5aSCb44. Estimated relative vocal tract lengths from vowel spectra based on fundamental frequency adaptive analyses and their relations to relevant physical data of speakers

Mayuko Kobayashi, Ryuichi Nisimura, Toshio Irino and Hideki Kawahara*

*Corresponding author's address: Design Information Sciences, Wakayama University, Wakayama, 640-8510, Wakayama, Japan, kawahara@sys.wakayama-u.ac.jp

A Japanese vowel database of males, females and children speakers (385 speakers in total) along with relevant physical data (Deguchi et al. 2011) was analysed using a set of F0 adaptive procedures, which were developed for a speech analysis, modification and synthesis framework TANDEM-STRAIGHT (Kawahara et al. 2008) and its extensions. By restricting spectral region in estimating ratios for VTLN (vocal tract length normalization), cepstrally weighted distance measure yielded estimates with 2% standard error, by assuming relative vocal tract lengths estimated using regression analysis as the ground truth. A set of regression analyses of the estimated relative vocal tract lengths, average F0s, ages, gender and physical data (speakers’ height, weight) were conducted. The results suggest that the proposed analysis procedures applied to the Japanese five vowels may provide sufficient information for estimating speakers' physical data. Possible applications of the proposed estimation procedures will also be discussed.
INTRODUCTION

Humans well outperform automatic speech recognition systems in everyday-life conditions, where they have to adapt to divergent speakers usually in adverse conditions [1]. Vowel-based vocal tract length normalization may play an important role in this superior human performance. There are several supporting findings. Size perception of the sounding object, which is believed to be built-in our auditory perception [2, 3], enables humans to normalize effects due to vocal tract length of different speakers [4]. Very rapid adaptation to unknown speakers in monosyllabic speech perception [5, 6] suggests that human adaptation is using information embedded in vowels.

In this article, a new method to estimate vocal tract length ratio between two set of vowel spectra is proposed based on an interference-free power spectral envelope estimation procedure [7]. The proposed procedure was tested by a set of analyses using a Japanese vowel database of men, women and children speakers (385 speakers in total) along with relevant physical data [8] and indicated that the standard error in vocal tract length ratio estimation is around 2%. The proposed estimation method combined with the vowel database with the relevant physical data enables to estimate speaker’s height and weight from vowel information.

BACKGROUND

Vocal tract length normalization is a commonly used technique in automatic speech recognition systems [9]. However, many of them are not literally normalize vocal tract length alone, since the spectral shape variation is affected not only by the vocal tract length but also by other factors such as individual and developmental shape variations [10], acoustic branching by the piriform fossa [11], glottal source waveform [12] and three dimensional vocal tract shape [13]. Recently, a reliable vocal tract length ratio estimation method was proposed [14] by combining frequency range selection and the auditory model based on the dynamic compressive gammachirp filter bank (dcGCFB) [15], for automatic spectral alignment in temporally variable multi-aspect speech morphing [16].

A new vowel-based method replaces this auditory model with a set of simple spectral representations based on a short time Fourier transform and only uses vowel information. This simplification makes the proposed method practically useful and easy to be designed for target applications.

Vowel Database with Relevant Physical Data

Recently a large Japanese vowel database was compiled using men, women and children speakers ranging between 6 and 56 ages in standard Japanese dialect [8]. More specifically, there are 199 female speakers ranging from 6 to 56 years old and 186 male speakers from 6 to 47 years old. The utterances were acquired using the omni-directional miniature condenser microphone (DPA-4061) and edited to select isolated syllables /haa hii huu hee hoo/, which were originally embedded in each carrier sentence. In this database, relevant physical data such as height and weight are available for 284 (146 male and 138 female) speakers. This provides an important scientific basis for speech production and perception research and is an ideal test material for our proposed method.

VOCAL TRACT LENGTH RATIO ESTIMATION USING VOWELS

This section introduces the proposed method. The proposed method consists of three sub-procedures; vowel template preparation, spectral equalization and smoothing, and
minimization of the distance measure for ratio estimation. The following sections introduce details.

**Vowel Template Preparation**

Stable vowel portions are selected from the target utterance and used to calculate the vowel template. An F0 (fundamental frequency) adaptive procedure for estimating the interference-free spectral representation is used to calculate the frame-wise spectral slice. The interference-free representation $P_S(\omega, t)$ is calculated from a power spectrum $P(\omega, t)$ based on short time Fourier transform using following equations.

$$P_S(\omega, t) = \frac{1}{\omega_0} \int_{-\omega_0}^{\omega_0} h(\lambda)P(\omega - \lambda, t)d\lambda$$

$$h(\omega) = \begin{cases} 1 - \frac{\omega}{\omega_0}, & (|\omega| \leq \omega_0) \\ 0, & (|\omega| > \omega_0) \end{cases}$$

where $\omega_0 = 2\pi f_0$ is the fundamental angular frequency. The time window for calculating the power spectrum is set long enough to separate individual harmonic components. The vowel template of the $k$-th speaker $G(k)$ is defined as a set of averaged logarithmic power spectra $L(v,k)(\omega)$ in each vowel segment, where $k$ and $v \in V = \{/a/, /i/, /u/, /e/, /o/\}$ represent the speaker and the vowel type (phoneme) respectively.

$$G(k) = \{L(/a/,k)(\omega), L(/i/,k)(\omega), L(/u/,k)(\omega), L(/e/,k)(\omega), L(/o/,k)(\omega)\}$$

$$L(\omega) = \frac{1}{\#(F)} \sum_{n \in F} 10 \log_{10}(P_S(\omega, t(n)))$$

where $F$ represents a set of indices of frames associated with a specific speaker and a specific vowel. The function $\#(F)$ returns the cardinal number of the set $F$ and $t(n)$ returns the temporal location of the frame $n$.

**Spectral Equalization and Smoothing**

Prior to distance calculation, spectra in each template are equalized and smoothed to eliminate disturbing factors in vocal tract length comparison. The equalization process removes global spectral shape from the observed vowel template spectra, because the global spectral slope of vocal tract transfer functions is essentially zero. The equalized spectrum $P_E^{(v,k)}(\omega)$ is calculated from a constituent template spectrum $P^{(v,k)}(\omega) = 10^{L^{(v,k)}(\omega)/10}$.

$$P_E^{(v,k)}(\omega) = \frac{P^{(v,k)}(\omega)}{w_G(\lambda)} \int_{-\omega_W}^{\omega_W} w_G(\lambda)d\lambda$$

where $w_G(\omega)$ is the smoothing kernel for calculating global spectral shape. The size of kernel $2\omega_W = 4\pi f_W$ is two to four times wider than the average spacing of formant frequencies. In the current implementation, a raised cosine function $(1 + \cos(\pi \omega/\omega_W))$ is used as the smoothing kernel.

The following smoothing process removes spectral details such as dips due to glottal source waveform, acoustic branching by piriform fossa and three dimensional wave propagation. The final equalized and smoothed spectrum $P_N^{(v,k)}(\omega)$ is calculated from the equalized spectrum
The proposed method consists of four parameters $f_W$, $f_N$, $f_L$ and $f_H$. They have to be properly selected, because the accuracy of the vocal tract length ratio is dependent on these parameters.

A regression-based method is used to assess the accuracy of the spectrally estimated values, because no ground truth is directly observable. By integrating all available ratio data using the following procedure provides the best available approximates of relative vocal tract lengths.

**TUNING OF PERFORMANCE DETERMINING PARAMETERS**
Estimation of Relative Vocal Tract Lengths from VTL Ratios

Because the vocal tract length ratio is linearized by logarithmic conversion, the following set of linear equation provides the estimate of the logarithmic conversion of the relative vocal tract length.

Let the vector $r$ represent the set of logarithmic conversion of the vocal tract length ratio of all combinations of the speakers in the database and let vector $l$ represent the set of logarithmic conversion of the normalized vocal tract lengths of all speakers in the database. Then, $r$ is modeled by the following equation by introducing the connection matrix $H$ and the normalization condition represented in the last row of $H$.

$$ r = Hl + n, $$

where $n$ represents the observation noise. The vectors and the connection matrix are defined as follows.

$$ r = \begin{bmatrix} \log(r_{1,2}), \log(r_{1,3}), \ldots, \log(r_{k,p}), \ldots, \log(r_{N,N-1}), 0 \end{bmatrix}^T $$

$$ l = \begin{bmatrix} \log(l_1), \log(l_2), \ldots, \log(l_N) \end{bmatrix}^T $$

$$ \langle H \rangle_{m,n} = \begin{cases} 1 & (m = k) \\ -1 & (n = p) \\ 0 & (m \neq k, n \neq l) \end{cases} $$

where $k$ and $p$ are from: $\langle r \rangle_m = \log(r_{k,p})$

$$ H|_{\text{last}} = [1, 1, \ldots, 1], $$

where $N$ represents the number of speakers and $\langle H \rangle_{m,n}$ represents the element of $H$ on the $m$-th row and the $n$-th column. Also, $H|_{\text{last}}$ represents the last row of $H$.

The least square estimate $\hat{l}$ provides the approximate ground truth of the relative vocal tract length $\hat{l}_k$.

$$ \hat{l} = (H^T H)^{-1} H^T R $$

$$ \hat{l}_k = \exp (\{l \}^T_k), $$

where $\{l \}_k$ represents the $k$-th element of the vector $\hat{l}$.

Performance Measure of Spectrum-Based Estimation

Using the approximate ground truth of relative length $\hat{l}_k$ and $\hat{l}_p$, also the approximate ground truth of the ratio $\hat{r}_{k,p} = \hat{l}_k/\hat{l}_p$ is defined. The normalized root mean squared deviation from this value provides the performance measure $\eta(f_W, f_N, f_L, f_H)$ of the spectrum-based estimation.

$$ \eta(f_W, f_N, f_L, f_H) = \left( \frac{1}{N(N-1)} \sum_{k \in S} \sum_{p \in (S-\{k\})} \left| \hat{r}_{k,p}(f_W, f_N, f_L, f_H) - \hat{r}_{k,p}(f_W, f_N, f_L, f_H) \right|^2 \right)^{\frac{1}{2}}, $$

$$ \hat{r}(f_W, f_N, f_L, f_H) = \frac{1}{N(N-1)} \sum_{k \in S} \sum_{p \in (S-\{k\})} r_{k,p}(f_W, f_N, f_L, f_H) $$

where the set of parameters $(f_W, f_N, f_L, f_H)$ is explicitly denoted for the spectral-based estimate and the regression-based estimate. The symbol $S$ represents the set of the speaker IDs and $S - \{k\}$ represents the set of the speaker IDs excluding the $k$-th speaker's. Preliminary tests using the vowel database indicated that $f_W = 2500$ Hz, $f_N = 300$ Hz, $f_L = 600$ Hz and $f_H = 4000$ Hz provides the best performance. The performance measure using this set of
parameters was 0.133 and the standard error of the spectrum-based vocal tract length ratio was 0.0196. Figure 1 shows the landscape of the performance measure for parameter combinations.

The left plot of Fig. 2 shows the distribution of the relative vocal tract length $\hat{l}_k$ for each gender. The vertical axis of the plot is the probability $P_{l_k}(l_\theta | l_k < l_\theta)$, where $l_\theta$ is the value on the horizontal axis. The middle plateau in the male distribution may reflects the shape change due to puberty.

The right plot of Fig. 2 shows the histogram of the difference between the spectrum-based estimate $r_{k,p}$ and the regression-based estimate $\hat{r}_{k,p}$. The size of the histogram bin of the right plot is 0.005. The total number of speaker pairs is 147,840. The histogram indicates that majority of the spectrum-based ratio are within ±5% error.

**ESTIMATION OF HEIGHT AND WEIGHT FROM VOWELS**

Relation between the regression-based estimates of relative vocal tract length and the speakers’ height and weight were analyzed. First of all, the recorded speakers’ height and
weight are shown in Fig. 3 as functions of the speakers' age. These physical data are not recorded for speakers with age 13 and 14 in this database. The plot marks are color coded based on the speakers' gender (blue for male and red for female speakers).

The left plot of Fig. 4 shows relation between the speakers' age and their relative vocal tract lengths. The right plot of Fig. 4 shows relation between the speakers' age and their average fundamental frequencies. The plot marks are also color coded based on their gender. All 385 speakers are shown in these plots. In Fig. 3 and Fig. 4, there seems to be monotonic relations with the age up to puberty. These plot suggest that the estimated vocal tract length will show clear relation with these relevant physical data, if the estimates reflect the actual physical data.

Figure 5 shows relations between the regression-based vocal tract length (VTL) and speakers' height and weight. Table 1 shows linear regression results. Linear regression lines are also plotted in Fig. 5. Analysis results indicated that the regression coefficient and the intercept are both statistically significant. These plots and tables illustrate that it is possible to estimate the speakers' height and weight from vowel information. Multiple regression analyses using vocal tract length and fundamental frequency indicated that contribution of fundamental frequency is not statistically significant.
CONCLUSION

A new method to estimate relative vocal tract length ratio between two set of vowel spectra is proposed. It is based on an interference-free representation of power spectra and yields vocal tract length ratio estimates with standard error around 2%. The estimated relative vocal tract length is found to correlate well with the physical data recorded on the database. The proposed method is implemented using computationally efficient FFT and linear interpolation. This performance and efficient implementation are useful aspects of the proposed method for various applications such as vowel-based voice conversion, speech recognition, speaker authentication and so on.

ACKNOWLEDGMENTS

The work was supported partly by JSPS KAKENHI Grant Number 24650085, 21300069 and 24300073.

REFERENCES


