Estimating Speaker Height and Subglottal Resonances Using MFCCs and GMMs

Harish Arsikere, *Student Member, IEEE*, Steven M. Lulich, *Member, IEEE*, and Abeer Alwan, *Fellow, IEEE*

***Abstract*—This letter investigates the use of MFCCs and GMMs for 1) improving the state of the art in speaker height estimation, and 2) rapid estimation of subglottal resonances (SGRs) without relying on formant and pitch tracking (unlike our previous algorithm in [1]). The proposed system comprises a set of height-dependent GMMs modeling static and dynamic MFCC features, where each GMM is associated with a height value. Furthermore, since SGRs and height are correlated, each GMM is also associated with a set of SGR values (known *a priori*). Given a speech sample, speaker height and SGRs are estimated as weighted combinations of the values corresponding to the N most-likely GMMs. We assess the importance of using dynamic MFCC features and the weighted decision rule, and demonstrate the efficacy of our approach via experiments on height estimation (using TIMIT) and SGR estimation (using the Tracheal Resonance database [15]).**

***Index Terms*—GMMs, MFCCs, rapid estimation, speaker height, subglottal resonances.**

I. Introduction

SPEAKER height is known to have strong negative correlations (on the order of -0.8) with subglottal resonances (SGRs) [1]. Therefore, this letter treats height estimation and SGR estimation as related problems and proposes a novel framework to solve them *simultaneously*. In contrast, our previous approach [1] estimates height *using* SGR estimates. We show experimentally that the proposed approach: (1) improves the state of the art in height estimation, and (2) provides accurate and rapid SGR estimates.

A. Height Estimation

Estimating the height of an unknown speaker from his/her speech sample is a challenging task. This is because traditional speech parameters such as the fundamental frequency ($F0$), formant frequencies and Mel-frequency cepstral coefficients (MFCCs) correlate only weakly with height [2]–[4]. Therefore, reliable height estimation requires sophisticated learning algorithms [5] and/or a knowledge of height-related physiological parameters that can be determined from speech [1].

Manuscript received June 29, 2013; revised November 11, 2013; accepted December 15, 2013. Date of publication December 19, 2013; date of current version December 26, 2013. This work was supported in part by National Science Foundation Grant 0905381. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ali Taylan Cemgil.

H. Arsikere and A. Alwan are with the Electrical Engineering Department, University of California, Los Angeles, CA 90095 USA (e-mail: hari.arsikere@gmail.com; alwan@ee.ucla.edu).

S. Lulich is with the Department of Speech and Hearing Sciences, Indiana University, Bloomington, IN 47405 USA (e-mail: slulich@indiana.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2013.2295397

In [5], SVM-based regression is used to estimate height from a 50-dimensional feature vector consisting mostly of the means, standard deviations, percentiles and quartiles of MFCCs, $F0$ and voicing probability. On the other hand, the algorithm in [1] estimates height (h_{est}) using Eq. (1):

$$h_{est} = \alpha_K \times SgK_{est} + \beta_K, \tag{1}$$

where SgK_{est} is the Kth SGR ($K = 1, 2$ or 3) estimated from speech, and α_K and β_K are model parameters that are determined from 'ground truth' measurements of SGRs (obtained from accelerometer recordings of subglottal acoustics) and height. The algorithms in [5] and [1] are known to be equally accurate on the TIMIT database (yielding an RMS error of 6.8 cm), but the SGR-based method is more efficient because it requires only one feature.

Despite its acceptable error performance, the SGR-based method, as observed in [1], yields a poor within-gender correlation (~ 0.2) between actual height and estimated height. This is partly because the method relies on speech-based estimates of SGRs that are often prone to error (see Section 5.1 of [1] for a detailed argument). The first goal of this study is to improve the state of the art in height estimation, especially with regard to the within-gender correlation between actual height and estimated height.

The approach we propose here is inspired by the work of Pellom and Hansen [6], whose algorithm relies on a set of height-dependent Gaussian mixture models (GMMs) modeling MFCC distributions. This study re-evaluates (for comparison purposes) the approach of [6] because it uses the entire TIMIT database for training as well as evaluation. Despite the fact that MFCCs correlate only moderately with speaker height [4], we demonstrate their effectiveness for height estimation via a careful selection of features and decision rules.

In [6], the TIMIT database (containing 630 speakers with known heights) was partitioned into 11 height groups and each group was modeled using a 128-component GMM. The GMMs were trained using 19 MFCCs (c_1 to c_{19}) computed every 10 ms during voiced speech activity. Each GMM was assigned an average height value based on the data used for training it, and the estimated height for a given speech sample was taken to be the value associated with the most-likely GMM. The maximum-likelihood (ML) decision rule of [6] is probably not optimal because MFCCs correlate only moderately with height. We show that better performance can be achieved by incorporating more than one GMM in the decision-making process. In addition, we show that it is important to use both static and dynamic MFCC features (note that [6] uses only static features).

B. SGR Estimation

SGRs are useful not only for estimating height, but also for speaker normalization in automatic speech recognition (ASR)

[7]. Our previous SGR estimation algorithm (proposed in [1]) is based on certain well-established phonological relations between SGRs and formant frequencies [8]–[10], and is known to be reasonably accurate (yielding RMS errors of less than 5%) in estimating the first three SGRs. However, owing to its dependence on automatic formant and pitch tracking (which incurs delays and computational overhead), it is not well suited to real-time applications. Therefore, the second goal of this study is to design a rapid SGR estimation algorithm. We show that the proposed MFCC-GMM approach can estimate SGRs efficiently by exploiting their correlation with height.

II. The Proposed Approach

The proposed system comprises a set of GMMs $\{\lambda^{(1)}, \ldots, \lambda^{(M)}\}$ (corresponding to M different height groups) that are trained on a subset of the TIMIT database. The GMMs are associated with two sets of numbers: average 'ground truth' heights $\{h^{(1)}, \ldots, h^{(M)}\}$ obtained from the TIMIT database, and average 'ground truth' SGRs $\{SgK^{(1)}, \ldots, SgK^{(M)}\}$ ($K = 1, 2, 3$) obtained from the WashU-UCLA corpus [11]. Since SGRs are known to correlate strongly with height [1], the association of average SGRs with GMMs $\{\lambda^{(1)}, \ldots, \lambda^{(M)}\}$ is well motivated. Given a speech signal, speaker height and SGRs are estimated as follows.

1: Extract MFCCs from the given speech signal.
2: Detect voiced frames (T in number) and form a sequence of feature vectors: $\mathcal{O} = \{O_1, O_2, \ldots, O_T\}$.
3: Compute the log-likelihoods of $\mathcal{O}$ (normalized by T) with respect to GMMs $\{\lambda^{(1)}, \ldots, \lambda^{(M)}\}$:

$$\ell^{(j)} = \frac{1}{T}\sum_{t=1}^{T} \log P(O_t|\lambda^{(j)}), \quad j \in \{1, \ldots, M\}. \tag{2}$$

4: Pick GMMs corresponding to the N highest likelihoods:

$$\{i_1, i_2, \ldots, i_N\} = \underset{j\in\{1,2,\ldots,M\}}{\text{argN-highest}}\, \ell^{(j)}. \tag{3}$$

5: Estimate speaker height and SGRs using Eq. (4):

$$h_{est} = \sum_{n=1}^{N} w_n h^{(i_n)}; \quad SgK_{est} = \sum_{n=1}^{N} w_n SgK^{(i_n)}, \tag{4}$$

where $\{w_1, \ldots, w_N\}$ are weights that sum to 1.

Fig. 1 depicts the proposed MFCC-GMM system. Based on the above description, the salient features of the proposed approach can be summarized as follows:

- Speaker height and SGRs are estimated *simultaneously*.
- Unlike the ML approach of [6], height is estimated using the N most-likely GMMs; this is expected to compensate for the lack of a strong correlation between MFCCs and height. In addition, the height estimates have better resolution: $\binom{M}{N} \times N!$ possible outputs (N-highest method) versus M possible outputs (ML method).
- SGRs can be estimated rapidly without relying on formant and pitch tracking algorithms (unlike in [1]). Therefore, SGR-based speaker normalization (for ASR) [7] can be implemented efficiently using the same MFCC features for SGR estimation as well as recognition.

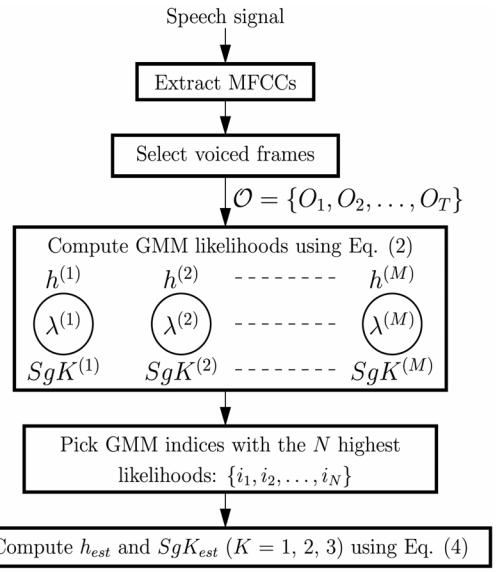

Fig. 1. The proposed system for estimating speaker height and SGRs using MFCCs and GMMs. Each GMM $\lambda^{(j)}$, $j \in \{1, 2, \ldots, M\}$, is associated with an average height, $h^{(j)}$, and a set of average SGRs, $SgK^{(j)}$, $K \in \{1, 2, 3\}$.

TABLE I
Training Parameters for the Proposed Approach: Height Range; Number of Speakers and Mixtures; and the Average Height and SGRs for Each of the 10 Height Groups (TIMIT Database)

Grp. #	Height Range (*cm*)	# Spkr.	# Mix.	Avg. Ht. (*cm*)	Avg. SGRs (*Hz*)		
					$Sg1$	$Sg2$	$Sg3$
1	[145, 155)	6	64	150.0	676	1541	2487
2	[155, 160)	10	64	157.5	665	1528	2444
3	[160, 165)	26	128	162.5	641	1498	2387
4	[165, 170)	40	128	167.5	617	1459	2368
5	[170, 175)	53	256	172.5	595	1420	2324
6	[175, 180)	62	256	177.5	563	1351	2180
7	[180, 185)	67	256	182.5	544	1319	2175
8	[185, 190)	32	128	187.5	514	1282	2166
9	[190, 195)	17	64	192.5	486	1213	1943
10	[195, 203]	4	64	197.5	518	1247	2043

III. System Implementation

This section provides the details of our implementation with regard to: (a) training data, (b) feature extraction, (c) acoustic modeling (using GMMs), and (d) decision rule.

A. Training Data

The system is trained using the TIMIT database and the WashU-UCLA corpus. Speech data from 317 (out of 630) TIMIT speakers (93 female, 224 male) are used for acoustic modeling. As shown in columns 1 and 2 of Table I, 10 height groups ($M = 10$) are created such that each of them, except groups 1 and 10, spans a height range of 5 cm. Each height group is associated with an average height value (column 5) that is used in the estimation process. It is clear from column 3 that the database has only a few speakers towards the extremes (i.e., for height < 160 cm and > 190 cm). Although this is somewhat undesirable, we use the TIMIT database to enable comparisons with the state of the art.

The WashU-UCLA corpus contains SGR measurements (obtained from accelerometer recordings) and height information for 50 adult speakers (25 female, 25 male). SGRs corresponding to speakers in each height group (of column 1) are averaged to obtain the numbers shown in columns 6–8. As expected, the

average SGRs decrease with an increase in average height; the anomalous behavior of group 10 is presumably due to its small sample size (only 4 speakers).

B. Feature Extraction

Speech signals are down sampled to 8 *kHz* and pre-emphasized with the filter: $H(\omega) = 1 - 0.97e^{-j\omega}$. MFCCs c_1 to c_{12} and their first- and second-order derivatives (Δ and $\Delta\Delta$) are computed every 10 ms during voiced speech activity, using 25 ms frames and a 26-channel Mel filterbank. To account for microphone and/or channel effects, mean and variance normalization are applied at the utterance level. Unlike [6], we use the time derivatives of MFCCs along with the static features, which, as we will show empirically (Section IV), leads to slightly better results.

To detect voiced frames, we use an unbiased estimate of the autocorrelation function $R(l)$:

$$R(l) = \frac{1}{L-l}\sum_{v=0}^{L-1-l} x(v)x(v+l) \qquad l = 0, \ldots, L-1, \quad (5)$$

where l is the lag in samples and $\{x(v)\}_{v=0}^{L-1}$ is a speech frame of length L. A frame is declared as voiced if $R(l_{p1})/R(0)$ is greater than 0.4 (a commonly-used threshold [12]), where l_{p1} is the lag corresponding to the first autocorrelation peak.

C. Acoustic Modeling

The expectation-maximization algorithm is used for training GMMs with diagonal-covariance components. Unlike in [6], the number of components in a GMM is roughly proportional to the number of speakers in its corresponding height group (columns 3 and 4 of Table I). This ensures that the modeling accuracy is similar across height groups.

As mentioned earlier, we use the Δ and $\Delta\Delta$ MFCC features in addition to the static coefficients $c_1 - c_{12}$. To see if the proposed feature set leads to better acoustic models than those trained using $c_1 - c_{19}$ (as in [6]), we measure the *separability* between GMMs corresponding to adjacent height groups.

The Kullback-Leibler (KL) divergence is a well known measure of the dissimilarity or 'distance' between two probability density functions [13]. Since closed-form expressions of the KL divergence are not available for Gaussian mixtures, approximate solutions are often used in practice. Here, we use the approximation proposed in [14]. The separability of $\lambda^{(j)}$ with respect to its adjacent GMMs $\lambda^{(j-1)}$ and $\lambda^{(j+1)}$ (denoted by $\eta^{(j)}$) is computed using Eq. (6):

$$\eta^{(j)} = (\sigma^{(j-1),(j)} + \sigma^{(j+1),(j)})/2, \quad (6)$$

where $\sigma^{(j-1),(j)}$ and $\sigma^{(j+1),(j)}$ denote, respectively, the approximate KL divergences of $\lambda^{(j-1)}$ and $\lambda^{(j+1)}$ with respect to $\lambda^{(j)}$. Fig. 2 shows the value of $\eta^{(j)}$ ($j \in \{2, 3, \ldots, 9\}$) for GMMs trained using the two feature sets under consideration: $c_1 - c_{12} + \Delta + \Delta\Delta$; $c_1 - c_{19}$. The proposed feature set clearly results in models that are more separable; it is therefore expected to yield better estimates of speaker height and SGRs.

D. Decision Rule

The most important feature of the proposed approach is the weighted decision rule of Eq. (4). The weights $\{w_1, \ldots, w_N\}$ can be chosen in several ways (based on model likelihoods, empirical adjustments, etc.). In order to incorporate the contributions of non-ML models without over emphasizing them, we use a linearly-decreasing function of n:

$$w_n = \frac{2}{N}\left(1 - \frac{n}{N+1}\right) \qquad n = 1, 2, \ldots, N. \quad (7)$$

Note that the weights defined by Eq. (7) sum to 1. Our height estimation experiments (Section IV) revealed that $N = 4$ (weights = {0.4, 0.3, 0.2, 0.1}) yields the best results on TIMIT.

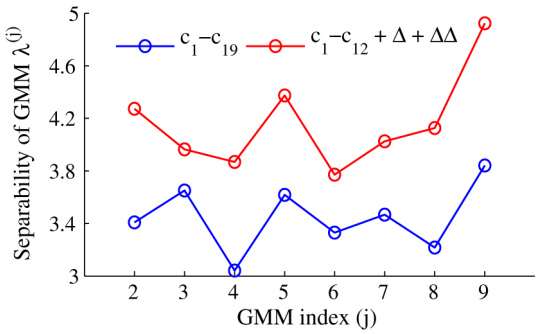

Fig. 2. Separability of $\lambda^{(j)}$ ($j \in \{2, 3, \ldots, 9\}$) with respect to $\lambda^{(j-1)}$ and $\lambda^{(j+1)}$ (Eq. (6)) for the proposed feature set and the feature set in [6], using the training subset of the TIMIT database.

IV. Experiments and Results

This section demonstrates the efficacy of the proposed approach via: (a) height estimation using the TIMIT database, and (b) SGR estimation using the Tracheal Resonance (TR) database from MIT [15].

A. Estimation of Speaker Height

Data from 313 TIMIT speakers (214 male, 99 female; different from the training set) are used for evaluation. TIMIT contains 10 utterances per speaker: 2 "shibboleth"*sa* sentences, 5 phonetically-compact *sx* sentences, and 3 phonetically-diverse *si* sentences. While GMMs are trained using all available utterances, only the *si* sentences are used for evaluation. Each *si* sentence is 2–3 *seconds* long, meaning that the height of each speaker is estimated using 6–9 *seconds* of speech.

Table II shows the RMSEs and correlation coefficients (between actual height and estimated height) for the proposed algorithm and the algorithms in [5] (using SVMs), [1] (using $Sg1$ in Eq. (1)) and [6] (using $c_1 - c_{19}$ with the ML decision rule). Note that the RMSE for the SVM-based approach [5] is taken from that paper (which does not report correlations and errors by gender). We prefer RMSE over mean absolute error because large estimation errors are captured better by RMSE. Note that the proposed algorithm with $N = 1$ (row 4) uses the ML rule like [6] but with different features; a value of $N = 4$ (row 5) is found to yield the best results.

Compared to [1], the proposed algorithm (with $N = 4$) yields statistically-significant improvements in the within-gender correlations ($p < 0.05$) and also an overall RMSE reduction of 9%. The achieved RMSE of 6.2 cm may not always be acceptable in practice, but, to our knowledge, it is currently the best result on TIMIT. Note that the overall correlations are much larger than the within-gender correlations; this is the result of large gender differences in speaker height. Male speakers benefit mostly from the dynamic MFCC features (row 4 vs. row 3), while female speakers benefit mostly from the weighted deci-

TABLE II
RMSEs (*cm*) and Correlations for Height Estimation using TIMIT (313 Speakers; 3 *si* Sentences per Speaker). The Proposed Algorithm is Compared with the Algorithms in [5], [1] and [6]

Algorithm	RMSE (*cm*)			Correlation		
	Male	Female	Overall	Male	Female	Overall
Using SVMs [5]	-	-	6.8	-	-	-
Using $Sg1$ [1]	6.8	6.6	6.8	0.17	0.25	0.70
ML method of [6]	6.8	6.7	6.8	0.25	0.32	0.72
Proposed (N = 1)	6.5	6.4	6.4	0.33	0.32	0.74
Proposed (N = 4)	6.4	5.7	6.2	0.37	0.55	0.76

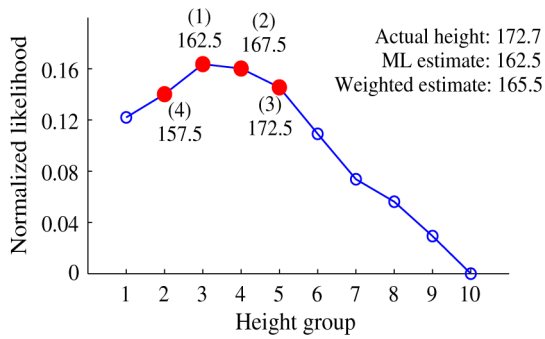

Fig. 3. Normalized likelihoods for a particular female test speaker (*vkb0* in TIMIT) showing the 4 most-likely GMMs (labeled 1–4 and highlighted in red) and the corresponding average heights. This speaker belongs to height group 5 (cf. Table I), but the most-likely estimate is group 3. By computing a weighted height estimate with $N = 4$, the ML estimate is bettered by 3 cm.

sion rule (row 5 vs. row 4). Fig. 3 illustrates with an example as to how the weighted decision rule mitigates estimation error, when the ML rule is inaccurate.

B. Estimation of SGRs

Data from 14 speakers (7 male, 7 female) in the TR database are used for this experiment. The database consists of utterances of the form "——, *say* —— *again*," and each utterance has a corresponding accelerometer recording from which the 'ground truth' SGRs can be measured (further details can be found in [15]). The TR database has been used previously to evaluate the SGR estimation algorithm in [1]. Here, the SGRs of each speaker are estimated using 6 randomly-chosen utterances (each being about 1.5 *seconds* long).

Table III shows the RMSEs and within-gender correlations (averaged over males and females) for the estimation of $Sg1$, $Sg2$ and $Sg3$. Compared to [1], the proposed algorithm (with $N = 4$) performs better with regard to $Sg2$ and $Sg3$ (the RMSE reductions and correlation improvements are statistically significant; $p < 0.05$), but worse with regard to $Sg1$. The performance drop for $Sg1$ is probably due to the fact that $Sg1$ does not correlate as strongly with height as do $Sg2$ and $Sg3$ [1]. The improvement achieved for $Sg3$ can be attributed to the fact that the proposed approach is more direct compared to that of [1], which estimates $Sg3$ via an estimate of $Sg2$. The main advantage of the proposed algorithm, however, is its ability to estimate SGRs rapidly; MFCC computation and GMM-based scoring are better suited to real-time implementation than are automatic formant and pitch tracking. Also, when SGRs are used for speaker normalization [7], the same features can be used for both SGR estimation and recognition.

TABLE III
RMSEs (*Hz*) and Within-gender Correlations (Averaged Over Males and Females) for the Estimation of $Sg1$, $Sg2$ and $Sg3$ using the TR Database (14 Speakers; 6–9 *seconds* of Speech per Estimate). The Proposed Algorithm is Compared with that of [1]

Algorithm	RMSE (*Hz*)			Correlation		
	$Sg1$	$Sg2$	$Sg3$	$Sg1$	$Sg2$	$Sg3$
Using $F0$ and formants [1]	26	60	109	0.68	0.35	0.15
Proposed (N = 4)	33	50	92	0.37	0.62	0.58

V. Conclusions

In this letter, an MFCC-GMM system is proposed for the simultaneous estimation of speaker height and SGRs. To compensate for the lack of a strong correlation between MFCCs and height, the system uses a weighted decision rule (instead of the ML rule) involving the N most-likely GMMs. With $N = 4$, the proposed algorithm improves upon the state of the art in height estimation, especially with regard to within-gender correlations. It also estimates $Sg2$ and $Sg3$ better than our previous estimation algorithm; more importantly, SGRs can be estimated rapidly without relying on formant and pitch tracking. The methods developed here can possibly be improved further by: (1) using more training data for the extreme height ranges (< 160 cm and > 190 cm), and/or (2) training GMMs using discriminative criteria (e.g., maximum mutual information).

References

[1] H. Arsikere, G. K. F. Leung, S. M. Lulich, and A. Alwan, "Automatic estimation of the first three subglottal resonances from adults' speech signals with application to speaker height estimation," *Speech Commun.*, vol. 55, pp. 51–70, 2013.
[2] H. J. Künzel, "How well does average fundamental frequency correlate with speaker height and weight?," *Phonetica*, vol. 46, pp. 117–125, 1989.
[3] J. González, "Formant frequencies and body size of speaker: A weak relationship in adult humans," *J. Phonetics*, vol. 32, pp. 277–287, 2004.
[4] S. Dusan, "Estimation of speaker's height and vocal tract length from speech signal," in *Proc. Interspeech*, 2005, pp. 1989–1992.
[5] T. Ganchev, I. Mporas, and N. Fakotakis, "Audio features selection for automatic height estimation from speech," in *Proc. Artificial Intell. Theories Models Appl.*, 2010, pp. 81–90.
[6] B. L. Pellom and J. H. L. Hansen, "Voice analysis in adverse conditions: The centennial Olympic park bombing 911 call," in *Proc. 40th Midwest Symp. Circuits Syst.*, 1997, pp. 873–876.
[7] H. Arsikere, S. M. Lulich, and A. Alwan, "Non-linear frequency warping for VTLN using subglottal resonances and the third formant frequency," in *Proc. ICASSP*, 2013, pp. 7922–7926.
[8] K. N. Stevens, "On the quantal nature of speech," *J. Phonetics*, vol. 17, pp. 3–45, 1989.
[9] S. M. Lulich, "Subglottal resonances and distinctive features," *J. Phonetics*, vol. 38, pp. 20–32, 2010.
[10] Y. Jung, "Acoustic articulatory evidence for quantal vowel categories: the features [low] and [back]," Ph.D. dissertation, Harvard-MIT Division of Health Sciences and Technology, Mass Inst. Technol., Cambridge, MA, USA, 2009.
[11] S. M. 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. Amer.*, vol. 132, pp. 2592–2602, 2012.
[12] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20, no. 5, pp. 367–377, Dec. 1972.
[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley-Interscience, 2012.
[14] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proc. ICCV*, 2003, pp. 487–493.
[15] M. Sonderegger, "Subglottal coupling and vowel space: An investigation in quantal theory," B.S. thesis, Mass. Inst. Technol., Cambridge, MA, USA, 2004.