Vocal Tract Length Estimation for Voiced and Whispered Speech Using Gammachirp Filterbank

Toshio Irino, Erika Okamoto, Ryuichi Nisimura, and Hideki Kawahara
Faculty of Systems Engineering, Wakayama University, Japan
E-mail: {irino, nisimura, kawahara}@sys.wakayama-u.ac.jp Tel: +81-43-457-8475

***Abstract*—In this paper, we demonstrate an auditory spectrogram based on a dynamic compressive gammachirp filterbank (GCFB) that enables accurate and robust estimation of vocal tract length (VTL) for both voiced and whispered speech. Normalized VTLs of 21 speakers were derived by using the least squared analysis of their VTL ratios (for all permutations, 420 = $_{21}P_{20}$) which were estimated by minimizing spectral distances in the auditory spectrograms. The frequency range was selected in the calculation and the range between 500 and 5000 (Hz) was most reasonable for both speech mode. The method based on GCFB was better than that based on the mel-frequency filterbank (MFFB). The estimated VTLs were compared with the VTL data measured in MRI to confirm the reliability.**

I. Introduction

Accurate and robust estimation of vocal tract length (VTL) is important for many speech applications including automatic speech recognition and speaker classification [1]. A VTL estimation method was also introduced in a voice morphing technique, which is widely used in researches for speech perception and singing voice manipulations [2], to improve the sound quality.

Speech sounds convey information about nonlinguistic speaker characteristics in addition to linguistic contents. Humans can easily identify a speaker as male, female, or child from the speech sounds, even with a single word or a monosyllable [3] and even with whispered speech[4]. This suggests that the auditory system effectively extracts VTL information separately from vocal tract shape information. The process was modeled as the stabilized wavelet- Mellin transform (SWMT) [5] in which the cochlear filtering is reasonably modeled by using the dynamic compressive gammachirp filterbank (GCFB) [6].

We previously proposed a VTL estimation method that uses GCFB as a simplified and effective version of SWMT. The method was applied to voiced speech to demonstrate the method based on GCFB outperformed other methods based on the conventional mel-frequency filterbank (MFFB)[7] [8].

In this paper, we demonstrate that the GCFB-based method is also successful in estimating the VTL from whispered speech sounds and outperforms the MFFB-based methods. The estimated VTLs were also compared with VTL data measured in magnet resonance image (MRI) [9] to evaluate the reliability.

II. VTL Estimation Method

The relative VTLs were estimated from auditory spectrograms as described below. The least squared (LS) analysis is then applied to evaluate the estimation error.

A. VTL ratio estimation based on spectral distance

The speech samples of speakers A and B with the same sentence were analyzed by using the gammachirp filterbank (GCFB), gammatone filterbank (GTFB), or mel-frequency filterbank (MFFB) (see section II-B3) to derive smoothed spectrograms $P_A(\tilde{f},t)$ and $P_B(\tilde{f},t)$, where $\tilde{f}$ is either ERB frequency f_{ERB} in GCFB and GTFB or mel-frequency f_{mel} in MFFB. Since the phoneme locations in the two spectrograms are different, we deformed the time axis of spectrogram B to align the phoneme boundaries for B with those for A. Deformed spectrogram B is denoted as $P_{Bn}(\tilde{f},t)$. To estimate the VTL ratio between A and B, $P_{Bn}(\tilde{f},t)$ is dilated or compressed along the linear frequency axis with a scaling factor, r, to become $P_{Bn}((\tilde{rf}),t)$. The spectral distance in the dB scale at time t is defined as the root mean squared (rms) difference as

$$D_{dB}(t,r)=\sqrt{\frac{D_P}{\tilde{f_H}-\tilde{f_L}}} \quad , \tag{1}$$

where

$$D_P=\int_{\tilde{f_L}}^{\tilde{f_H}}\left(10\log_{10}\frac{P_A(\tilde{f},t)}{\bar{P}_A(t)}-10\log_{10}\frac{P_{Bn}((\tilde{rf}),t)}{\bar{P}_{Bn}(t)}\right)^2 d\tilde{f},$$

where $\tilde{f_L}$ and $\tilde{f_H}$ are the warped versions of the lower and higher limits of the frequency region (f_L and f_H) and $\bar{P}_A(t)$ and $\bar{P}_{Bn}(t)$ are the values averaged across the frequencies.

The objective of the VTL estimation is to find the best r (scaling factor), which minimizes distance D_{dB}^{total}:

$$r=\mathrm{argmin}(D_{dB}^{total}(r)), \tag{2}$$

where total distance $D_{dB}^{total}(r)$ is defined by using frame-wise spectral distance $D_{dB}(t,r)$ in Eq. 1:

$$D_{dB}^{total}(r)=\sqrt{\frac{1}{T}\int_0^T D_{dB}^2(t,r)dt}, \tag{3}$$

where T represents the time for the final frame. Thus, the estimation is performed for the whole sentence regardless of voiced or whispered speech.

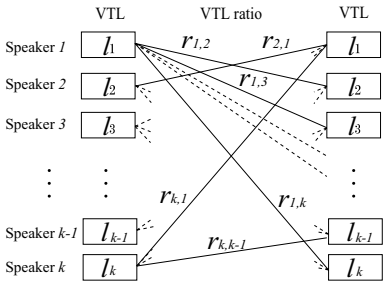

Fig. 1. *Strategy of VTL estimation from all permutation. VTL ratio between two speakers, $r_{m,n}$, is calculated from individual VTLs, l_n.*

B. Accuracy measure for VTL estimation

It is essential to define a measure to evaluate the accuracy of the estimated VTL. In this paper, we used the rms error between the individual VTL ratios estimated by using Eq. 2 and the VTL ratios calculated from the whole set of the individual VTL ratios as described shortly.

1) VTL estimation from individual VTL ratios: We estimated the relative VTLs between 21 speakers. When the m-th and n-th speakers' VTLs are l_m and l_n as shown in Fig. 1, the VTL ratio is defined as $r_{m,n} = l_m/l_n$. By introducing logarithmic conversion, it becomes subtraction as

$$\log(r_{m,n}) = \log(l_m) - \log(l_n). \tag{4}$$

It can be converted into a vector notation as

$$\begin{aligned} log(r_{m,n}) &= [0, 0, ..., 1, ..., -1, ...] \times \\ &\quad [\log(l_1), \log(l_2), ..., \log(l_m), ..., \log(l_n), ...]^T. \end{aligned} \tag{5}$$

We set a vector $\log(r_{m,n})$ as $\boldsymbol{r}_{log}$, a vector for all speaker $\log(l_m)$ as $\boldsymbol{l}_{log}$, and a coefficient matrix $\boldsymbol{H}$ as a set of vectors of 1, 0, and -1 together with a uniform vector for normalization. With this formation, it is possible to estimate the relative VTL between the individual speakers, but not the absolute VTL in cm. The relationship is then rewritten as

$$\boldsymbol{r}_{log} = \boldsymbol{H}\boldsymbol{l}_{log}. \tag{6}$$

The least squared (LS) analysis is applied to estimate the normalized VTLs, $\hat{\boldsymbol{l}}$ $(= [\hat{l_1}, \hat{l_2}, ..., \hat{l_{21}}])$.

$$\begin{aligned} \hat{\boldsymbol{l}}_{log} &= (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{r}_{log}, \\ \hat{\boldsymbol{l}} &= [\hat{l_1}, \hat{l_2}, ..., \hat{l_{21}}]^T = \exp(\hat{\boldsymbol{l}}_{log}), \end{aligned} \tag{7}$$

The VTL ratio, $\hat{\boldsymbol{r}}$, from the LS analysis is estimated as

$$\hat{\boldsymbol{r}} = \exp(\boldsymbol{H}\hat{\boldsymbol{l}}_{log}). \tag{8}$$

The estimation error is evaluated in terms of the rms difference or Euclidean norm d_{est} between the VTL ratios from the spectral distance, $\boldsymbol{r}$, and the VTL ratios from the LS analysis, $\hat{\boldsymbol{r}}$.

$$d_{est} = ||\boldsymbol{r} - \hat{\boldsymbol{r}}|| \simeq \sigma. \tag{9}$$

Note that d_{est} is almost the same as standard deviation σ around the identity mapping line ($\hat{\boldsymbol{r}} = \boldsymbol{r}$) when the bias is small. The estimation is accurate when d_{est} is small, since the the individual VTLs are consistent with the VTLs derived from the least squared analysis of all permutations.

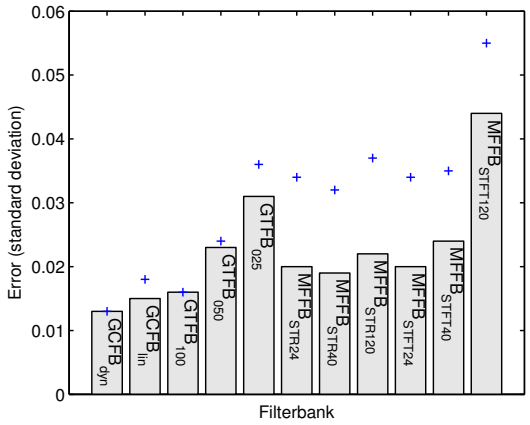

Fig. 2. *Comparison of filterbanks for VTL estimation from voiced sounds [1][8]. Bar shows minimum estimation error (Eq. 9) for each filterbank condition. Plus (+) shows error when $[f_L, f_H]$=[500, 5000].*

TABLE I

Filterbank in Fig. 2	Description	#Channel
$\mathrm{GCFB_{dyn}}$	dynamic compressive GCFB	100
$\mathrm{GCFB_{lin}}$	linear GCFB	100
$\mathrm{GTFB_{025}}$	gammatone filterbank	24
$\mathrm{GTFB_{050}}$	(linear)	50
$\mathrm{GTFB_{100}}$		100
$\mathrm{MFFB_{STR24}}$	mel-frequency filterbank	24
$\mathrm{MFFB_{STR40}}$	based on TANDEM-STRAIGHT	40
$\mathrm{MFFB_{STR120}}$	spectrogram (linear)	120
$\mathrm{MFFB_{STFT24}}$	mel-frequency filterbank	24
$\mathrm{MFFB_{STFT40}}$	based on STFT spectrogram	40
$\mathrm{MFFB_{STFT120}}$	(linear)	120

2) Selection of best frequency region: We need to select frequency region $[f_L, f_H]$ in Eq. 1 for the reliable estimation of VTL, because the VTL information in the low and high frequency regions are smeared by other speech characteristics. In low frequencies, the spectrum is largely affected by the rate and the shape of the glottal pulse. In high frequencies, there are spectral zeros caused by the resonances of the pyriform fossa [10] which differs individually. The spectral components in the middle frequency region are not largely affected by these factors and, thus, effective for VTL estimation. In other words, the individual VTL ratio, $\boldsymbol{r}$, is estimated as a function of the selected frequency region. The objective here is to select the best frequency region to minimize the estimation error d_{est} ($\simeq \sigma$) in Eq. 9.

3) Auditory spectrogram: The input speech sound is coverted into two-dimensional cochlear spectrograms by using GCFB or GTFB. The number of channels was 100 for sufficient filter density, and the center frequencies of the filters were equally spaced on the ERB_N-number axis between 100 and 6000 Hz. An equal-loudness contour (ELC) filter was also applied to simulate the sensitivity. The power of the filterbank outputs was summarized every 5 ms with a 25-ms hamming window to reduce the periodic components that are dependent on the fundamental frequency (F0).

For comparison, we also calculated auditory-like spectrograms from the output of mel-frequency filterbanks (MFFBs). The linear-frequency spectrogram for the MFFB was derived by using either STFT or TANDEM-STRAIGHT.

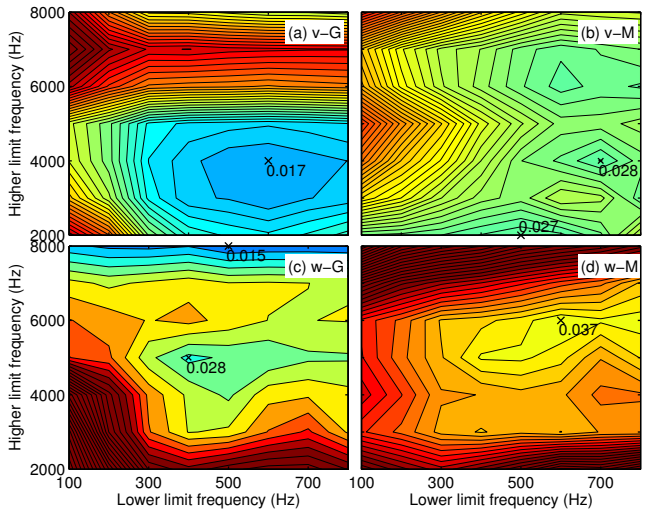

Fig. 3. *Rms error as a function of lower and upper limits of the frequency region* $[f_L, f_H]$*. (a) Voiced speech with* $\mathrm{GCFB_{dyn}}$*, (b) voiced speech with* $\mathrm{MFFB_{STR40}}$*, (c) whispered speech with* $\mathrm{GCFB_{dyn}}$*, and (d) whispered speech with* $\mathrm{MFFB_{STR40}}$*. ×: Global and local minima with error value.*

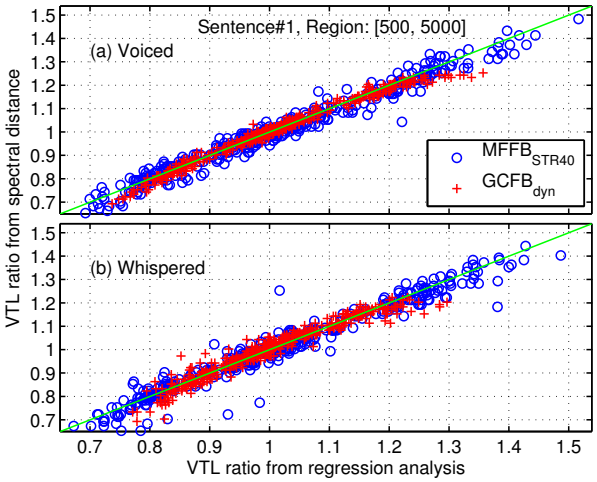

Fig. 4. *Scatter plot of VTL ratios estimated from (a) voiced and (b) whispered speech by using* $\mathrm{GCFB_{dyn}}$ *(+) and* $\mathrm{MFFB_{STR40}}$ *(○).*

III. Results

A. VTL estimation from voiced speech

In the previous papers [1][8], we reported a comparison of auditory spectrograms for VTL estimation from the voiced speech of 28 speakers. The result of the VTL estimation error is summarized in Fig. 2, where the type of filterbank is described in Table I. As a consequence, a dynamic compressive gammachirp filterbank ($\mathrm{GFFB_{dyn}}$) gave the smallest error. We selected $\mathrm{GFFB_{dyn}}$ and the best MFFB ($\mathrm{MFFB_{STR40}}$) for comparison in VTL estimation from whispered speech.

B. Voiced and whispered speech database

We collected voiced and whispered speech samples to produce a database to evaluate VTL estimation and to analyze the relationship between the estimated VTL and the speaker height. The speakers were14 males and 7 females aged between 21 and 24 years old and their heights are ranged between 147.0 cm and 186.0 cm. Each speaker pronounced 30 Japanese sentences both with voiced and whispered speech in a sound proof room. The speech was recorded monaurally at 48 kHz and in16 bit with a B&K 4003 microphone and Edirol R4-Pro recorder. The microphone was located 30 cm from the mouth of the speakers. In this study, we used speech samples of two sentences that consisted of 10 and 14 syllables.

C. Robust VTL Estimation from whispered speech

The VTL ratios between two speakers (Fig. 1) were calculated for two speech samples by using Eq. 2. We also considered the reverse order since the scaling factor, r, is applied to one side in Eq. 1. The total number of permutations was 420 ($=_{21}P_{20}$).

1) Dependency of frequency region: Figure 3 shows the contour maps of estimation errors for combinations of filterbanks ($\mathrm{GCFB_{dyn}}$ or $\mathrm{MFFB_{STR40}}$) and speech mode (voiced or whispered). The abscissa is a lower limit frequency, f_L, and the ordinate is a higher limit frequency, f_H, for the frequency region used in Eq. 1.

The minimum estimation errors for voiced speech were 0.017 with $\mathrm{GFFB_{dyn}}$ in panel (a) and 0.028 with $\mathrm{MFFB_{STR40}}$ in panel (b). The minimum estimation errors for whispered speech were 0.028 with $\mathrm{GFFB_{dyn}}$ in panel (c) and 0.037 with $\mathrm{MFFB_{STR40}}$ in panel (d). Therefore, $\mathrm{GFFB_{dyn}}$ outperformed $\mathrm{MFFB_{STR40}}$ independently of speech mode. The best frequency regions $[f_L, f_H]$ in the middle of maps were : (a) [600, 4000] Hz, (b) [700, 4000] Hz, (c) [400, 5000], and (d) [600, 6000]. Therefore, the frequency region should be restricted above about 500 Hz for accurate VTL estimation, although it is not a well-known fact in conventional studies on VTL estimation.

2) Relationship between VTL ratios : Figure 4 shows a scatter plot between VTL ratios estimated from the LS analysis $\hat{\boldsymbol{r}}$, and VTL ratios based on the spectral distance $\boldsymbol{r}$. The frequency region [500, 5000] was used for analysis here and in the rest of this paper since this gives small errors for all conditions. It is clear that the points for $\mathrm{GFFB_{dyn}}$ (red pluses) more compactly concentrate to the identity mapping line for both voiced and whispered speech. Moreover, the ratios above about 1.4 estimated with $\mathrm{MFFB_{STR40}}$ (blue circles) does not seem reliable since the maximum ratio of the speaker height was 1.26 (=186.0 cm/147.0 cm) and the height and VTL are linearly correlated as described in section III-D.

3) Robustness of VTL estimation: Figure 5 shows the correlation between VTLs, $[\hat{l_1}, \hat{l_2}, ..., \hat{l_{21}}]$, estimated from voiced and whispered speech for the 21 speakers. There was a strong correlation between two VTL estimates. The points for $\mathrm{GCFB_{dyn}}$ concentrated to the least squared line more, and the coefficient of determination, r^2, was greater than for $\mathrm{MFFB_{STR40}}$. The results imply that $\mathrm{GFFB_{dyn}}$ enabled robust VTL estimation independent of the speech mode.

The slope of the least squared line for $\mathrm{GCFB_{dyn}}$ was slightly less than the unity. The reason is not immediately clear due to the restricted number of samples. However, it is also

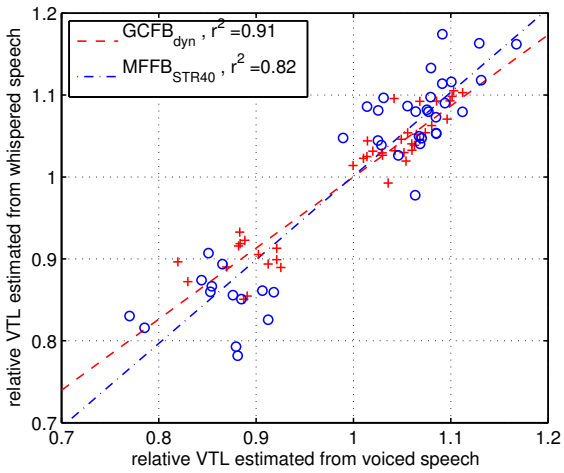

Fig. 5. *Correlation between relative VTLs estimated from voiced and whispered speech. Each point represents VTL combination for one speaker.*

possible to assume that the VTLs were estimated differently in accordance with the speech mode and speaker size.

D. Relationship between VTL and height

The speech database also has speaker height information. We analyzed the relationship between the estimated VTLs and the speaker heights, and compared them with the MRI data obtained by Fitch and Giedd [9]. They reported that the regression line estimated from 121 subjects with an age range from 2 to 25 years is

$$VTL = 2.7 + 0.068 \times Height\ (cm), \qquad (10)$$

where $r = 0.926$, adj. $r^2 = 0.86$, and $p < 0.0001$.

Figure 6 shows the relationship between heights and VTLs estimated with $\mathrm{GCFB_{dyn}}$ (red) and $\mathrm{MFFB_{STR40}}$ (blue) for voiced speech (a) and whispered speech (b). The regression lines are also plotted with the Fitch's data of Eq. 10 with its $\pm 10\%$ (green). The coefficients of determination, r^2, were less than about 0.7 in all cases and less than Fitch's result ($r^2 = 0.86$). This is mainly due to the small numbers of speech samples collected from adult speakers in which the range of height is relatively small. By using $\mathrm{GCFB_{dyn}}$, all VTLs except for one speaker (F05) in Fig. 6(a) were estimated within the variability of $\pm 10\%$ which is observed in Fitch's MRI-VTL data[9]. It is not, however, the case for $\mathrm{MFFB_{STR40}}$. The GCFB enables reasonable VTL estimation for both voiced and whispered speech.

IV. Conclusions

In this paper, we demonstrated a VTL estimation method based on a dynamic compressive gammachirp filterbank (GCFB) that enables accurate VTL estimation from voiced and whispered speech sounds. It was shown that the selection of the frequency range is important and that a range of about [500, 5000] is reasonable for both speech modes. The GCFB-based method was better than the MFFB-based method. The VTLs were reliably estimated within the range of VTLs reported in the MRI study.

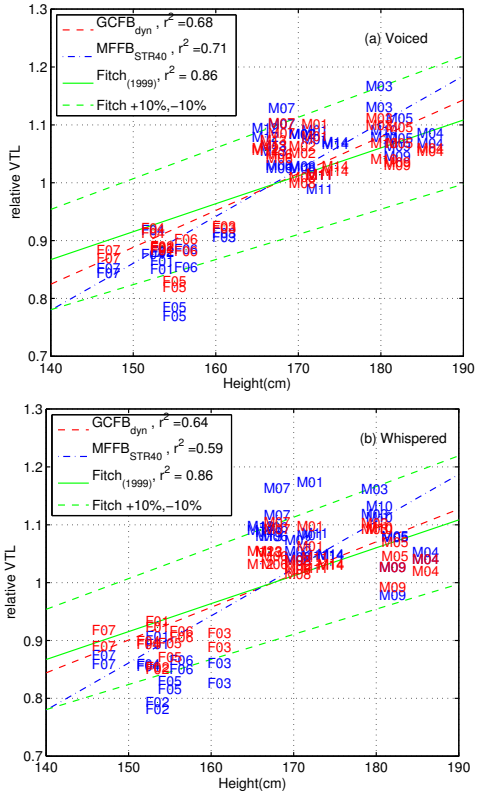

Fig. 6. *Relationship between height and VTL estimated for two sentences from voiced (a) and whispered speech (b). Each label shows speaker ID centered on estimated VTL for one sentence.*

Acknowledgements

This work was supported in part by Grants-in-Aid for Scientific Research 21300069 and 25280063 by JSPS. The authors wish to thank Haruka Kitade for assisting with data analysis.

References

[1] R. Nisimura, *et al.*, "Detecting child speaker based on auditory feature vectors for VTL estimation," *Proc. APSIPA 2012*, #117, Hollywood, California, 3-6 Dec., 2012.

[2] H. Kawahara *et al.*, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP'2009*, pp.3905–3908, 2009.

[3] D. R. Smith *et al.*, "The processing and perception of size information in speech sounds," *J. Acoust. Soc. Am.*, 117(1), pp. 305-318, 2005.

[4] T. Irino *et al.*, "Comparison of performance with voiced and whispered speech in word recognition and mean-formant-frequency discrimination," *Speech Commun.*, 54 (9), pp.998-1013, 2012.

[5] T. Irino and R. D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised Wavelet-Mellin Transform," *Speech Commun.*, 36 (3-4), pp. 181-203, 2002.

[6] T. Irino and R. D. Patterson, "A dynamic compressive gammachirp auditory filterbank," *IEEE Trans. Audio, Speech, and Language Process.*, 14(6), pp.2222–2232, Nov. 2006.

[7] E. Okamoto, *et al.*, "Evaluation of voice morphing using vocal tract length normalization based on auditory filterbank," *J. Signal Processing*, 15(4), pp.283–286, July, 2011.

[8] E. Okamoto, *et al.*, "Auditory filterbank improves voice morphing," *Proc. Interspeech 2011*, pp.2517 – 2520, Florence, Italy, 27-31 Aug., 2011.

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

[10] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans." *J. Acoust. Soc. Amer.*, 101(1), pp. 456–465, 1997.