Estimation of Speaker's Height and Vocal Tract Length from Speech Signal

Sorin Dusan

Center for Advanced Information Processing
Speech and Language Processing Laboratory
Rutgers University, Piscataway, NJ 08854, U.S.A.
sdusan@caip.rutgers.edu

10.21437/Interspeech.2005-625

Abstract

Estimation of speaker's height and vocal tract length (VTL) from speech signal can have forensic and automatic speech recognition applications. It was suggested for a long time that there is a correlation between speaker's VTL, on one side, and speaker's height and formant frequencies, on another side. Until recently, these putative relationships have been empirically examined in studies employing relatively small numbers of speakers. Scattered studies presented intriguing results about the correlations between speaker's height and various acoustic speech parameters. Due to lack of databases, few studies presented extensive comparative results between the actual speaker's VTL and the estimated one from speech signal. This paper presents an analysis of correlations between various acoustic speech parameters and speaker's height for a large number of speakers. It also presents a new method for an optimal estimation of speaker's height and VTL from various acoustic speech parameters.

1. Introduction

Although it was observed for a long time, the correlation between height and VTL of speakers was not extensively analyzed. Recent studies based on X-ray and magnetic resonance imaging (MRI) showed compelling evidence for the anatomical correlation between height and VTL in monkeys [1], and humans [2]. The acoustic theory of speech production suggests that there is also a correlation between speaker's formant frequencies and VTL [3]. Due to lack of extensive databases containing spoken utterances and VTL measurements this correlation was also not comprehensively examined on empirical bases. A very recent study analyzed the correlations between various acoustic features and body-size features in humans and baboons using significant numbers of subjects [4]. The examined acoustic features consisted of pitch and formant frequencies, whereas the body-size features consisted of height, third digit length, neck circumference, and weight. This analysis was made on data from 68 (34 male and 34 female) human subjects and 27 (11 male and 16 female) baboon subjects, independently. Significant average differences between sexes were found in both acoustic and body-size features in both species. However, significant correlations were observed only between certain acoustic and body-size features. The highest correlation of individual acoustic and body-size features was found between formant F4 and height in male human subjects for the vowel schwa.

Listener's capability of estimating speaker's height and weight was investigated in numerous studies. Such a study, analyzed the perceptual estimation depending on phonetic complexity (e.g. vowels, monosyllabic words, bisyllabic words, and sentences) [5]. Good correlations were found between estimated and actual mean values for both weights and heights. However, such results are considered controversial because they are based on mean values of estimated and actual body-size features. A recent reanalysis of these results showed that, in fact, listeners are not very accurate in estimating the height and weight of speakers [6].

A computational study based on data from 105 male speakers and 78 female speakers found no significant correlations between average fundamental frequency (pitch), on one side, and height and weight, on another side [7]. Another study [8] proposed an automatic method of estimating speaker's height from Mel-Frequency Cepstrum Coefficients (MFCC) and employed a large database (TIMIT) containing data from 630 speakers [9]. This study trained 11 height-dependent Gaussian mixture models (GMM) with speech from two read sentences (approximately 6 s) from each of the 630 speakers. Only data from speakers within the corresponding height range (mean ± 2.5 cm) of each height-dependent GMM were used to train each model. The estimation of speaker's height was based on two read sentences (approximately 6 s), different from those used in training. Sex dependent analyses showed relatively low correlation coefficients between estimated and actual heights of speakers (0.3143 for female and 0.3924 for male speakers). When collapsing the data from both sexes the analysis showed that only 42.1% of the estimated heights were within ± 2.5 cm and 72.1% of the estimated heights were within ± 5 cm. These authors suggested "…that the height of the unknown caller can not be conclusively predicted using the method outlined in Sec. IV."

Because extensive databases containing spoken utterances and VTL measurements are not available, a method of estimating VTL from speech signal was proposed in [10], based on speech data obtained from an accurate articulatory model by synthesizing vowel sequences using 250 different VTL values linearly distributed between 14.96 cm and 18.75 cm. The error in estimating VTL from synthetic speech was less than 1%, on average, and 3.2% maximum. This showed that if accurate VTL measurements are available for a relatively large number of speakers then accurate VTL values can be automatically estimated from speech signal, by averaging the results over different vowels. An unsupervised method of estimating VTL from formant frequencies over sentence level utterances was presented in [11] based on data from 164 speakers in TIMIT. Because actual VTL values for the TIMIT speakers are not available, correlations between the estimated VTL values and actual heights of speakers were computed instead. Using four different estimation methods

these correlations were a little higher than 0.7 in each case, when 8 sentences were used to average the estimated VTL values for each speaker.

The current paper presents an extensive analysis of the correlations between various acoustic speech features and heights of speakers from the TIMIT database. Then it proposes a new method for optimal estimation of speaker's height and VTL using a phone-based approach. The estimation of speaker's height from speech signal has direct applications in forensic analysis whereas the estimation of VTL is relevant to automatic speech recognition for speaker normalization or for employing appropriate acoustic models to the speaker's VTL.

2. Data

This study used the training part of the TIMIT American English speech corpus [9], containing utterances from 462 (326 male and 136 female) speakers. The speakers' heights, provided in the database at the resolution of one inch, range between 144.78 cm and 198.12 cm with an average value of 175. 58 cm. The speech signals, digitized at 16 KHz, from all the 10 sentences from each speaker were included in the correlation analyses. This database contains a phonetic transcription and segmentation using a set of 61 phonetic symbols [9].

3. Analysis of correlations between speaker's height and different acoustic speech features

Unlike most of the previous studies that analyzed the correlation between speaker's height and one type of acoustic feature (usually formant frequencies), this research analyzed the correlations between speaker's height and four types of acoustic speech features. These speech features are: Mel-Frequency Cepstrum Coefficients (MFCC), Linear Predictive Coding (LPC) coefficients, formant frequencies, and fundamental frequency (pitch).

The analysis of the correlation between speaker's height and speech features is the same for each type of feature. This analysis is based on the method of multiple linear regression using least squares. This work employed the MATLAB function *regress* to compute the multiple linear regression. If the speech feature vector P has dimension d, then the multiple linear regression function returns the vector of regression coefficients b of dimension n obtained by solving the linear model

$$y = Xb \tag{1}$$

where y is the $n \cdot 1$ observation vector containing speakers' heights and X is a $n \cdot (d+1)$ matrix containing in each row the speech feature vector P transposed followed by the constant 1. The *regress* function takes as input arguments y, X, and α. The last argument specifies the confidence level and it was set to 0.01 in this study, which corresponds to 99% confidence. The *regress* function returns along with the regression coefficients the R^2, F, and p values for the regression.

A phone-based method of identifying non-linguistic speech features such as, gender, speaker, and language, was proposed in [12]. This method employed phone-based models adapted during the training for each type of feature (e.g. male or female). Then, during the evaluation the incoming speech is first classified into phonemes and then used to compute the likelihood of the corresponding phone-based non-linguistic feature models. The final results are obtained by averaging the results over all phonemes in the incoming speech utterances.

A similar method to that in [12] was adopted in this research. However, [12] showed that using the given phonetic transcription in the TIMIT database does not improve the identification accuracy as compared with employing an automatic phoneme classification method. For this reason, in the current study the automatic classification of the incoming speech into phonemes was not performed, but, instead, the given phonetic transcription was used.

The first analysis focused on the correlation between speaker's height and MFCC features and it was performed independently for each individual vowel. 10 MFCC features, excluding $MFCC_0$ which represents the total energy, were computed as described in [13], on frames of 32 ms with a frame step of 10 ms for each of the 4620 utterances of the 462 speakers in the training part of TIMIT. This phone-based analysis used only the frames corresponding to one of the 20 vowels present in TIMIT. The goal was to identify the ability of each vowel to preserve the speaker's height characteristics in the MFCC features. Three MFCC frames from the center of each vowel were averaged to obtain a more reliable speech feature vector which, together with speaker's height, was applied to the regression analysis. The phone-dependent correlation results for the 20 vowels are presented in Table 1. The vowels exhibiting correlations above 0.7 between the MFCC features and speaker's height are /iy/, /ae/, /ey/, /ih/, and /eh/. The vowel /uh/ exhibited the smallest correlation R=0.55, corresponding to R^2=0.3039, which means that only 30.39% of the variation in speaker's height is embedded in the MFCC spectral features for this vowel.

Although each vowel preserves to some extent the speaker's height information, this information is redundantly distributed over these vowels. In a second analysis the speaker's height information contained in these vowels was combined in order to eliminate the redundant information. To combine this information only the vowels that have been uttered by all 462 speakers were clustered together. These vowels are /iy/, /ih/, /eh/, /ae/, /aa/, /ay/, /ao/, /ow/, and /ix/. The combined vowels exhibited a correlation between speaker's height and $MFCC_{1\text{-}10}$ features of R=0.7426, corresponding to R^2=0.5515, as presented at the bottom of Table 1.

A subsequent analysis focused on identifying the correlation of individual MFCC features with the speaker's height in the combined vowel case. These results are presented in Table 2. The highest correlation R=0.6879, was found for $MFCC_7$. However, as in the vowel case, the individual features exhibit a high redundancy in preserving speaker's height since the combined correlation is R=0.7426.

The next experiment examined the correlation between speaker's height and LPC features. 16 LPC features were computed, as described in [13], on 32 ms frames with 10 ms frame steps. Due to space limitations only the combined vowel results are presented for the LPC features in this paper. The correlation results between speaker's height and the LPC features are presented in Table 3. This correlation is smaller that the one obtained for the MFCC features, presented at the top of this table. Table 3 also presents the correlation results between speaker's height and the first 5 formant frequencies (F_1-F_5), as well as between speaker's height and the fundamental frequency (F_0). The formant frequencies were computed as the first 5 picks in the LPC spectra.

Table 1: Phone-based correlation results between speaker's height and $MFCC_{1-10}$ features

Vowel	R	R^2
/iy/	0.7254	0.5262
/ih/	0.7027	0.4938
/eh/	0.7014	0.4920
/ey/	0.7060	0.4984
/ae/	0.7165	0.5134
/aa/	0.6440	0.4148
/aw/	0.6468	0.4148
/ay/	0.6754	0.4561
/ah/	0.6395	0.4089
/ao/	0.5877	0.3454
/oy/	0.5845	0.3417
/ow/	0.6544	0.4283
/uh/	0.5513	0.3039
/uw/	0.5753	0.3310
/ux/	0.6788	0.4607
/er/	0.6981	0.4873
/ax/	0.6492	0.4214
/ix/	0.6898	0.4759
/axr/	0.6934	0.4808
/ax-h/	0.3626	0.1315
Overall	0.7426	0.5515

In the very recent study cited earlier [4], a detailed correlation analysis was performed between speaker's height and individual formant frequencies (F_1 – F_4) for the schwa vowels. The highest correlation was obtained between speaker's height and F_4 (R=0.7615, R^2=0.58). In the current study, a similar correlation analysis was performed for each individual formant frequency (F_1 – F_5). These correlations are as follows: R=0.7077 (F_1), R=0.6730 (F_2), R=0.6807 (F_3), R=0.6153 (F_4), and R=0.5805 (F_5).

The combined-vowel analysis showed similar correlations obtained for the LPC features and the first 5 formant frequencies. The correlation between speaker's height and fundamental frequency is smaller, but still quite significant.

Table 2: Correlations between speaker's height and individual MFCC features

Feature	R	R^2
$MFCC_1$	0.4671	0.2182
$MFCC_2$	0.5954	0.3545
$MFCC_3$	0.5048	0.2548
$MFCC_4$	0.6717	0.4512
$MFCC_5$	0.6361	0.4046
$MFCC_6$	0.6512	0.4240
$MFCC_7$	0.6879	0.4732
$MFCC_8$	0.6586	0.4338
$MFCC_9$	0.6778	0.4595
$MFCC_{10}$	0.6258	0.3916
$MFCC_{1-10}$	0.7426	0.5515

By building a combined feature vector containing all the four types of features the multiple regression analysis shows a higher correlation value, presented as Overall in Table 3. By excluding the fundamental frequency from this combined feature vector the correlation between the speaker's height and the combined speech features reaches the highest value R=0.7560 (at the bottom of Table 3). This means that about 57% of the variability in the speaker's height is embedded in the combined features (MFCC + LPC + formant frequencies).

Table 3: Correlations between speaker's height and MFCC features, LPC features, formant frequencies, and fundamental frequency

Feature	R	R^2
$MFCC_{1-10}$	0.7426	0.5515
LPC_{1-16}	0.7286	0.5309
F_{1-5}	0.7264	0.5277
F_0	0.5880	0.3458
Overall	0.7556	0.5709
All except F_0	0.7560	0.5715

4. Estimation of speaker's height and VTL

The multiple linear regression analysis can be used to obtain the phone-based (in this paper vowel-based) regression coefficients vector b which transforms the original d-dimensional speech feature vector into a one-dimensional value approximating speaker's height. The matrix equation can be expressed as follows

$$\hat{y} = Xb \qquad (2)$$

where $\hat{y}$ represents the estimated n-dimensional vector of height observations. $\hat{y}$ exhibits the same correlation R with the speakers' heights as the original speech feature vectors included in the X matrix. A scattered plot of the actual heights and the estimated heights of the 462 speakers based on the combined feature (MFCC + LPC + formant frequencies) is presented in Fig. 1. The solid line represents the regression line for $\hat{y}$ and y height values. The correlation coefficient R=0.7560, corresponding to R^2=0.5715, represents the highest correlation obtained in all these experiments. The average difference between actual and estimated height is 5.1 cm.

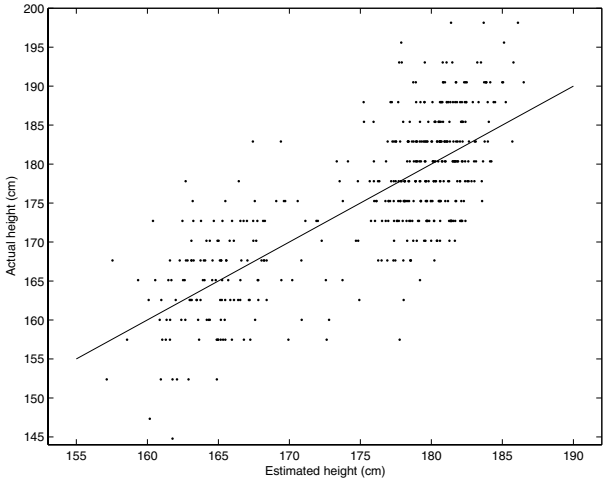

Figure 1: Scatterplot of the actual and estimated heights of the 462 speakers included in the training part of TIMIT.

The estimation of VTL from speech signal performed in this study was based on a classical computation method from

formant frequencies using the formula $L = c(2k-1)/(4F_k)$, where L is the length of the vocal tract, c is the speed of sound and F_k is the formant frequency number k. Then the estimated VTL is obtained by averaging the L values obtained from all formants. The key difference of this study is that the VTL computations were phone-based (vowel-based). The formant frequencies of three frames from the center of the vowels were averaged and the resulting 5 formants were used in the VTL computation for each vowel. Then, as in Section 5, the speaker's VTL was estimated by averaging the VTL results over the vowels in the speaker's utterances. Since TIMIT does not provide actual VTL measurements the absolute values of the estimated VTL are not relevant because they cannot be compared.

An indirect, but meaningful, evaluation of these estimated VTL values was performed by computing the correlation between the estimated VTL values and the actual height values of the 462 speakers. This correlation coefficient is R=0.7211 (R^2=0.5200) and it is similar with the values obtained in [11] based on four different methods (R=0.724, R=0.702, R=0.706, R=0.702) on a smaller group containing 164 speakers from TIMIT. That means that only 52% of the estimated VTL correlate with actual speakers' heights. It would be relevant to compare this correlation with real correlations obtained between actual VTL and actual height values from human subjects. In [2] the correlation between actual VTL and actual height values was found to be R=0.926 (R^2=0.86) for 129 subjects. This real, anatomical correlation is close to the correlation obtained in the current study between the estimated VTL and the estimated height values for 462 speakers (R=0.9517, R^2=0.9057). This strong correlation (close to 1.0) shows that the estimated VTL values are obtained with approximately the same accuracy (in fact a little smaller) as that for the estimated heights of speakers for which actual measurements are available.

5. Discussion

The analysis of the relationship between speaker's height and various acoustic speech features showed a correlation coefficient R=0.7560, which means that 57.15% of the variability of speaker's height can be accounted for by the combined features (MFCC + LPC + formant frequencies). This result is significant for a few reasons: (a) previous studies provided controversial results based on different numbers of speakers, (b) a recent study [4] provided a similar value (R=0.7615) between speaker's height and F_4 but only for a single vowel (schwa) and a smaller number of speakers (164), (c) this study combines not only various vowels but also various features, making the estimation more reliable and efficient (e.g. does not require a schwa vowel from speakers), (d) this study obtained the high correlation from a larger sample of speakers (462) than previously reported. The method proposed here can also be extended to all vowels or even to all phonemes regardless if all speakers uttered all of them or not. Averaging can be done on the actual phonemes uttered. This method can be directly applied to estimate the height and VTL of speakers from the speech signal alone.

6. Conclusions

This paper presents a method of obtaining a higher correlation between speaker's height and a combination of various speech features. This method is more robust than previous ones because it depends less on the accuracy of extracting a single feature (e.g. a formant frequency). It is also more efficient because uses phone-specific transformation functions and not a simple averaging over all phones. This method is directly applicable to speaker's height and VTL estimation.

7. Acknowledgements

This research was supported in part by the National Science Foundation under the Creativity Extension of the Knowledge and Distributed Intelligence grant NSF IIS-98-72995. The author thanks James L. Flanagan for this support.

8. References

[1] Fitch, W. T., "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaque", *J. Acoust. Soc. Amer*., 102(2): 1213-1222, 1997.

[2] Fitch, W. T., and Giedd, J., "Morphology and development of the human vocal tract: A study using magnetic resonance imaging", *J. Acoust. Soc. Amer*., 106(3): 1511-1522, 1999.

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

[4] Rendall, D., Kollias, S., and Ney, C., "Pitch (F0) and formant profiles of human vowels and vowel-like baboon grunts: The role of vocalizer body size and voice-acoustic allometry", *J. Acoust. Soc. Amer*., 117(2): 1-12, 2005.

[5] Lass, N. J., DiCola, G. A., Beverly, A. S., Barbera, C., Henry, K. G., and Baldi, M. K., "The effect of phonetic complexity on speaker height and weight identification", *Language and Speech*, 22(4): 297-309, 1979.

[6] Gonzalez, J., "Estimation of speaker's weight and height from speech: A reanalysis of data from multiple studies by Lass and colleagues", *Perceptual and Motor Skills*, 96, 297-304, 2003.

[7] Künzel, H., J., "How well does average fundamental frequency correlate with speaker height and weight?", *Phonetica*, 46: 117-125, 1989.

[8] Pellom, B. L., and Hansen, J. H. L., "Voice analysis in adverse conditions: The Centennial Olympic Park bombing 911 call", Proc. IEEE Midwest Symposium on Circuits and Systems, California, 1997.

[9] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[10] Dusan, S., and Deng, L., "Vocal-tract length normalization for acoustic-to-articulatory mapping using neural networks," *J. Acoust. Soc. Amer*., 106(4), Pt. 2, p 2181(A), 1999.

[11] Necioğlu, B. F., Clements, M. A., and Barnwell III, T. P., "Unsupervised estimation of the human vocal tract length over sentence level utterances," Proc. IEEE ICASSP 2000.

[12] Lamel, L. F., and Gauvain, J.-L., "A phone-based approach to non-linguistic speech feature identification," *Computer Speech and Language*, 9: 87-103, 1995.

[13] Davis, S., and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition," *IEEE Trans. on Acoust., Speech, and Signal Process.* 28 (4), 357-366, 1980.