speaker's occlusal plane. Finally, the oral cavity data were smoothed and interpolated using a robust filtering algorithm.

The front-back (x) and superior-inferior (y) positions of TD, UL, LL, and JW in the coordinate system were measured for each target sentence. The lip protrusion (LP), the opening (LO), and the front cavity (FC) were calculated by the following formulas. Then, the mean and range of LP, LO, and FC were calculated for each target sentence. All measures are in millimeters (mm).

$$\text{Lip Protrusion (LP)} = LL\_x - JW\_x \quad (1)$$

$$\text{Lip Opening (LO)} = UL\_y - LL\_y \quad (2)$$

$$\text{Front Cavity (FC)} = \text{JW\_x}–\text{TD\_x} \tag{3}$$

3. RESULTS

3.1 Correlation analysis

The Pearson's correlation coefficients (pcc) were calculated between the acoustic characteristics and body size and oral cavity measurements, using SPSS software. As is shown in Table 1, no (i.e., $p > 0.05$) or weak (i.e., $p < 0.05$ and pcc < 0.3) correlations were found between body size/oral cavity and the majority of acoustic characteristics (e.g., intensity, jitter, shimmer, HNR, and H1-H2), except for the mean and the extrema of F0. To be specific, height and weight showed significant strong (i.e., pcc > 0.6) negative correlations with F0 mean, F0 max, and F0 min, while FC range showed significant moderate negative (i.e., 0.3 < pcc < 0.6) correlations with F0 mean and F0 range.

Table 1. Pearson's coefficients between acoustic features and body size/oral measures.

Acoustic feature	Height	Weight	LP mean	LP range	LO mean	LO range	FC mean	FC range
F0 mean	-0.8**	-0.9**	-0.1**	-0.1**	-0.3**	-0.1**	-0.3**	-0.5**
F0 max	-0.7**	-0.7**	0	0	-0.3**	0	-0.3**	-0.3**
F0 min	-0.6**	-0.7**	-0.1**	-0.2**	-0.3**	-0.1**	-0.2**	-0.5**
F0 sd	-0.1**	-0.1**	0.1**	0.1**	0	0.1**	-0.1**	0.1**
F0 range	-0.1**	-0.1**	0.1**	0.2**	0	0.1**	-.0.2**	0.1**
Intensity mean	-0.1**	0	-0.1*	0.1**	-0.1**	0.1**	-0.1**	0.1**
Intensity max	-0.1**	0	0	0.2**	0.1*	0.2**	-0.1**	0.2**
Intensity min	-0.2**	-0.3**	0	-0.2**	-0.2**	-0.2**	-0.2**	-0.3**
Intensity sd	0.1**	0.2**	0	0.2**	0.2**	0.3**	0	0.3**
Intensity range	0.1**	0.2**	0	0.3**	0.2**	0.3**	0	0.3**
jitter	0.2**	0.2**	0.1**	0	0.1**	-0.1**	0.1**	0.1**
shimmer	-0.1**	-0.1**	0.1**	-0.1**	0	-0.1**	0	-0.1**
HNR	-0.3**	-0.3**	-0.2**	-0.1**	-0.3**	0	-0.1**	-0.2**
H1-H2	0.2**	0.1**	-0.1**	-0.1**	-0.1**	-0.1**	0.2**	0

3.2 Random forest classification analysis

To conduct classification analysis, height, weight, and oral cavity measurements were first equally divided into subcategories according to the data distribution (as shown in Table 2).

A random forest model was built using the R package rondomForest with height, weight, and oral cavity features as dependent variables and the 14 acoustic characteristics as factors, respectively. The proportion of the training and test sets of the data was 7:3. To minimize the OOB (out-of-bag) prediction error rate, the mtry parameters were optimized by the tunTF function built in the rondomForest package.

The results of the random forest are shown in Table 3. The accuracies of classification were higher than 65% for all body size and oral cavity characteristics. Moreover, the classification rates were approximately higher than 70% for height (69.82%), LP range (69.41%), LO range (78.31%), and FC range (73.31%).

Table 2. The subcategories of body size and oral cavity measurements.

Feature	Subcategories							
Height (cm)	160 below	160-165	165-170	170-175	175-180	180-185		
Weight (kg)	45-50	50-55	55-60	60-65	65-70	70-75	75-80	80-85
LP mean (mm)	8-11	11-13	13-16	16 above				
LP range (mm)	6-10	10-14	14-18	18 above				
LO mean (mm)	15-20	20-25	25-30	30 above				
LO range (mm)	10-20	20-30	30-40	40-50				
FC mean (mm)	20-25	25-30	30-35	35-40	40 above			
FC range (mm)	15 below	15-20	20-25	25-30	30 above			

Table 3. The classification results of random forest model.

	Height	Weight	LP mean	LP range	LO mean	LO range	FC mean	FC range
Optimal mtry	4	4	4	3	4	4	6	6
Accuracy	69.82%	67.61%	66.81%	69.41%	68.22%	78.31%	67.25%	73.31%

3.3 Decision tree classification analysis

Decision tree models were further developed to explore the decision boundaries for estimating the subcategory probabilities of body size and oral cavity characteristics. The fourteen acoustic features were involved as factors, and body size and oral cavity characteristics were involved in the model as dependent variables, respectively. The proportion of the training and test sets of the data was 7:3.

The classification accuracies of body size and oral cavity features were 64.3% for height, 60.6% for weight, 51.6% for LP mean, 63.1% for LP range, 58% for LO mean, 69.9% for LO range, 48.1% for FC mean, and 56.1% for FC range. It is noteworthy that the overall classification accuracy for height was below 65%, while the accuracies were particularly high for subcategories 160-165cm (i.e., 93.4%) and 175-180cm (i.e., 86.4%). Taking height as an example, as shown in Table 4, the decision boundaries for height were heavily dependent on F0, intensity, and H1-H2.

Table 4. Decision boundaries for height based on the acoustic features.

	160 below	160-165	165-170	170-175	175-180	180-185
F0 mean		> 12.28	8.99 - 12.28		-0.70 - 8.99	< -0.70
F0 max				> 9.10		
F0 range	> 8.99					
Intensity mean		< 64.01			> 64.01	
H1-H2		< 3.18	3.18 - 6.11		> 6.11	

4. DISCUSSION

The current paper aims to explore whether the speaker profile characteristics, viz., body size (i.e., height and weight), and oral cavity (i.e., lip protrusion LP, lip opening LO, and front cavity FC), could be predicted by the acoustic features of speech signals. The results of correlation analysis revealed that body size was strongly negatively correlated with the mean and the extrema of F0, and the front cavity was moderately negatively correlated with the mean and the min of F0.

Moreover, based on the results of random forest and decision tree models, it was found that good classification accuracies (i.e., around 65%-70%) could be achieved for height, LP range, LO range, and FC range based on the acoustic features. Further, the random forest model provided higher accuracy for the classification tasks than the decision tree model.

A comparison of the findings with those of other studies confirms that body size is negatively correlated with F0[10]. However, our findings do not support the findings of Cao et al.[11], in which body size is found to be significantly correlated with F0 standard deviation. The inconsistent findings might result from the different research paradigms, which suggest more studies on the current topic.

In the meantime, it is also found that the front cavity is negatively correlated with F0, which provides evidence to previous hypotheses that the larger the vocal tract the lower the speech frequency[6].

Another important finding is that height, LP range, LO range, and FC range could be reasonably classified based on the acoustic characteristics, which further confirms that body size and oral cavity could be salient speaker profile characteristics. The classification accuracies of two height subcategories (i.e., 160-165cm and 175-180cm) were higher than those of other subcategories. This phenomenon may be due to the imbalanced data distribution of the speaker's height. Besides, according to the results of the decision tree model, the decision boundaries for height classification coincide with the general voice impression of people of different heights, viz., tall people usually have lower pitch levels than short people.

Further, this study found that the oral cavity characteristics, as more in-depth physiological measurements of individuals, could be predicted based on the acoustic features. Note that only the ranges of LP, LO, and FC received reasonable accuracies in the classification analysis. This finding implies that the movement range of lips and volume of the front cavity are speaker-dependent characteristics; however, no definitive conclusion can be drawn.

The generalisability of the results in this study is subject to several limitations. For instance, the amount of subjects can be further expanded in future studies to make a more precise examination. Secondly, other machine learning models need to be further tested for identifying body size and oral cavity characteristics, as the precisions were found different between the random forest and decision tree models in the current study.

5. CONCLUSION

The current study was set to determine whether the speaker's body size and oral cavity characteristics could be predicted based on speech signals. The results reveal not only significant correlations between body size/oral cavity and acoustic characteristics but also good accuracies in predicting height and range of oral cavity measurements. In general, therefore, it seems that body size and oral cavity are salient speaker profile characteristics. This study thus highlights the significance of speech for identifying individuals and provides foundations for automatic speaker recognition techniques.

ACKNOWLEDGMENTS

This study is jointly funded by the National Social Science Fund of China (21CYY011), and the grants from the Fundamental Research Funds for the Central Scientific Research Institutions [grant numbers GY2021G-9]. This study was approved by the Committee for the Protection of Human Subjects at the University of California, Berkeley (No. 2018-12-11643).

REFERENCES

[1] Kulshreshtha, M., Singh, C. P., and Sharma, R. M., "Speaker profiling: The study of acoustic characteristics based on phonetic features of Hindi dialects for forensic speaker identification," In [Forensic Speaker Recognition], Springer, 71-100(2012).

[2] Jessen, M., "Speaker classification in forensic phonetics and acoustics," In [Speaker classification I], Springer, 180-204(2007).

[3] Lee, J., Kim, K., Lee, K., and Chung, M., “Gender, age, and dialect identification for speaker profiling,” 22nd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), (2019).

[4] Kalluri, S. B., Vijayasenan, D., and Ganapathy, S., “Automatic speaker profiling from short duration speech data,” Speech Communication, 121, 16-28(2020).

[5] Aung, T., and Puts, D., “Voice pitch: A window into the communication of social power,” Current Opinion in Psychology, 33, 154-161(2020).

[6] Ohala, J. J., “An ethological perspective on common cross-language utilization of F_0 of voice,” Phonetica, 41(1), 1-16(1984).

[7] Reddon, A. R., Dey, C. J., and Balshine, S., “Submissive behaviour is mediated by sex, social status, relative body size and shelter availability in a social fish,” Animal Behaviour, 155, 131-139(2019).

[8] Pisanski, K., and Bryant, G. A., “The evolution of voice perception,” In [The Oxford Handbook of Voice Studies], Handbooks of Oxford, 269-300(2019).

[9] Rendall, D., Vokey, J. R., and Nemeth, C., “Lifting the curtain on the Wizard of Oz: Biased voice-based impressions of speaker size,” Journal of Experimental Psychology: Human Perception and Performance, 33(5), 1208(2007).

[10] Evans, S., Neave, N., and Wakelin, D., “Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice,” Biological Psychology, 72(2), 160-163(2006).

[11] Cao, H., Kong, J., and Wang, Y., “Relationship between the fundamental frequency and the speaker’s physiological parameters,” Journal of Tsinghua University (Science and Technology), 53(06), 848-851(2013).

[12] Fitch, W. T., “Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques,” The Journal of the Acoustical Society of America, 102(2), 1213-1222(1997).

[13] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M., [Montreal Forced Aligner], (2017).

[14] Boersma, P., and Weenink, D., [Praat: Doing Phonetics by Computer], (2021).