INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Relating estimated cyclic spectral peak frequency to measured epilarynx length using Magnetic Resonance Imaging

Elizabeth Godoy, Andrew Dumas, Jennifer Melot, Nicolas Malyska, Thomas F. Quatieri

MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA USA 02420
{elizabeth.godoy,andrew.dumas,jennifer.melot,nmalyska,quatieri}@ll.mit.edu

Abstract

The epilarynx plays an important role in speech production, carrying information about the individual speaker and manner of articulation. However, precise acoustic behavior of this lower vocal tract structure is difficult to establish. Focusing on acoustics observable in natural speech, recent spectral processing techniques isolate a unique resonance with characteristics of the epilarynx previously shown via simulation, specifically cyclicity (i.e. energy differences between the closed and open phases of the glottal cycle) in a 3-5kHz region observed across vowels. Using Magnetic Resonance Imaging (MRI), the present work relates this estimated cyclic peak frequency to measured epilarynx length. Assuming a simple quarter wavelength relationship, the cavity length estimated from the cyclic peak frequency is shown to be directly proportional (linear fit slope =1.1) and highly correlated ($\rho = 0.85$, pval$<10^{-4}$) to the measured epilarynx length across speakers. Results are discussed, as are implications in speech science and application domains.
Index Terms: lower vocal tract, epilarynx, MRI, cyclic peak

1. Introduction

Situated between the glottis and the pharynx, the epilarynx is a region of the lower vocal tract (VT) that plays an important role in speech production, regulating vocal loudness and coloring timbre in addition to carrying speaker-specific information [1, 2, 3, 4]. Though typically overlooked in many speech processing applications, the resonance pattern of the lower VT cavities (epilarynx and piriform fossa) shapes the speech spectrum in a prominent and largely static way, occupying a frequency range above 3kHz that avoids highly dynamic lower formants (F1, F2) [1, 2]. Ultimately, in addition to having distinct spectral characteristics, resonance features of the lower VT hold important information both about the individual speaker and the manner in which he or she is speaking.

Inspired by these observations, recent speech processing algorithms have been proposed to estimate VT features that reflect acoustics of the epilarynx and piriform cavities [5]. At the crux of the proposed spectral processing techniques is the exploitation of significant VT resonance differences between the glottal closed and open phases at mid-to-high frequencies (e.g. 3-5kHz), targeting the *cyclicity* property attributed to the (first) epilarynx resonance [6]. Thus, the spectral peak in this frequency range that exhibits this cyclicity property (i.e. the estimated *cyclic* peak) can be treated as an acoustic proxy to the epilarynx resonance. The goal of this work is to relate parameters of this cyclic peak, automatically estimated from speech, to the epilarynx cavity structure measured using high-resolution Magnetic Resonance Imaging (MRI).

Despite being the topic of many studies on the singer's/speaker's formant and "resonant" voice, detailed analyses of the acoustics and morphology of the epilarynx (i.e. laryngeal ventricle and vestibules) are very challenging and have been limited previously by lack of access to high-resolution MRI [7, 3, 8, 9, 10, 11, 12, 13]. Though general behavior of the epilarynx and corresponding acoustics have been established via modeling and simulations (specifically noting a cyclic resonance near 3kHz for males that appears consistently across different vowels [6, 14, 15]), the precise relationship between observed acoustics and measured cavity structure is not clear. For example, some works propose modeling the epilarynx as a Helmholtz resonator [3, 2], while others argue that it behaves more like a quarter wavelength resonator [1, 16]. The relationship between the resonance acoustics and dimensions of the epilarynx, specifically cross-sectional area of its airway with respect to that of the bottom of the pharynx (i.e. hypopharynx), is also disputed. Prior tube models based on coarse X-ray and MRI measurements suggest that a 1:6 (or larger) epilarynx-to-pharynx cross sectional area ratio difference would generate the most clearly observable epilarynx resonance (as in production of a singer's or speaker's formant) [3, 1]. However, studies on real (i.e. not simulated) data across different speakers report a much lower ratio (e.g. 1:3) [4, 17]. The precise means of quantifying these area measurements is also not obvious, as shown in [4].

Though analyses of the epilarynx morphology have recently become more detailed with access to high-resolution 3D MRI, as in [4, 15], examination of the observable (i.e. not simulated) acoustics is typically limited to coarse measurements of long-term average spectral energy in a broad frequency range (e.g. level above 2kHz) [4, 18]. Spectral analyses in these studies have not targeted specific acoustic properties of the epilarynx. Consequently, the recently proposed estimation of a specific cyclic peak in [5] provides an opportunity for more detailed analysis of observable epilarynx acoustics.

Consistent across all previous work is a premise that the epilarynx resonance frequency is inversely proportional to the cavity length. The simplest model proposed for the epilarynx that captures this relationship (independently of area dimensions) is a quarter wavelength resonator [1]:

$$F_e = \frac{c}{4L_e} \tag{1}$$

which describes the resonant frequency F_e as proportional (via a constant $c/4$ where $c = 350m/s$ is the speech of sound in the VT) to the inverse of the cavity length L_e. The present work

*This work is sponsored under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

http://dx.doi.org/10.21437/Interspeech.2016-1362

adopts this quarter-wavelength resonance relationship in order to compare the cavity length estimated from the cyclic peak frequency to the measured epilarynx length. Results indicate that the cavity lengths automatically estimated from the cyclic peak frequencies of several speakers are directly proportional to the epilarynx cavity lengths measured from 3D MRI.

Sections 2 and 3 respectively describe the acoustic and image analyses. Sec. 4 then compares the cavity lengths estimated from the cyclic peak frequencies to the measured epilarynx lengths. Discussion and conclusions follow.

2. Cyclic Peak Estimation

2.1. Speech Audio Data from Real-Time MRI

The speech analyzed in this work is from a real-time MRI (rtMRI) collect done by USC SAIL [19], an extension of which was done for MIT Lincoln Laboratory. First, speech articulation for 15 native English-speaking participants (9 female, 6 male) was captured using rtMRI in the midsagittal plane [20]. Audio was simultaneously recorded and denoised using a model-based approach [21] with a resulting sampling rate of 20 kHz. Participants were instructed to say a sentence that consisted of a series of vowel-consonant-vowel (VCV) words with the vowels /aa/ (“aa”), /uw/ (“oo”), /iy/ (“ee”) and the consonants /th/, /s/, /sh/, /m/, /n/, /l/. The stimuli sentence began with “aathaa oothoo eethee aasaa oosoo eesee.” Data included 3 instances of this sentence for a total of 18 spoken vowel instances per speaker.

Phone-level segmentation of the audio was obtained automatically via forced alignment of subjects’ speech to text of the VCV stimuli. The alignment was performed using an MIT Lincoln Laboratory scala package, that wraps HTK [22], with a neural network acoustic model trained on the publicly available WSJ1 corpus [23], with a phone error rate of 14.8.

2.2. Cyclic Peak Estimation

The speech frames analyzed are from the vowels (/aa/ , /uw/ , and /iy/) in the VCV stimuli. In order to mitigate segmentation errors in the automatic speech-to-text alignment, an additional acoustic measure is used to isolate frames labeled as vowels that also have a high probability of voicing, as in [24]. Specifically, the probability of voicing was calculated here using the PEFAC algorithm [25], designed to be robust to high levels of noise, and a threshold of 0.8 was applied to select the most highly voiced frames [24].

Spectral analyses of the selected frames were carried out using the same algorithms in [5], beginning with spectral envelope analyses (True Envelope [26], cepstral order 40) of full resolution (3 pitch period) frames as well as isolated closed and open phases of the glottal cycle. Candidate frequencies for the cyclic resonance are spectral peaks that exhibit the cyclicity property (maximal difference in spectral energy between the closed and open phases between 3-5kHz). A refinement stage selects an overall cyclic peak frequency that appears most consistently across the voiced speech and refines the frame estimates to be observed peaks nearest this frequency.

In the present work, one adjustment from the approach in [5] was made in the spectral peak-picking step to account for residual gradient noise in the data [21]. At this step, before removing spectral tilt, half of the mean power spectrum was removed (half-scaling of the power represents a compromise between no-noise and extreme-noise reduction). The noise power spectrum was estimated per utterance as the mean spectral envelope of silence frames. Mirroring the method for vowel frame selection, silence frames were considered those that were labeled as silence by the automatic alignment and that also had a low probability (<0.2) of voicing [24].

An example of the spectral analyses run on the selected vowel frames is shown in Fig. 1. On the left are the mean spectral envelopes (full resolution, closed and open phase). The estimated cyclic peak is indicated with a black star. On the right is the mean difference between spectral envelopes estimated over the closed and open glottal phases, respectively. This closed-open phase spectral difference clearly displays a concentration of energy between 3-4kHz, reflecting the epilarynx cyclicity property. Fig. 2 illustrates the corresponding spectrograms (full resolution and closed-open difference) for the selected frames. The refined cyclic peak estimate for each frame is indicated with a black star and the overall estimate for the speaker is shown with a black line. The spectral envelopes on the left of Fig. 2 are from different vowels, as can be seen most clearly in the variation of lower formants (e.g. energy between 500-2500Hz). The corresponding closed-open difference on the right of Fig. 2 shows cyclicity consistently observable across the different vowels, as expected from [6, 14, 5].

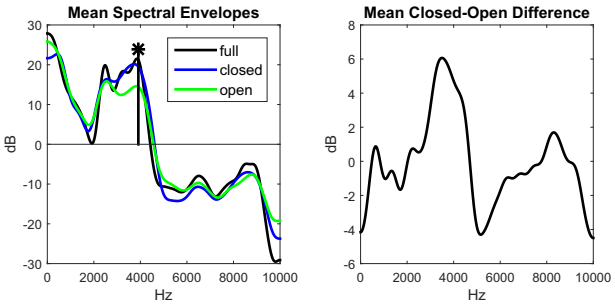

Figure 1: *Mean spectral envelopes (left) and closed-open spectral difference (right) for an utterance. The overall cyclic peak estimate is shown with a black star (*).*

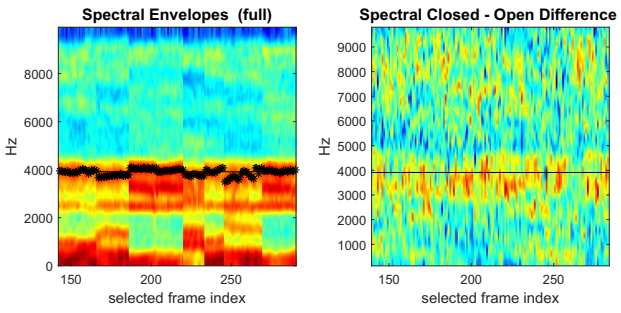

Figure 2: *Full-resolution (left) and closed-open difference (right) spectrograms. The cyclic peak estimates are shown in black (overall-line, frame-star*).*

Fig. 3 plots the estimated overall cyclic peak frequency for the speakers along with a breakdown of mean cyclic frequency for the individual vowels. As can be seen in Fig. 2-3, for the same speaker, there is little variation in the cyclic peak frequency between vowels. Comparing across speakers in Fig. 3, there is a clear trend towards higher cyclic peak frequency estimates for the female speakers (1-9) versus the male speakers (10-15), consistent with acoustics of a shorter VT and subsequently shorter epilarynx.

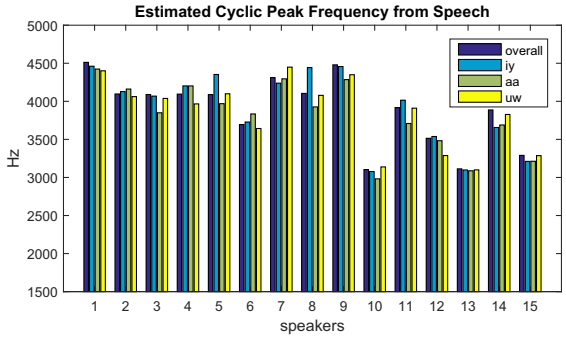

Figure 3: *Cyclic peak frequency, estimated from speech, across speakers with a breakdown by vowel.*

3. Epilarynx Length Measurement

3.1. Vocal Tract Image Data from High-Resolution 3D MRI

Unfortunately, image resolution of the rtMRI data described in Sec. 2.1 is too low to provide detailed estimates of the epilarynx length [19]. In particular, the achieved resolution of 2.9mm for a single voxel [19] represents 15% of the 20mm average epilarynx length reported in [1]. Fortunately, however, higher-resolution 3D MRI data was collected by USC SAIL for the same speakers. Specifically, static 3D MRI were acquired for sustained vowels lasting 8 seconds using an undersampled 3DFT gradient echo pulse sequence described in [27], yielding volumes with 1.25 mm^3 isotropic spatial resolution over a FOV of 200 x 200 x 100 mm^3. Additionally, the 3D MRI data were interpolated linearly by a factor of three in the transverse and coronal directions and by a factor of two along the sagittal planes, similarly to the processing in [4]. Consequently, the image pixel size on the sagittal plane corresponds to approximately 0.4 mm (1.25/3 mm). Finally, the stimuli used for epilarynx length annotations were speakers uttering the American English words "bought," "boot," "beet," corresponding to the respective spoken vowels /aa/, /uw/, and /iy/ analyzed in Sec. 2.

3.2. Epilarynx Length Measurement

As made evident in [4], measurement of the dimensions of lower VT cavities rely on decisions about how to define limits and orientation of the epilarynx. The method adopted in this work for measuring epilarynx length is as follows. First, the left plot of Fig. 4 shows an example of the epilarynx Region Of Interest (ROI) with labeled regions of the airway (dark/black) and tissue (light/white) indicating the bottom of the pharynx (hypopharynx), top of the trachea and glottis, epilarynx and arytenoid cartilages. As indicated in Fig. 4, the epilarynx cavity (i.e. laryngeal ventricle and vestibule) is the airway between the hypopharynx and the glottis, formed around the arytenoid cartilages [1].

For measurements, the length of the epilarynx is the distance between two points respectively annotating the cavity *exit* and *entrance*. First considering the epilarynx exit [4, 13], the epilarynx/hypopharynx limit is approximated here as the uppermost visible point of the arytenoid cartiledges on the midsagittal view, corresponding to the transverse view shown on the right of Fig. 4. Specifically, the transverse view shows the limit of the formation of the epilarynx cavity, visible by the arytenoid cartilages closing in to the left and right of the cavity, beginning to separate it from the piriform fossa [15]. The *exit point* is at this uppermost limit near the central part of the airway to the left of the arytenoid cartilages, indicated by a red star in both plots of Fig. 4. Next, the lower limit of the epilarynx cavity used for the length measurement is the *entrance point* at the lower, anterior limit of the visible airway at the glottis, shown with the pink star on the left of Fig. 4. The glottis is identified by hyperintense airway pixels near the bottom of the arytenoid, which is unfortunately difficult to view clearly with the contrast in Fig. 4 and 5. Finally, the measured epilarynx length is the distance between labeled cavity entrance and exit points. Note that, as observed in [4, 15], the precise limits of the epilarynx are not concretely defined. However, a concerted effort was made in this work to provide consistent annotations across the speakers and vowels.

Fig. 5 shows the manually labeled epilarynx entrance and exit points, connected by a line illustrating the measured length, for the vowels of a male speaker. The (40x40mm) ROIs are each centered in the coronal plane of the epilarynx exit point, half the distance to the glottis. As can be seen in Fig. 5 for each of the speakers, the epilarynx shape, length and orientation remains similar across the vowels, as is consistent with [15].

Unlike all previous studies on the epilarynx that consider the cavity as a vertical structure, the length measurements in this work account for the observed cavity orientation. Specifically, the epilarynx orientation angles (between the measured line and the horizontal axis) ranged from 61°to 77°across speakers (mean of 72°$\pm$ 5°), where 90°corresponds to vertical. Across vowels, the standard deviation of these measured angles for each speaker was under 6°for all speakers, confirming that the epilarynx orientations for each speaker are consistent across vowels.

Finally, Fig. 6 plots the measured epilarynx cavity length for the speakers, where the overall is the mean length over the vowels. As indicated by Fig. 5, there is little variation across vowels for each speaker. Specifically, the standard deviation of length measurements across vowels for each speaker ranged from 0.16 mm to 1.23 mm, with a mean of 0.67 mm. Additionally, the measured cavity lengths for the female speakers (1-9) are shorter than those for the males (10-15). This trend is opposite of that observed for the estimated cyclic peak frequencies in Fig. 3, as is expected and consistent with an inverse relationship between resonant frequency and epilarynx cavity length.

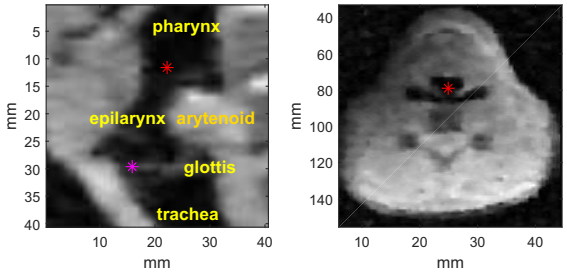

Figure 4: *Epilarynx ROI (left) with labeled exit (red*), entrance (pink*) points. Transverse view (right) at exit point (red*).*

4. Estimated vs Measured Cavity Lengths

Considering the cyclic peak frequency as an acoustic proxy to the epilarynx resonance, a quarter wavelength relationship (eq 1) is used to map the cyclic peak frequencies estimated from speech to an estimated cavity length. This estimated cavity length from the speech acoustics is plotted against the measured epilarynx length in Fig. 7 for all of the speakers, considering the overall values (including all vowels). A linear fit of the data was performed (shown by the red line), yielding a slope of 1.1 with

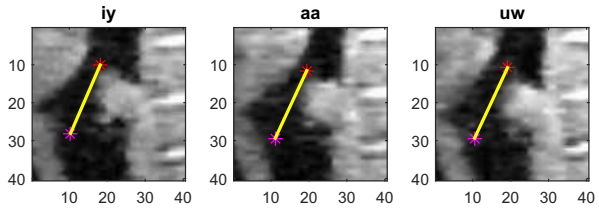

Figure 5: *Epilarynx ROI with labeled length measurements across vowels for a male speaker.*

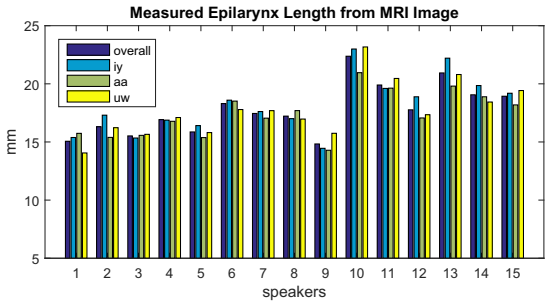

Figure 6: *Epilarynx length, measured from 3D MRI, across speakers with a breakdown by vowel.*

an offset of 3 mm. The (Pearson) correlation between the length estimate from the cyclic peak in speech and the measured epilarynx length is $\rho = 0.85$ with a $pval = 6.5x10^{-5}$. This linear trend and correlation was also observed consistently across the individual vowels, reflecting the low variation across vowels observed in both the acoustic and image analyses.

It should be noted that the mean cavity length across the speakers estimated from the acoustics is 22.8 mm whereas the mean epilarynx length measured from the 3D MRI images is 17.8 mm, yielding an overall difference of 5 mm. This overestimation was observed for all vowels and speakers: potential causes are discussed in Sec. 4.1. Ultimately, the results in Fig. 7 and the corresponding statistics indicate that the acoustic-estimated and image-measured lengths are directly proportional (linear fit slope near one) and highly correlated.

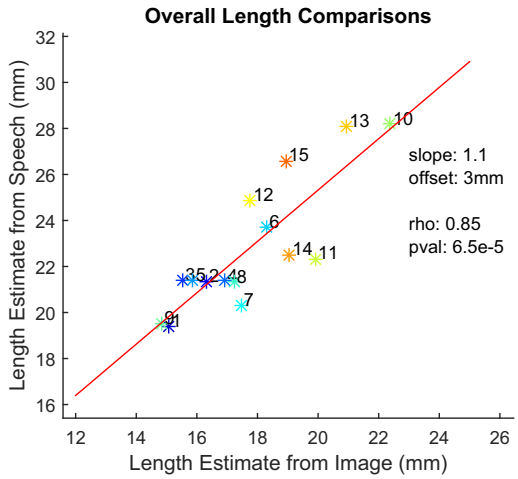

Figure 7: *Plot of estimated cavity length from the cyclic peak frequency versus measured epilarynx length for each speaker (colored, numbered stars *). A linear fit is shown in red.*

4.1. Discussion

Given its grounding in studies on epilarynx cavity behavior (established via simulation [1, 2]), up to this point, the cyclic peak has been considered an observable acoustic proxy to the epilarynx resonance. For ease of interpreting results, the remaining discussion assumes that the cyclic peak frequency is definitively the epilarynx resonance.

While the linearly proportional relationship in Fig. 7 is clear, the absolute lengths estimated from the speech acoustics are greater than those measured from the images. There are several potential reasons for this overestimation. First, the epilarynx exit is a smooth. Consequently, the effective length of the epilarynx could be longer than that measured in the MRI [3], either because the measurement is biased (towards visible landmarks in the MRI) or because the smooth transition results in an effective extension of the epilarynx cavity (e.g. an addition of a shorter, slightly larger area tube at the epilarynx exit). Similarly, from an impedance standpoint, conditions at the epilarynx exit are not ideal (i.e. free space), as the cavity terminus opens into the remainder of the VT. Consequently, interactions between the epilarynx and pharynx could impact observable acoustics[1]. Thirdly, the laryngeal cavity morphologies (including widths, lengths, cross-sectional area, curvature) undoubtedly impact the corresponding resonance patterns [4], though it is unclear precisely how and to what degree. In essence, while the quarter wavelength resonance model might capture the basic behavior of the epilarynx, deviations from this simple relationship could vary with morphology.

In interpreting results, there are also sources of error to mention due to the nature of MRI image measurements seeking to define simple parameters from structures with complex shapes. Moreover, the labeling is inherently limited by the voxel resolution of 1.25 mm achieved at the MRI data acquisition. Consequently, detailed analyses of differences between observed acoustics and measured lengths quickly encroaches on this limit. However, the specific approaches used in this work have been outlined and acoustic and image analyses were performed independently in order to avoid biasing results comparing both domains.

5. Conclusions and Future Work

Analyses in this work illustrated a clear linear relationship between the cavity length estimated from the cyclic spectral peak frequency in speech (using a simple quarter wavelength resonance relationship) to the MRI-based measurement of epilarynx length across speakers. Trends observed in both the acoustic and image domains were consistent across vowels. Consequently, results represent an important step direct linking specific, observable acoustics to VT cavity structures seen via MRI.

Future work will include examination of morphological variations of the epilarynx, including cavity widths, and corresponding links to observable acoustics, helping to further bridge speech science modeling and application domains. Extending from the epilarynx, the estimated cyclic peak frequency could also be correlated with related speaker characteristics such as VT length and height [1]. For practical applications in recognition and synthesis, given its link to epilarynx length, the cyclic peak could also be used for VTL Normalization (VTLN) [28] or frequency warping in voice conversion [29].

6. Acknowledgements

Thanks to Professor Shrikanth Narayanan and his group at USC-SAIL for their work collecting the MRI data.

7. References

[1] I. Titze and B. Story, "Acoustic interactions of the voice source with the lower vocal tract," *J. Acoust. Soc. Am.*, vol. 101, no. 4, pp. 2234–2243, 1996.

[2] K. Honda, T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, S. Takano, Y. Nota, H. Hirata, Y. Shimada, I. Fujimoto, S. Masaki, S. Fujita, and J. Dang, "Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling," *Comp. Methods in Biomech. & Biomed. Engineering*, vol. 13, no. 4, pp. 443–453, 2010.

[3] J. Sundberg, "Articulatory interpretation of the "singing formant"," *J. Acoust. Soc. Am.*, vol. 55, no. 4, pp. 838–844, 1974.

[4] A. Mainka, A. Poznyakovskiy, I. Platzek, M. Fleischer, J. Sundberg, and D. Murbe, "Lower vocal tract morphologic adjustments are relevant for voice timbre in singing," *PLos One*, vol. 10, no. 7, pp. 1–19, 2015.

[5] E. Godoy, N. Malyska, and T. F. Quatieri, "Estimating lower vocal tract features with closed-open phase spectral analyses," in *Interspeech*, 2015, pp. 771–775.

[6] T. Kitamura, H. Takemoto, S. Adachi, P. Mokhtari, and K. Honda, "Cyclicity of laryngeal cavity resonance due to vocal fold vibration," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2239–2249, 2006.

[7] W. T. Bartholomew, "A physical definition of good voice quality in the male voice," *J. Acoust. Soc. Am.*, no. 1, pp. 24–33, 1934.

[8] I. Titze, "Acoustic interpretation of resonant voice," *J. of Voice*, vol. 15, no. 4, pp. 519–528, 2001.

[9] J. Sundberg and C. Romedahl, "Text intelligibility and the singer's formant–a relationship?" *J. of Voice*, vol. 23, no. 5, pp. 539–545, 2009.

[10] T. Leino, A. Laukkanen, and V. Radolf, "Formation of the actor's/speaker's formant: a study applying spectrum analysis and computer modeling," *J. of Voice*, vol. 25, no. 2, pp. 150–158, 2011.

[11] T. Nawka, L. C. Anders, M. Cebulla, and D. Zurakowski, "The speaker's formant in male voices," *J. of Voice*, vol. 11, no. 4, pp. 422–428, 1997.

[12] K. Verdolini, D. Druker, P. Palmer, and H. Samawi, "Physiological study of "resonant voice"," National Center for Voice and Speech Status and Progress Report, Tech. Rep. 6, 1994.

[13] M. Guzman, A. M. Laukkanen, P. Krupa, J. Horacek, J. G. Svec, and A. Geneid, "Speaker identification based on nasal phonation," *J. Voice*, vol. 27, no. 4, pp. 523e19–523e34, 2013.

[14] H. Takemoto, S. Adachi, T. Kitamura, P. Mokhtari, and K. Honda, "Acoustic roles of the laryngeal cavity in vocal tract resonance," *J. Acoust. Soc. Am.*, vol. 120, no. 4, pp. 2228–2238, 2006.

[15] T. Kitamura, K. Honda, and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects," *Acoust. Sci. Tech.*, vol. 26, no. 41, pp. 16–26, 2005.

[16] B. H. Story, "Using imaging and modeling techniques to understand the relation between vocal tract shape to acoustic characteristics," in *Stockholm Music Acoustics Conf.* Citeseer, 2003.

[17] R. Detweiler, "An ivestigation of the laryngeal system as the resonance source of the singer's formant," *J. of Voice*, vol. 8, no. 4, pp. 303–313, 1994.

[18] B. Monson, A. Lotto, and B. Story, "Analysis of high-frequency energy in long-term average spectra of singing, speech and voiceless fricatives," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1754–1764, 2012.

[19] S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. Ghosh, A. Katsamanis, and M. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (tc)," *The Journal of the Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[20] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[21] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

[22] *The HTK Book*. Cambridge University, 2002.

[23] P. L. D. Consortium, "Csr-ii (wsj1) complete ldc94s13a," 1994.

[24] E. Godoy and Y. Stylianou, "Unsupervised acoustic analyses of normal and lombard speech, with spectral envelope transformation to improve intelligibility," in *Interspeech*, 2012, pp. 1472–1475.

[25] S. Gonzalez and M. Brookes, "Pefac - a pitch estimation algorithm robust to high levels of noise," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[26] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Digital Audio Effects (DAFx)*, 2005, pp. 30–35.

[27] Y.-C. Kim, J. Kim, M. Proctor, A. Toutios, K. Nayak, S. Lee, S. Narayanan *et al.*, "Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging," in *ISCA Workshop on Speech Production in Automatic Speech Recognition, Lyon, France*, 2013, pp. 2–5.

[28] S. Lulich, J. R. Morton, H. Arsikere, M. S. Sommers, G. K. F. Leung, and A. Alwan, "Subglottal resonances of adult male and female native speakers of american english," *J. Acoust. Soc. Am.*, vol. 132, no. 4, pp. 2592–2602, 2012.

[29] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.