Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models

John H. L. Hansen,[a] Keri Williams, and Hynek Bořil
Center for Robust Speech Systems, Erik Jonsson School of Engineering and Computer Science,
University of Texas at Dallas, Richardson, Texas 75083, USA

(Received 22 May 2013; revised 1 June 2015; accepted 17 July 2015; published online 21 August 2015)

Estimating speaker height can assist in voice forensic analysis and provide additional side knowledge to benefit automatic speaker identification or acoustic model selection for automatic speech recognition. In this study, a statistical approach to height estimation that incorporates acoustic models within a non-uniform height bin width Gaussian mixture model structure as well as a formant analysis approach that employs linear regression on selected phones are presented. The accuracy and trade-offs of these systems are explored by examining the consistency of the results, location, and causes of error as well a combined fusion of the two systems using data from the TIMIT corpus. Open set testing is also presented using the Multi-session Audio Research Project corpus and publicly available YouTube audio to examine the effect of channel mismatch between training and testing data and provide a realistic open domain testing scenario. The proposed algorithms achieve a highly competitive performance to previously published literature. Although the different data partitioning in the literature and this study may prevent performance comparisons in absolute terms, the mean average error of 4.89 cm for males and 4.55 cm for females provided by the proposed algorithm on TIMIT utterances containing selected phones suggest a considerable estimation error decrease compared to past efforts.
[http://dx.doi.org/10.1121/1.4927554]

[CYE] Pages: 1052–1067

I. INTRODUCTION

Recent years have witnessed an increased interest in the use of voice-based biometrics as complementary traits for person identification, access authorization, forensics, and surveillance. While the identity traits extracted from a person's voice do not reach the level of distinction as seen in fingerprints, iris patterns, or DNA, they have the potential to complement and benefit the accuracy and robustness of other biometric processes (Jain *et al.*, 2004). In comparison with other biometric domains, voice acquisition is less intrusive and its implementation and deployment are easier and less costly (Mporas and Ganchev, 2009). Moreover, in some application domains (e.g., emergency calls), voice may represent the only accessible biometric source.

Current automatic voice-based speaker identification systems, while very useful, are limited in the information they provide. A typical speaker identification system can identify an individual from a specific set of speakers (Reynolds, 1995) or claim the individual does not belong to the group of target speakers and is out of set (Angkititrakul and Hansen, 2007), or verify whether the speaker is the claimed individual or an impostor (Kinnunen and Li, 2010; Hasan *et al.*, 2013). These systems require access to a sufficient amount of training samples from each in-set speaker during the design stage, and when exposed to an out-of-set speaker during identification, besides rejecting the individual from the in-set, they do not generate any additional cues that could be used in the follow-up identification efforts.

By extracting supplementary physical traits from the speaker's voice, if the subject is an out-of-set speaker, or if insufficient training data are available, useful information about the individual can still be determined for further forensic purposes or simply for general analysis (Pellom and Hansen, 1997) (i.e., determining the gender balance of all users of a system; estimating a height or age distribution, etc.). The physical characteristic that will be estimated in this study is height. The overall goal is to determine a speaker's height strictly from an input audio sequence with minimal error and to evaluate the strengths and weaknesses of the algorithm to determine the appropriateness for other speech and language applications.

II. BACKGROUND

A majority of studies on height estimation from voice rely on the assumed correlation between individual's height and vocal tract length (VTL), supported by the evidence from magnetic resonance imaging (MRI) (Fitch and Giedd, 1999). Among other speech production features, low frequency energy (van Dommelen and Moxness, 1995), glottal pulse rate (Smith *et al.*, 2005), subglottal resonances (Arsikere *et al.*, 2012), fundamental frequency (Lass and Brown, 1978; Künzel, 1989; van Dommelen and Moxness, 1995; Rendall *et al.*, 2005; Ganchev *et al.*, 2010a), formants (van Dommelen and Moxness, 1995; Rendall *et al.*, 2005; Greisbach, 1999), and Mel frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) (Pellom and Hansen, 1997; Dusan, 2005) were studied in the context of height.

[a]Electronic mail: John.Hansen@utdallas.edu

The acoustic theory of speech production suggests that formant center frequencies are inversely proportional to VTL (Lee and Rose, 1996). In this sense, the strong correlation of VTL with height found in Fitch and Giedd (1999) could be expected to transfer also to formant frequencies. In reality, formant frequencies are not determined solely by VTL but rather depend on the complex configuration of the vocal tract cavity. For example, the first formant F_1 is known to vary inversely with the tongue height and the second formant F_2 varies with the posterior-anterior dimension during vowel articulation (Kent and Read, 1992). On the other hand, higher formants tend to be more steady and better reflect the actual VTL, a property successfully exploited in parametric VTL normalization techniques in automatic speech recognition (Eide and Gish, 1996).

In Greisbach (1999), a linear regression-based height prediction from the first four formant frequencies revealed significantly higher correlations of the third and four formants F_3 and F_4 with height compared to F_1 and F_2 for the best height estimation-suited sustained vowels. The study also observed varying suitability of different vowels for the task. The authors hypothesized that this might be attributed to the phenomenon of free variation, i.e., linguistically tolerable variations of vowel quality across speakers and social or geographical environments. The best result from that study was a standard error of 6.83 cm for males and 6.20 cm for females. Only long sustained vowels were considered; this might not be practical in some applications but could establish a reasonable upper bound on performance.

Other studies used multiple regression techniques on a large feature vector, which included prosodic and spectral features obtained from the open SMILE toolkit (Eyben *et al.*, 2009), resulting in a mean average error (MAE) of 5.3 cm for males and 5.2 cm for females from the TIMIT database (Ganchev *et al.*, 2010a; Mporas and Ganchev, 2009). A more recent study (Arsikere *et al.*, 2012) considered using the second subglottal resonance, which was suggested as being more stable as the phoneme sequence changes, unlike formant frequencies, which are phone dependent. An MAE of 5.33 cm for males and 5.45 cm for females was achieved using the TIMIT database with a traditional regression technique. That study was further explored, whereby estimating the first through third subglottal resonances an MAE as low as 5.3 cm was achieved, which is comparable to the top performing height estimation systems (Arsikere *et al.*, 2013).

One of the first studies in the area of automatic height estimation from speech used a statistical approach based on a Gaussian mixture model (GMM) class structure with 19 static MFCCs as the feature vector (Pellom and Hansen, 1997). Using the TIMIT corpus, an accuracy of 70% was achieved within 5 cm, but it should be noted that the speaker independent height models were trained on selected sentences from all available TIMIT speakers and hence, the evaluation set contained samples from the same speakers (yet different sentences). This approach achieved text independence, which is beneficial for practical applications but, unlike regression techniques, did not produce a specific numeric height, only a height bin range. A later study (Dusan, 2005) investigated correlations between height and various acoustic features including fundamental frequency, the first five formants, and MFCC and LPC features in a phone-dependent approach. The results confirmed the good correlation of MFCC features with height as previously suggested in Pellom and Hansen (1997) and also demonstrated benefits of combining cepstral features with formant frequencies.

The approach taken in this study is to employ a regression technique that utilizes formant frequency location and line spectral pair frequency structure (LSF) as well as a secondary MFCC-GMM system with data driven dynamic height bins that includes a confidence score. These two systems are then combined to achieve the best overall accuracy. By using a regression based method leveraged with a GMM based solution, the strengths and weaknesses of these alternative schemes will help balance error in the overall height estimation system. The regression system is phoneme dependent and provides a point estimate of height, while the GMM system is phone independent and provides a height class (i.e., high interval) estimate. It is noted that the regression system requires occurrence of at least one of the phones from the selected phone set to perform height estimation. LSFs have not been used previously in height estimation, but they were chosen due to the fact that they effectively represent spectral data including formant related structure. In addition, using data driven dynamic height bins and estimating a confidence score for the GMM based solution is also novel because previous GMM based solutions (Pellom and Hansen, 1997) employed uniformly distributed height bins with non-uniform speaker training data and also did not incorporate any confidence score. A preliminary version of the Gaussian mixture model height distribution based classification (GMM-HDBC) algorithm proposed here was considered in Williams and Hansen (2013). In the current study, further algorithm development as well as a variety of experiments are performed to judge performance. The distribution of error as well as average error for each height is established for each method to determine under what circumstances each approach performs well or fails. The effect of speaker session-to-session variability with respect to performance consistency for height estimation is determined as well as the consistency of the necessary regression coefficients when the training speakers change. In addition, open set height estimation, which employs speech from 10 male and 10 female actors obtained from the audio portions on YouTube video, is presented.

III. CORPUS

Little to no formal data collection for height estimation has been undertaken in the field, so the primary data used in this study is from the TIMIT (National Institute of Standards and Technology, 1988) database. TIMIT was chosen because it includes height information for each speaker, and it has also been used in previous studies (Pellom and Hansen, 1997; Ganchev *et al.*, 2010a; Mporas and Ganchev, 2009; Arsikere *et al.*, 2012, 2013; Dusan, 2005; Williams and Hansen, 2013). The height distribution of the TIMIT speakers closely resembles that of the U.S. population height distribution, which would allow for more effective testing.[1] The heights available from the TIMIT corpus, however, are self-reported; this could potentially introduce error. However, studies have shown that while people tend to overestimate their height, a majority of the people only overestimate by a small amount (Perry *et al.,* 2011). Therefore the error introduced by any potential self-reported bias is expected to be minimal. Moreover, because Institutional Review Board protocol was followed in the collection of TIMIT where specific individuals cannot be identified with speaker labels, there would be less incentive to overestimate or underestimate their height.

IV. MODIFIED FORMANT/LINE SPECTRAL FREQUENCY TRACK REGRESSION (MFLTR)

A. Formant and LSF estimation

As is shown Fig. 2, the MFLTR height estimation approach employs both LSF based height estimation as well as direct formant location estimation. It is a well known speech processing challenge that when estimating formant locations from LPC analysis, there is an inherent uncertainty in ordering the resulting all-pole model (because the all-pole speech model lies in the z-plane, the poles are not naturally ordered in two-dimensional space). As such, for very strong sharp formants, a double pole is possible so direct formant location estimation from LPC analysis needs to know when to assume a single pole pair is a true representation of a formant location or if a double pole pair represents the formant or if a pole pair is actually contributing to an overall shaping effect of the vocal tract response.

Many speech processing engineers/scientists address this issue by reducing the number of pole pairs in the LPC analysis with the hope of having the resulting LPC model focus more on the formant peaks and less on the overall vocal tract response/shape (i.e., for 8 kHz sample speech, a tenth order LPC analysis is reduced to eighth order in the hopes of having four pole-pairs to represent the four formants over a 4 kHz frequency range). In the MFLTR solution, we do perform LPC analysis to directly estimate the formants and height information (subsequent removal of "extreme values" was done partly because of this ordering issue). In addition to the direct formant estimation approach via LPC analysis, we also performed LSF—line spectral pair analysis. The convenient property of LSFs is that the two-dimensional set of LPC poles are projected onto the unit circle, so the resulting LSF position parameters can be more associated with formant locations. Basically, LSFs are associated with LSF position and difference parameters (i.e., the even and odd LSF frequencies become the position and difference parameters; one is the position, the other represents the movement from the LSF position, making it a "delta" shift in frequency). Small values of the LSF difference parameter suggest the LSF pair is associated with a sharp resonant peak corresponding to a formant location, while large difference parameter values associate the LSF pair more with an overall shaping pole pair. LSF also have better interpolation properties and more reliable in characterizing formant tracks. Figure 1 shows an example of LSF position (solid lines) and difference parameters (these have been thresholded in this example where if the difference parameter is within a delta of the position parameter, they are marked as "+" or "−" depending on if they are closer to the upper or lower LSF position parameter track. It is easy to see those tracks associated with formant tracks when the +/− difference parameters are present. The ease by which LSFs can be analyzed along a one-dimensional space (i.e., unit circle) make them ideal for automatic analysis of formant structure.

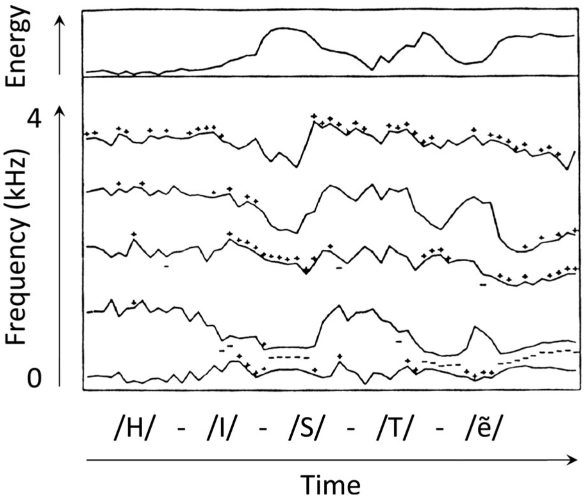


FIG. 1. Example of LSF position (Solid Lines) and difference parameters (Hansen, 1988).

To estimate formant frequencies, the poles from the all-pole speech model need to be determined. To accomplish this, the number of coefficients is determined as

$$n = \frac{f_s}{1000} + 2, \tag{1}$$

where n is the number of coefficients and f_s is the sampling frequency. The LPC coefficients can then be calculated and the roots of the all-pole model found. Once the roots of the all-pole model have been determined, the formant frequencies are estimated by

$$\hat{F}_i = \frac{f_s}{2\pi} \arctan\left(\frac{\mathrm{Im}(r)}{\mathrm{Re}(r)}\right), \tag{2}$$

where r refers to the roots of the all-pole model. With the formant frequencies estimated, the LSFs can then be calculated. LSFs are a robust way of representing the all-pole speech model (Itakura, 1975)

$$H(z) = \frac{1}{A(z)} = \frac{G}{1 - \sum_{k=0}^{n-1} a_k z^{-k}}. \tag{3}$$

The polynomial $A(z)$ is then split into two different polynomials as follows,

$$P(z) = A(z) - z^{n+1} A(z^{-1}), \tag{4}$$

$$Q(z) = A(z) + z^{n+1} A(z^{-1}). \tag{5}$$

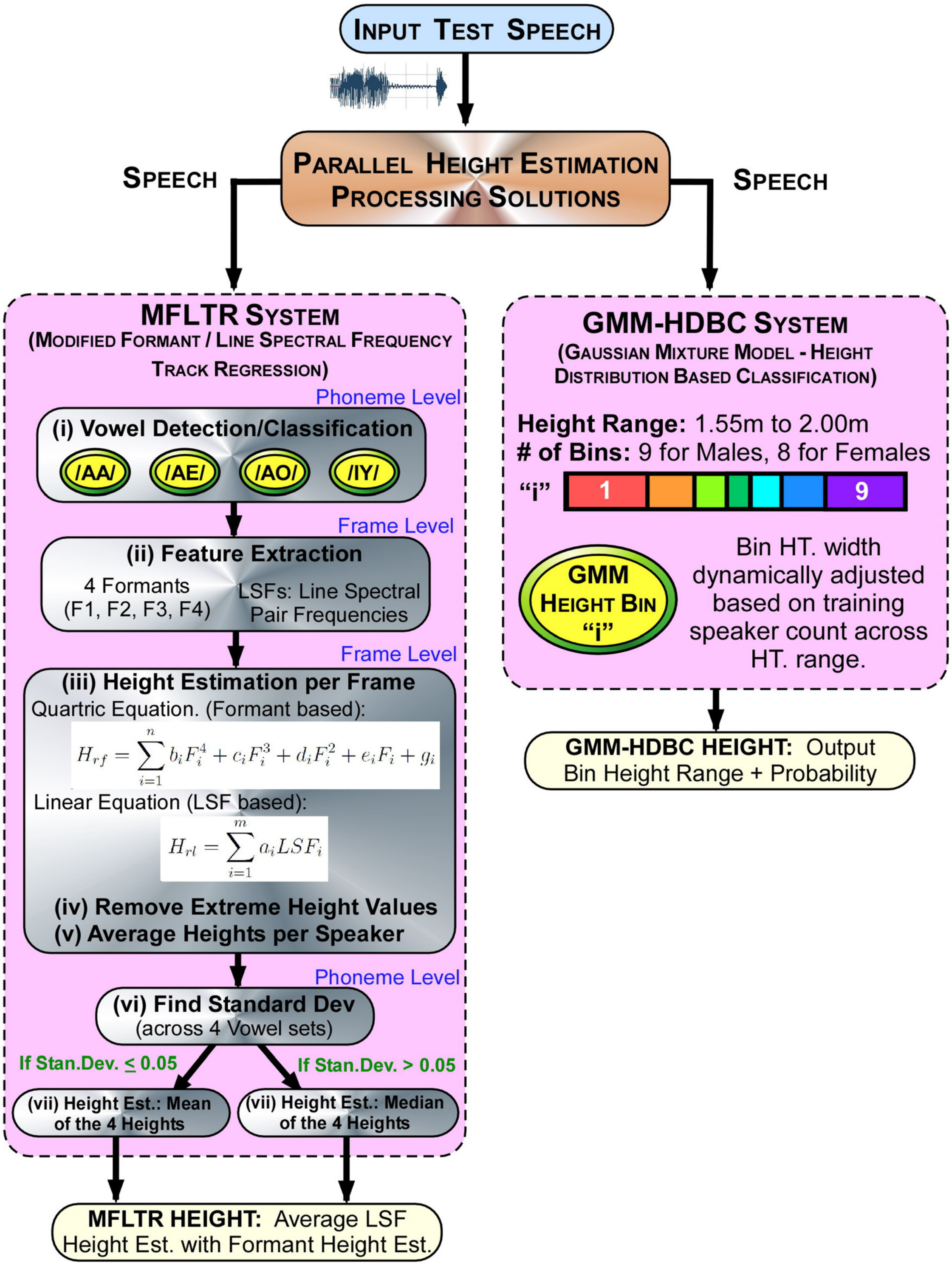

INPUT TEST SPEECH
PARALLEL HEIGHT ESTIMATION PROCESSING SOLUTIONS
SPEECH
SPEECH
MFLTR SYSTEM
(MODIFIED FORMANT / LINE SPECTRAL FREQUENCY TRACK REGRESSION)
Phoneme Level
(i) Vowel Detection/Classification
/AA/
/AE/
/AO/
/IY/
Frame Level
(ii) Feature Extraction
4 Formants (F1, F2, F3, F4)
LSFs: Line Spectral Pair Frequencies
Frame Level
(iii) Height Estimation per Frame
Quartric Equation. (Formant based):
$H_{rf} = \sum_{i=1}^{n} b_i F_i^4 + c_i F_i^3 + d_i F_i^2 + e_i F_i + g_i$
Linear Equation (LSF based):
$H_{rl} = \sum_{i=1}^{m} a_i LSF_i$
(iv) Remove Extreme Height Values
(v) Average Heights per Speaker
Phoneme Level
(vi) Find Standard Dev (across 4 Vowel sets)
If Stan.Dev. ≤ 0.05
If Stan.Dev. > 0.05
(vii) Height Est.: Mean of the 4 Heights
(vii) Height Est.: Median of the 4 Heights
MFLTR HEIGHT: Average LSF Height Est. with Formant Height Est.
GMM-HDBC SYSTEM
(GAUSSIAN MIXTURE MODEL - HEIGHT DISTRIBUTION BASED CLASSIFICATION)
Height Range: 1.55m to 2.00m
# of Bins: 9 for Males, 8 for Females
"i"
1
9
GMM HEIGHT BIN "i"
Bin HT. width dynamically adjusted based on training speaker count across HT. range.
GMM-HDBC HEIGHT: Output Bin Height Range + Probability

FIG. 2. (Color online) Flow diagram of modified formant track/LSF regression algorithm (MFLTR, left) and GMM height distribution based classification (GMM-HDBC, right) method.

From Eqs. (4) and (5), n is found by Eq. (1). The all-pole model can then be rewritten as

$$H(z) = \frac{2}{P(z) + Q(z)}. \tag{6}$$

Next, the zeros of $P(z)$ and $Q(z)$ are found, which results in the individual LSFs which effectively project the roots of $A(z)$ onto the unit circle in the z-plane. When examining the LSFs, the formant information can be inferred by two closely paired LSFs (Itakura, 1975; Crosmer and Barnwell, 1985). Once the raw formant locations and LSFs have been calculated, they can be smoothed over time to reduce estimation errors that are known to occur. The first step in smoothing the LSF tracks and formant tracks is to represent the raw data as a cubic function (Greenwood and Goodyear, 1994). The cubic track is then sorted from lowest to highest value. Once the sorting is complete, half of the frames will be eliminated by removing the upper 25% of the frames with the highest values as well as removing the lower 25% with the smallest values for each track. The final tracks are expected to be smooth and fairly constant with estimation errors being minimized.

B. Algorithm—FLTR

Having estimated the formant and LSF structure, the next step is to estimate height. The MFLTR for height estimation is based on solving an equation that represents the height of a speaker in terms of the first four formants along with 18 LSFs and then performing a post-processing cleanup phase for the height estimates (see the left hand side of Fig. 2).

Past literature on non-linguistic speech feature extraction (e.g., gender or speaker information) often utilizes phonetic segmentation (Lamel and luc Gauvain, 1995) to address signal variability during phonation of different speech sounds. Studies focused on height estimation from speech typically favor the use of vowel segments for their relative stationarity and regular structure compared to other phone classes (Greisbach, 1999; Dusan, 2005). In Dusan (2005), a set of vowels well represented in the TIMIT recordings were utilized in height estimation. Corresponding utterance segments were parameterized by formant center frequencies and MFCC and LPC features. The study provides a rank-ordered list of vowels with respect to the correlation of their parameterized segments with height.

Following Greisbach (1999) and Dusan (2005), the MFLTR algorithm proposed in this section focuses on vowel segments. Four vowels /aa/, /ae/, /ao/, and /iy/ were chosen due to the quantity of TIMIT speakers that produced them as well as their separation in the F_1–F_2 space. In addition, these four vowels ranked among the top ten (/iy/ being the rank number one) in Dusan (2005). Vowel formants are generally consistent over time because there is little vocal tract articulatory movement during their production. As a result, the features extracted from the vowel segments can be expected to be more stable compared to other phone classes. At the same time, due to the distinct articulatory configurations across the four vowels, it is beneficial to perform a separate modeling for each vowel class.

Time boundaries of the selected vowels in TIMIT utterances can be estimated with high accuracy using for example a freely accessible BUT phone recognizer (Schwarz, 2009). Because the information about phone boundaries is already available in the TIMIT transcription files and Lamel and luc Gauvain (1995) have demonstrated that their usage does not provide any advantage over automatically extracted phone boundaries in terms of paralinguistic information extraction, we follow the approach in Dusan (2005) and directly utilize the available labels rather than re-extracting them through an external phone recognizer. In general, the accuracy of the automatic phone alignment will be affected by the presence of environmental noise and channel variation as well as speaker variability. In the cases of recordings acquired under adverse conditions that might break the phone recognition, the MFLTR method can be utilized in a semi-supervised fashion where the phone boundaries would be perceptually labeled by a human expert.

After the smoothed formant location tracks and LSF tracks have been calculated in the vowels segments, they are incorporated into separate equations that relate height on a frame by frame basis. In Greisbach (1999) and also our preliminary study (Williams and Hansen, 2013), a subject's height was estimated through a linear regression function of the first four formants. It may be reasonable to assume that a person's height and VTL exhibit linear relationship. However, as discussed in Sec. II, center frequencies of higher formants tend to be inversely proportional to VTL, and lower formants are rather driven by the articulation of individual speech sounds. In this sense, representing height simply as a linear function of the formant center frequencies may be a rather crude approximation. For this reason, we propose to extend the linear function from Greisbach (1999) and Williams and Hansen (2013) to a polynomial to better accommodate the complex interdependencies of the formant structure, phonation, and height.

The height estimation from formants is defined in Eq. (7) where the coefficients b_i, c_i, d_i, e_i, g_i are found by linear regression,

$$H_{rf} = \sum_{i=1}^{n} b_i F_i^4 + c_i F_i^3 + d_i F_i^2 + e_i F_i + g_i, \tag{7}$$

and where n is the number of formants, and F_i refers to the formant center frequencies. For LSFs, the equation is shown in Eq. (8) where the coefficients are also found by linear regression,

$$H_{rl} = \sum_{i=1}^{m} a_i LSF_i, \tag{8}$$

where m is the number of LSFs and a_i are the coefficients found by regression. Once the height is estimated at a frame level, where frames here are 20 ms in duration, the frame-level heights are sorted from smallest to highest and the bottom 25% and upper 25% height values are eliminated. This is performed to alleviate possible effects of coarticulation at the phone boundaries on phone-specific height estimation and, more generally, to reduce the number of outliers in the frame-level height estimates.

In the following step, the height estimates are averaged together to obtain a height estimate for each speaker for each phoneme. To combine the phoneme speaker height estimates, the heights mean or median is found depending on the standard deviation of the four height samples. For a high standard deviation, it is assumed that the phoneme-specific height estimates contain outliers that would decrease the accuracy of the mean height estimation. In this case, a median of the estimates is taken. Median represents a non-linear filter and is less sensitive to outliers (in our case, the shortest and tallest outliers) than mean. For a low standard deviation, all phoneme-specific height estimates are assumed to be meaningful and their mean is taken. Once this step is complete, there are two height estimates available for each speaker, one based on formant locations and another on LSFs. To combine these two estimates, the average is calculated, resulting in one overall height estimate for each speaker.

C. Training

A total of 268 male and 127 female sessions from the TIMIT corpus sampled at 16 kHz were utilized in the evaluations. The gender-dependent sets were split approximately in half to form training and test sets with non-overlapping speakers. The lists of training and open test set sessions are available online (CRSS, 2015). The training and test sets were specifically designed to have speakers across the entire height range to achieve effective training and independent testing. Not all data could be used due to the phoneme dependence of the MFLTR method. The four vowels were chosen because a large number of speakers produced them for the *sa1* sentence ("She had your dark suit in greasy wash water all year"). For the other nine sentences produced by each of these speakers, all were examined to see if they contained any of the four vowels of interest and if so, these sentences were included.

It is noted that the partitioning of the TIMIT data into training and test sets here differs from those in Dusan (2005), Ganchev *et al.* (2010a), and Arsikere *et al.* (2013) and also that each of these studies introduced their own unique partitioning. In this sense, the height estimation accuracies reported by these studies (including the present one) cannot be compared in strictly absolute terms. However, it can be argued that the task difficulty should be comparable as the samples are drawn from the same corpus and the height ranges captured in the sets can be expected to be similar.

V. HEIGHT DISTRIBUTION BASED CLASSIFICATION—GMM-HDBC

In this section, a second alternative height estimation scheme, GMM-HDBC, is formulated based on statistical modeling concepts.

A. Feature estimation

The feature used for this method consists of 19 static MFCC coefficients including normalized energy. MFCCs have been shown in a previous study to be effective in reflecting a speaker's height (Pellom and Hansen, 1997; Dusan, 2005). This is possible because the static MFCC coefficients tend to be related to a person's vocal tract configuration (Pellom and Hansen, 1997; Dusan, 2005). The normalized energy is included to accommodate thresholding-based silence and low energy speech segments since those are not expected to provide any useful information.

B. Algorithm

This method (see the right hand side of Fig. 2) is focused on a sentence level analysis and extracts 19 static MFCC coefficients as described in Sec. V A. From there, the features are processed into different GMMs with pre-defined mixture sizes. For the GMM structure to be effective, the speaker heights need to be grouped within height ranges. Instead of employing an equally spaced scale where heights are distributed along a uniform scale (as was performed in Pellom and Hansen, 1997), the height bins were partitioned using a data driven approach based on the amount of data available for each height (see Fig. 3).

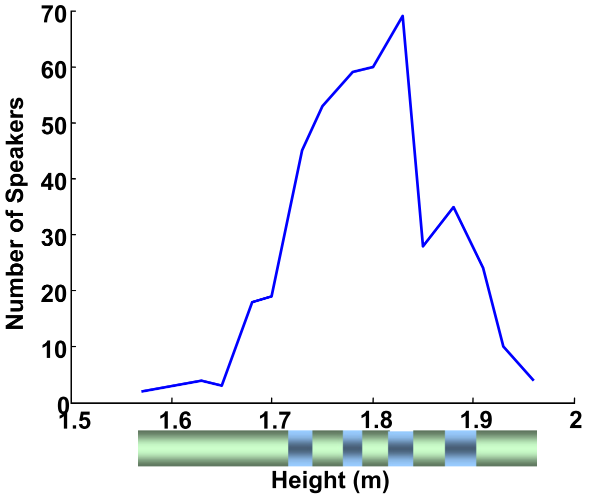


FIG. 3. (Color online) Height classes with height distribution for male speakers. Boxes below *x* axis represent height bins modeled by individual gender-specific GMMs.

In this manner, the intrinsic *a priori* probability of the height distribution of the population under train/test would be incorporated that also allows for data balancing of the height models. Because some heights have significantly more data than others, especially around the centroid of the height distribution scale, this ensures less speaker dependent characteristics within each height range GMM model. By using a linear partitioned scale, the tails of the height models do not have as much training data, so the height GMMs become more speaker dependent versus central height models that are more speaker independent. To address this problem, a minimum threshold was set for the number of speakers needed to construct each height range GMM. Using this strategy, height bin groups are formed based on the distribution of how many speakers are present for each height, and if insufficient speaker data are available, that bin group is added to the neighboring group. The minimum number of speakers for males was set to 30, and for females, it was set to 15. These thresholds consider the total number of available speakers per bin including pooled training and test speakers and were arbitrarily chosen to provide a reasonable speaker variability for speaker-independent height bin model training and evaluation. The disproportion between gender samples reflects the gender population in TIMIT.

In this way, a balanced height coverage for both training and evaluation is assured. This configuration will reduce the speaker dependency and ensure an effective height class estimate for each speaker. The nine centroids in meters for males are as follows: (1.635, 1.73, 1.75, 1.78, 1.8, 1.83, 1.85, 1.88, and 1.935), while the eight centroids for females are: (1.51, 1.6, 1.63, 1.65, 1.68, 1.7, 1.73, and 1.79). It seems useful to also include a confidence measure to communicate how likely that height class might be for the user. The confidence measure used here is the probability closeness measure (Rabiner and Schafer, 2011),

$$\text{confidence\_}s_1 = \frac{\frac{1}{p_1}}{\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3}}, \tag{9}$$

where the confidence measure for a single speaker s_1 is calculated by using the probability of the most likely class p_1, the probability of the second most likely class p_2, and the probability of the third most likely class p_3. This confidence measure will state how separable the top three height models' probabilities are; this reflects the confidence in the model choice. The greater the top height model probability is compared to the second and third models, the closer the confidence measure is to one. With this strategy, each speaker is assigned a detected height bin class as well as a corresponding confidence measure within (0, 1).

C. Training

Due to the text independent nature of the GMM-HDBC solution, all TIMIT utterances per speaker could be used. The training set for males contains 15 speakers for each height bin class and the training set for females contains 9 speakers for each height bin class. The remaining speakers are set aside for the test set. The TIMIT session partitioning follows the one from Sec. IV C and is detailed in CRSS (2015).

Acoustic models in automatic speech recognizers typically utilize 16–32 mixture GMM states to model speaker/gender independent phone distributions. Similar choices can be found in phone modeling for paralinguistic information extraction (Lamel and luc Gauvain, 1995). In Pellom and Hansen (1997), 128 mixture GMMs were employed in phone-independent height estimation. To accommodate the fact that the GMMs in the current study are phone-specific, which would suggest to use less mixtures than in Pellom and Hansen (1997), yet should provide a more detailed resolution of the non-linguistic content than speech recognition-oriented models, all height-bin GMMs here are facilitated with 64 mixtures. It is noted that this choice was found meaningful also in our preliminary study (Williams and Hansen, 2013).

VI. FUSION METHOD

Two independent algorithms have been determined for height estimation based on speech. In this section, an approach to fuse these estimates for an overall single estimate is formed.

A. Algorithm

The MFLTR algorithm produces a specific height estimate for each speaker, while the GMM-HDBC method assigns a height class along with a confidence score. This section describes a fusion strategy proposed to incorporate both method outputs into a single height estimate for each speaker (see Fig. 4). The first step in combining the two methods is to find the lower and upper height boundaries for the top two height classes and average the two lower and upper boundaries together. Next the height value from the MFLTR method is compared to the new upper and lower boundary to determine which is closer. If the MFLTR height value is within the upper and lower boundary, then the height value is calculated as

$$H_F = (1-C)H_R + CB, \tag{10}$$

where C is the confidence score and B is the closest boundary to the MFLTR height value, H_R. For greater confidence measures, more emphasis is placed on the boundary while for low confidence measures, more emphasis is placed on the regression result. If the MFLTR height value is not within the upper and lower boundaries, then the final height is calculated as

$$H_F = \frac{B+H_R}{2}. \tag{11}$$

This results in a compromised height estimate because it effectively averages the closest boundary height B with the estimated MFLTR height output H_R. With this method, there will be a single height result per speaker, H_F.

B. Training

The data set used for tuning the height fusion solution is the same as the data set used for training the MFLTR system because that method has the smallest number of training and testing speakers. So for the GMM-HDBC method only a portion of the entire test speakers are used.

VII. RESULTS

A. MFLTR

The results for the MFLTR method are displayed in Tables I and II. MAE (in cm) is the measure chosen to reflect performance of the MFLTR method because it has been used in previous studies for height estimation (Ganchev *et al.*, 2010a; Mporas and Ganchev, 2009; Arsikere *et al.*, 2012, 2013; Williams and Hansen, 2013). MAE was calculated on a per speaker basis. As mentioned in Sec. IV C, different reference studies utilized different partitioning of the TIMIT data sets; this makes the comparison of the reported MAEs somewhat difficult. However, it should be safe to assume that the samples tested in all these studies are drawn from the same population of speakers and present a comparable level of difficulty in estimating speakers height. It should also be noted that test duration will also influence overall system performance, so care should be exercised when comparing results if test durations are not equivalent.

Table I summarizes height estimation error results at the phoneme level to demonstrate the performance early in the MFLTR method after the frame level is complete (see Fig. 2). The results show consistent, close results for the

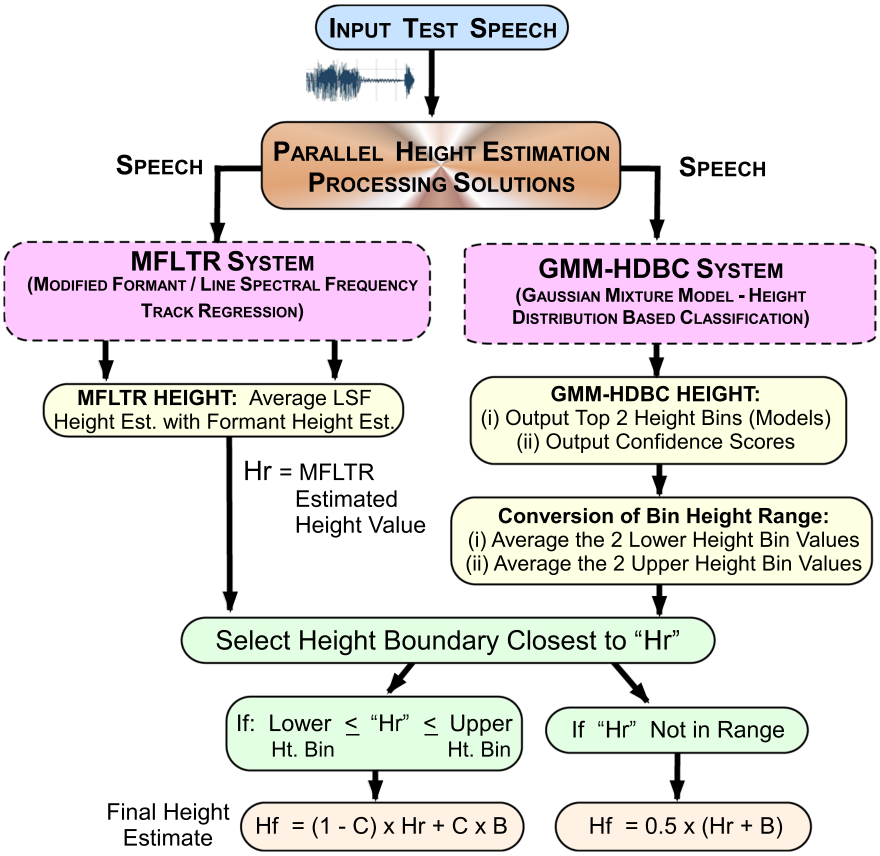

INPUT TEST SPEECH
PARALLEL HEIGHT ESTIMATION PROCESSING SOLUTIONS
SPEECH
SPEECH
MFLTR SYSTEM
(MODIFIED FORMANT / LINE SPECTRAL FREQUENCY TRACK REGRESSION)
GMM-HDBC SYSTEM
(GAUSSIAN MIXTURE MODEL - HEIGHT DISTRIBUTION BASED CLASSIFICATION)
MFLTR HEIGHT: Average LSF Height Est. with Formant Height Est.
GMM-HDBC HEIGHT: (i) Output Top 2 Height Bins (Models) (ii) Output Confidence Scores
Hr = MFLTR Estimated Height Value
Conversion of Bin Height Range: (i) Average the 2 Lower Height Bin Values (ii) Average the 2 Upper Height Bin Values
Select Height Boundary Closest to "Hr"
If: Lower Ht. Bin ≤ "Hr" ≤ Upper Ht. Bin
If "Hr" Not in Range
Final Height Estimate
Hf = (1 - C) x Hr + C x B
Hf = 0.5 x (Hr + B)

FIG. 4. (Color online) Flow chart for fusion method. Left branch represents MFLTR system and right branch GMM-HDBC system. Bottom part outlines fusion process of soft height scores from MFLTR and heights bins and confidences from GMM-HDBC.

MFLTR approach for both formant and LSF features for males and females. Performance can differ between phonemes due to coarticulation effects because these phonemes were extracted from within words in spoken sentences. Error in terms of MAE (in cm) ranges from 5.14 to 5.53 cm for males and 4.72 to 6.32 cm for females. Again note, these height estimates are based on a single vowel (∼0.25–0.5 s). In Table II, the results shown are obtained after the phoneme level in Fig. 2. Here the solution combines estimates from one to four phonemes to obtain an overall result for the LSF feature as well as the formant feature. The combination result is the final accuracy of the MFLTR method, which achieves very effective performance for males (4.93 cm) and females (4.76 cm).

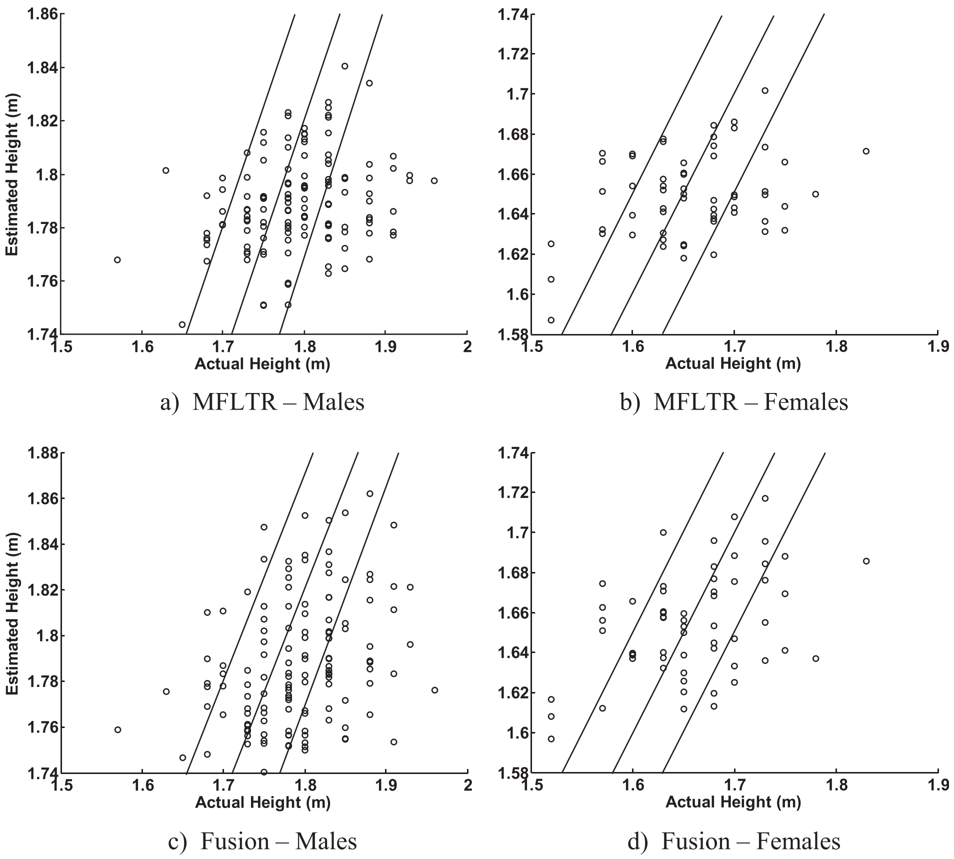
It is clear that LSF and formant based results are more accurate when combinations of four phonemes are used versus single phoneme results. This is expected because the LSF and formant based results consider multiple phonemes and if height information from one phoneme is erroneous for certain speakers, the results from the other phonemes can compensate for that. The combined result is approximately the same as the LSF result but better than the formant result for male speakers, and for female speakers, the combined result is slightly better than both the LSF and formant results individually. To demonstrate the accuracy of the MFLTR method, the estimated heights are plotted against the actual heights of all speakers in Fig. 4. The center line demonstrates perfect accuracy while the outer lines represent the error range being 5 cm. All of the points in between the lines show

TABLE I. Comparison of height estimation MAEs (in cm) in MFLTR method on the level of individual phones—LSF vs. formants.

		Male		
		LSF		
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.39	5.29	5.53	5.07
		Formants		
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.14	5.26	5.45	5.49
		Female		
		LSF		
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.55	5.61	6.32	5.08
		Formants		
Phoneme	/AA/	/AE/	/AO/	/IY/
MAE (cm)	5.06	4.72	5.44	5.35

TABLE II. Comparison of Height estimation MAEs (in cm) in MFLTR method after phone combination—LSF vs formants.

		Male	
	LSF	Formants	Combined
MAE (cm)	4.92	5.06	4.93
		Female	
	LSF	Formants	Combined
MAE (cm)	5.23	4.80	4.76



a) MFLTR – Males

b) MFLTR – Females

c) Fusion – Males

d) Fusion – Females

FIG. 5. Estimated height vs actual height for MFLTR method and fusion method with accuracy line and ±5 cm error line. Note distinctive gender-specific characteristics of height distributions.

speakers with height estimates that have less than 5 cm of error.

These plots suggest that the majority of the speakers have error less than 5 cm (59.2% of the male speakers, 55.9% of the female speakers). The correlation coefficient for males was determined to be 0.26, while for females the correlation coefficient is 0.34. These coefficients are not considerably high, but when considering both males and females together, the correlation coefficient is 0.72, which demonstrates a better relationship between estimated and actual height. The poor gender dependent performance can be explained by considering the MAE and Fig. 5. With a large number of speakers' errors ranging from ±5 cm from the actual height, this causes the ideal linear relationship to become more spread out. When males and females are grouped together, the range of heights is increased; this counteracts the spreading effect and improves the overall correlation coefficient.

B. GMM-HDBC

The results for the statistical GMM-HDBC height estimation method are determined by considering accuracy within a 5 cm range. For each confidence measure, only those speakers with at least that number are considered. As a result, a reduced number of speakers are included in the results as the confidence measure increases. The results are summarized in Fig. 6. The vertical lines represent 25% speaker elimination and 50% speaker elimination, respectively.

Here the male speaker results demonstrate a steady increase in accuracy as the confidence measure increases and achieve perfect accuracy for the highest confidence score, which is an ideal situation. The female speaker results remain fairly consistent for most values of the confidence measure and only increase slightly toward the higher end. Higher accuracy results when considering the top two models as displayed in Fig. 6. The accuracy results are displayed in such a way so as to compare the accuracy of the top model and the top two models. The female results do not increase

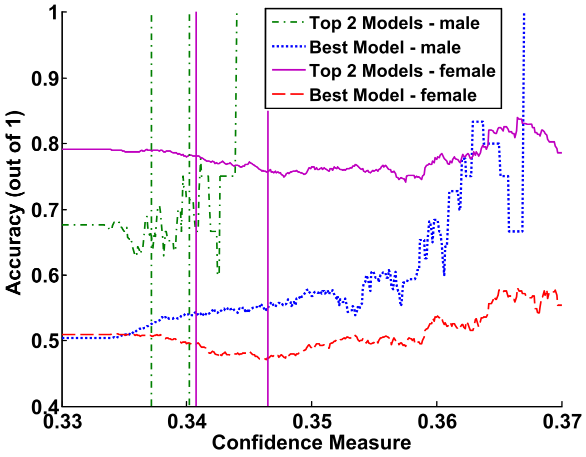


FIG. 6. (Color online) GMM-HDBC results with top model and top 2 models within 5 cm classification accuracy. Trends demonstrate accuracy of height bin assignment with respect to confidence score. Each confidence measure point represents accuracy across all samples that achieved equal or higher confidence score.

TABLE III. Height estimation MAEs (in cm) after fusion of MFLTR and GMM-HDBC methods.

		Fusion
MAE (cm)	Male	4.89
	Female	4.55

as much when the confidence measure increases as compared to the male speakers for the top model and top two models. However, accuracy does experience a significant rise when considering the top two models as opposed to only the single model (50.37% to 67.69% accuracy for males, 50.9% to 79.08% accuracy for females) and approaches the maximum accuracy sooner.

C. Fusion of MFLTR and GMM-HDBC

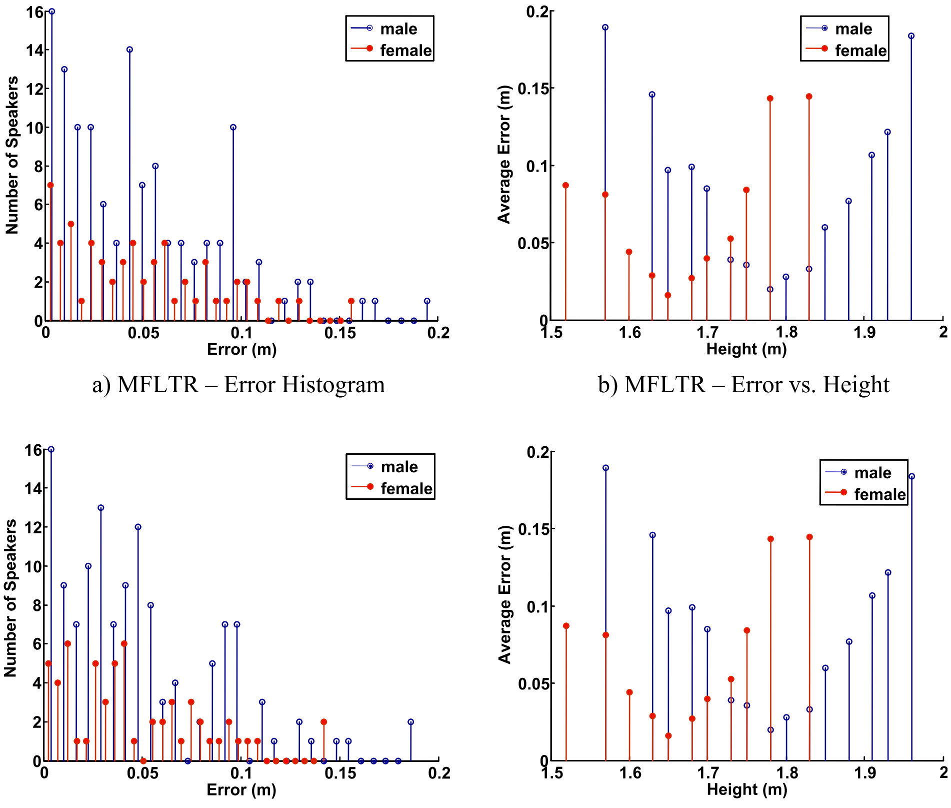
Having demonstrated individual MFLTR and GMM-HDBC performance, the fusion of these systems are now considered. The fusion result is shown in terms of MAE to compare performance with MFLTR. The fusion result is tabulated in Table III.

The combined fusion method achieves an MAE with the highest accuracy of all methods. The GMM-HDBC method when combined with the MFLTR method provides an added level of assurance when the phoneme results are combined. This scheme can be effective when both systems are sufficiently accurate but can still maintain a meaningful performance when a single system does poorly. The improved accuracy of fusion over the MFLTR method demonstrates the benefits of combining the individual MFLTR and GMMHDBC systems. To demonstrate accuracy of the fusion method, the estimated heights are plotted against the actual heights for all speakers in Fig. 5. The center line illustrates perfect accuracy while the outer two lines represent error within a 5 cm range. All points in between the lines show speakers that have height estimates which are less than 5 cm in error. These plots show that a majority of the speakers have error less than 5 cm (63.1% of male speakers, 62.7% of female speakers). Compared to MFLTR, the fusion results are very similar. Considering the close proximity of MAEs for both methods, it is reasonable to expect the plots showing the estimated versus actual height to appear very similar as well. The correlation coefficient for male speakers is 0.33, while for female speakers it is 0.43. When both males and females are grouped together, the overall correlation coefficient increases to 0.73. This is an improvement compared to the MFLTR system alone. The male speaker's correlation coefficient increased by 27% and females by 26%. The combined correlation coefficient, however, only increased from 0.72 to 0.73. This means that the tighter clustering around the perfect accuracy line for males and females was not as beneficial as the increased range of heights when combining male and female speakers for the correlation coefficient calculation.

VIII. CONSISTENCY

A. Error

To explore why each system achieves a particular MAE, as well as any potential shortcomings of a particular method, it is meaningful to explore the specific distribution of error



a) MFLTR – Error Histogram

b) MFLTR – Error vs. Height

c) Fusion – Error Histogram

d) Fusion – Error vs. Height

FIG. 7. (Color online) Histogram of error for MFLTR method and fusion method.

for each speaker. To examine this error more closely, two different graphs are examined. One is a histogram of speaker error and the second is the average error versus height. For the GMM-HDBC system, the within 5 cm classification accuracy is considered per class because an MAE would not be feasible a system that assigns speech samples to height bins rather than soft height scores. Ideally, the method should have a histogram clustered within zero to low error marks and have the same low error across all heights.

The first method examined is the MFLTR method. The MFLTR error profiles for the MFLTR method are displayed in Figs. 7 and 8.

The error distributions show that most of the error is concentrated in the short and tall ends of the height range for male and female speakers with the best performing heights being in the middle. This could be due to the fact that there are a very small number of speakers for the short and tall heights, and those few speakers do not provide an excellent model to represent the general speaker in these height distribution tails. The histogram demonstrates that for male and female test speakers, most of their heights are estimated with high accuracy and the number of speakers that are estimated poorly are very few. The next method examined is the GMM-HDBC method. The accuracy within 5 cm is displayed for each class in Fig. 9 to see which height classes are more accurate.

The GMM-HDBC method for male speakers has a single poor performing class (class bin 1) representing the shortest speaker heights. The average and tall bin classes ranging from 3 to 9 perform well, with the best being classes 3, 4, and 6. The performance tends to increase or stay the same as the confidence measure increases and also has higher initial accuracies. For female speakers, class bin 1 also displays the poorest performance. However, classes 2 through 4 decrease in accuracy as the confidence measure increases. The best performing classes for females are classes 5 through 8, which happen to be the class bins related to average heights and taller heights.

The next method examined is the dual system fusion method. The histogram and error profile are displayed in Figs. 7 and 8. For the fusion method, most of the error is concentrated in the short and tall ranges. Most speakers have low height estimation errors, which are also seen for the MFLTR method. By combining the MFLTR and GMM-HDBC systems, the height error has decreased for all but three of the heights for males, 1.96, 1.8, and 1.78 m and all but three of the heights for females, 1.63, 1.65, and 1.78 m. For female speakers using the GMM-HDBC method, the upper half of the height range shows superior performance with accuracy improving when combined with MFLTR for taller heights. For male speakers, the heights that perform poorly can be matched to a poor performing class in the GMM-HDBC method or one that experienced a decline in accuracy as the confidence measure increased. The histogram of error for MFLTR and fusion methods has the same overall general shape with a higher number for speakers having low estimation errors with a sloping downward trend as the error increases.

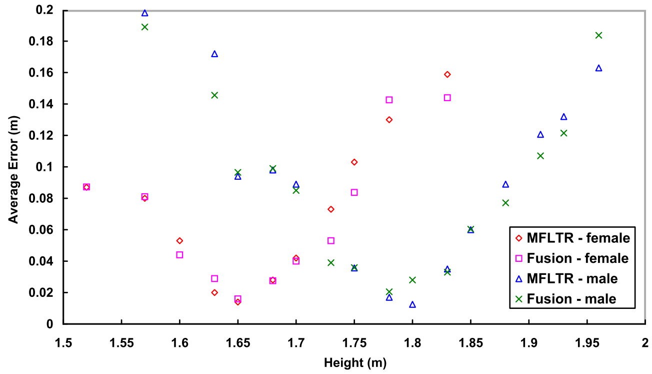


FIG. 8. (Color online) Average error vs height for MFLTR method and fusion method.

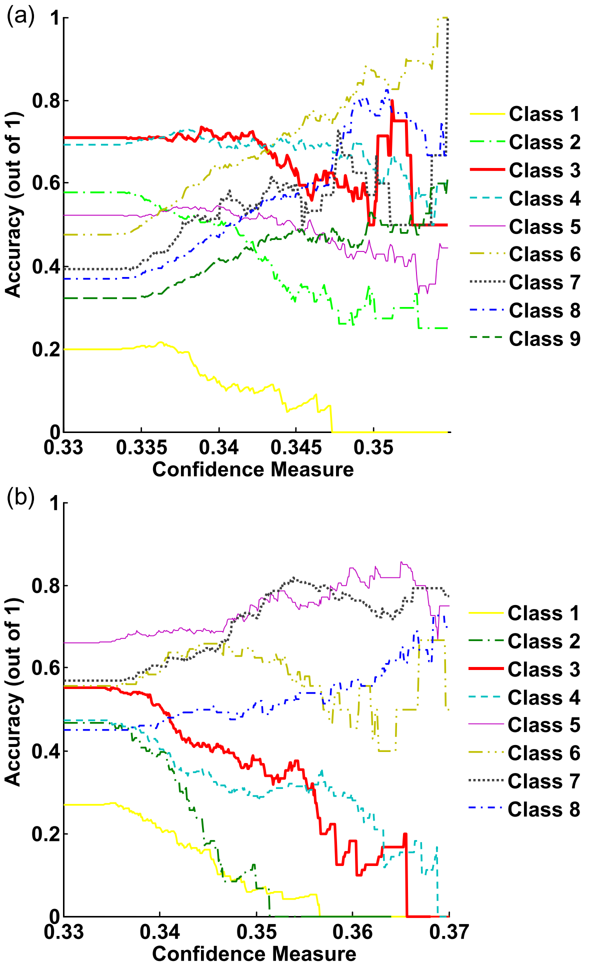


FIG. 9. (Color online) Comparative analysis of accuracy within 5 cm across height classes in GMM-HDBC method, (a) male speakers and (b) female speakers.

B. Coefficients

Another way to assess MFLTR consistency is to analyze the rate of changes in the correlation coefficients and system performance changes when the training set changes (i.e., exploring the repeatability of the evaluations based on the training speakers). To evaluate this, five different training and testing sets are considered. Each set is produced by switching a portion of training and test speakers falling into

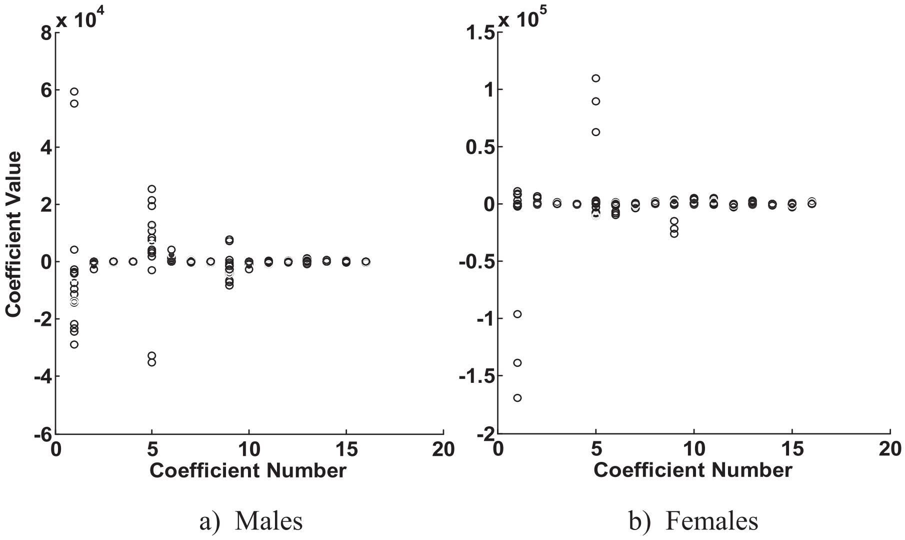


FIG. 10. Coefficients for formant equation [Eq. (7)] with five different training sets; left/right, coefficients for males/females, respectively. Note high variation in coefficients 1, 5, 9.

the same height category. In this way, new training sets that contain different speakers but exhibit the same height distributions are generated. The first set is the original set. The second, third, fourth, and fifth sets contain a 10.9%, 18.8%, 26.1%, and 33.3% change in training speakers, respectively, from the original for male speakers and a 16.17%, 29.41%, 42.65%, and 52.94% change in training speakers for female speakers. To compare the differences caused by this change in training set speakers, the MAE, regression coefficients, histogram of height error, and average error versus height are compared for the five sets. In ideal case, the error performance would be consistent for all five sets, confirming the reliability of the MFLTR system. The first measure examined is the regression coefficients stemming from the training of the regression equations in the MFLTR approach. The coefficients are displayed for each phoneme, and for all five sets, in a scatter graph in Figs. 10 and 11. The graph with 16 coefficients refers to the formant equation [e.g., Eq. (7)] while the graph with 18 coefficients refers to the equation using LSFs [e.g., Eq. (8)].

For both male and female speakers, all regression coefficients for the formant equation are extremely close with the exception of 1, 5, and 9, which applies for both males and females. Regression coefficients 1, 5, and 9 refer to the coefficients in front of the first three formants raised to the 4th power. The effect of this discrepancy will be examined by comparing height performance. The regression coefficients for the LSF equation are generally consistent for male speakers but have a wider variability for female speakers. For male speakers, regression coefficients 1 and 2, which are the coefficients for the first and second LSF, are spread out, which was similarly seen for female LSF equation coefficients. Again the effect of this will be examined in the performance metrics. The next evaluation will compare the overall MAE and phoneme/feature dependent MAE for the five different sets (e.g., see Tables IV and V). The LSF refers to LSF Eq. (8) and 4 F refers to the formant Eq. (7).

The MAEs for the five training sets for the LSF equation are not as tightly clustered as that for the formant equation. However, some of the formant equation sets have a significant shift as in the /AE/ phoneme for both male and female speakers. This matches the observations made for regression coefficients. The regression coefficients for the formants were generally tightly clustered, and the LSFs regression

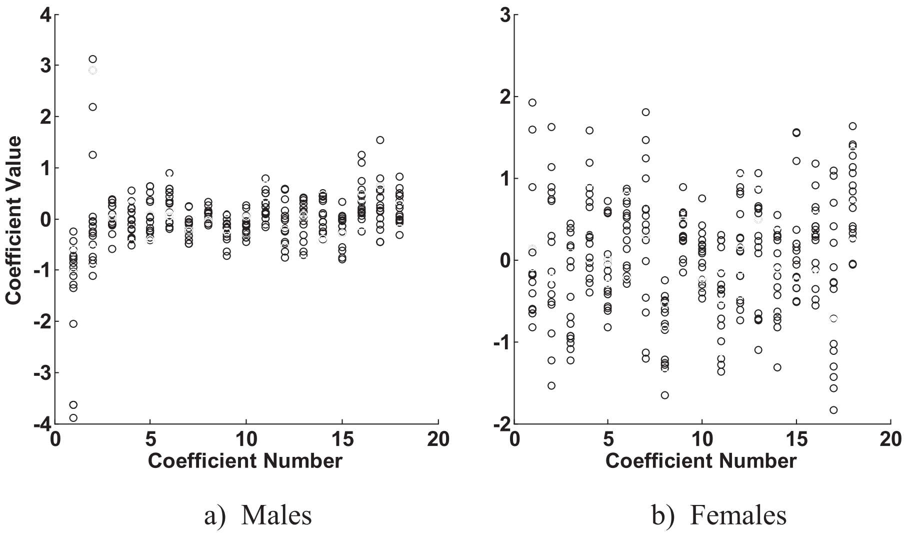


FIG. 11. Coefficients for LSF equation with five different training sets; left/right, coefficients for males/females, respectively.

TABLE IV. Overall MAE (in cm) and phone/feature dependent MAE for MFLTR method for five different male speaker train sets; LSF refers to LSF height prediction Eq, (8) and 4 F to formant prediction Eq. (7).

	MAE (cm)					
Set	1	2	3	4	5	σ^2
/AA/ LSF	5.39	5.76	5.85	5.62	5.77	0.181
/AA/ 4 F	5.14	5.13	5.11	5.26	5.15	0.059
/AE/ LSF	5.29	5.69	5.53	5.77	5.66	0.188
/AE/ 4 F	5.26	5.33	6.91	5.02	4.91	0.814
/AO/ LSF	5.53	5.66	5.47	5.48	5.33	0.119
/AO/ 4 F	5.45	5.38	5.52	5.44	5.30	0.083
/IY/ LSF	5.07	5.47	5.46	5.73	5.56	0.242
/IY/ 4 F	5.49	5.41	5.25	5.26	5.25	0.111
Overall	4.93	5.18	5.10	5.10	5.05	0.092

coefficients while close were not as compact. The next item examined is the error histogram and the average error versus height for all five sets (e.g., see Figs. 12 and 13).

The average error versus height for the five different training sets show the same general shape with greater average error in both the short and tall end of the height scale. The average errors for each height are also quite close except for 1.65 m for males and 1.52 and 1.75 m for females. The histogram for all five sets follows the same pattern where the number of speakers with that error decreases as the error increases. Between the five training sets, the frequency for the greater errors stays practically the same while for the smaller errors there are more changes. Overall between the MAE, histogram, average error versus height, and regression coefficients, the five training sets perform in an extremely consistent manner. From this, it can be concluded that varying the speakers in the training set has little impact on the MFLTR method which is what is expected.

C. Session to session variability

Another way to consider system consistency is to examine a single speaker across multiple sessions recorded on different days. This is a typical performance challenge in achieving effective algorithm performance for speaker recognition (Godin and Hansen, 2010). In repeated trials for the same speaker over different recording sessions with the same data capture setup, height estimation schemes should result in the same height. To examine this, an in-house corpus, Multi-session Audio Research Project (MARP), was used that consisted of speakers reading sentences in various sessions throughout a 3 yr period (Godin and Hansen, 2010). For consistency consideration, three male and three female speakers were chosen as well as one sentence from three different sessions, see Tables VI and VII. The regression coefficients and height GMMs used are those trained from the independent TIMIT data. The actual heights of these speakers are unknown, so the absolute accuracy cannot be examined. As a result, the consistency of the estimated height results across the sessions is the only item considered for the MFLTR method and the GMM-HDBC method.

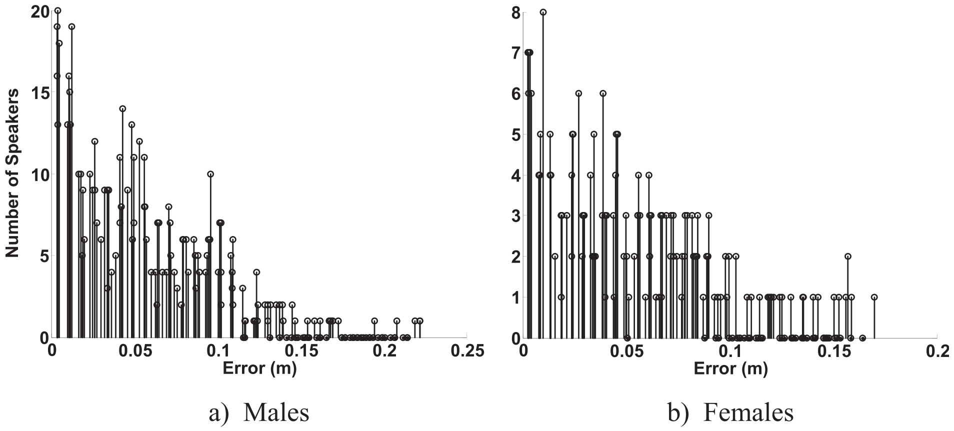
TABLE V. Overall MAE (in cm) and phone/feature dependent MAE for MFLTR method for five different female speaker train sets; LSF refers to LSF height prediction Eq, (8) and 4 F to formant prediction Eq. (7).

	MAE (cm)					
Set	1	2	3	4	5	σ^2
/AA/ LSF	5.55	5.43	4.98	4.84	5.25	0.298
/AA/ 4 F	5.06	5.10	4.95	5.28	5.00	0.127
/AE/ LSF	5.61	4.88	4.78	4.71	4.55	0.411
/AE/ 4 F	4.72	4.22	4.83	4.72	4.57	0.238
/AO/ LSF	6.32	7.14	6.95	7.50	6.87	0.431
/AO/ 4 F	5.44	5.24	5.24	5.82	5.63	0.252
/IY/ LSF	5.08	5.66	5.66	6.13	5.77	0.377
/IY/ 4 F	5.35	5.04	4.85	5.75	4.90	0.374
Overall	4.76	4.76	4.67	4.79	4.51	0.114

The MFLTR results for each subject are all consistent with no more than a 4 cm difference among all of the session heights for any speaker. The GMM-HDBC results show similar level of consistency. Speaker AAA and ABA result in the same class for all three sessions, while the remaining speakers are assigned neighboring classes for at least one of the sessions. It should be noted that the time difference between each session in MARP is 1 month with a total elapsed time across all sessions of as much as 36 months. Not all speakers were captured each month due to travel or other personal reasons. Between the two different height estimation methods, the results are very close for all but speaker ABA. The remaining speakers for MFLTR results are either close to the boundaries of the height range or within the limit. For speaker ABA, there is approximately a 10 cm difference between the upper height range and the MFLTR results. Considering that the overall system is not flawless in height estimation, having the two individual methods disagree for some speakers can be expected. Both methods demonstrate strong consistency even with open test data not seen by the system during training.

IX. OPEN SET TESTING

As a final exploration, the individual height estimation solutions are evaluated on open public speaker data. Earlier results in this study for the MFLTR and GMM-HDBC methods considered open training and test speakers; however, both of these speaker groups were drawn from the same speech corpus, TIMIT. A true open test of the height estimation solution would be to use data that the system was in no way originally trained to be able to handle. It should be noted that microphone, communication hand-set, communication channel, or other acoustic mismatch could impact performance for any speech based system. To accomplish this last evaluation, eight male and eight female movie/TV actors were chosen to be test speakers due to the availability of speech from interviews, movies, etc. The speech was drawn from YouTube where at least one of the four vowels, /AA/, /AE/, /AO/, and /IY/, had to be included in the specific test speech. The speech data were chosen to have minimal background noise. This constraint was very easy to accomplish due to the plethora of words associated with these vowels. The male and female actors' names with actual heights were obtained from the internet movie database and summarized



a) Males

b) Females

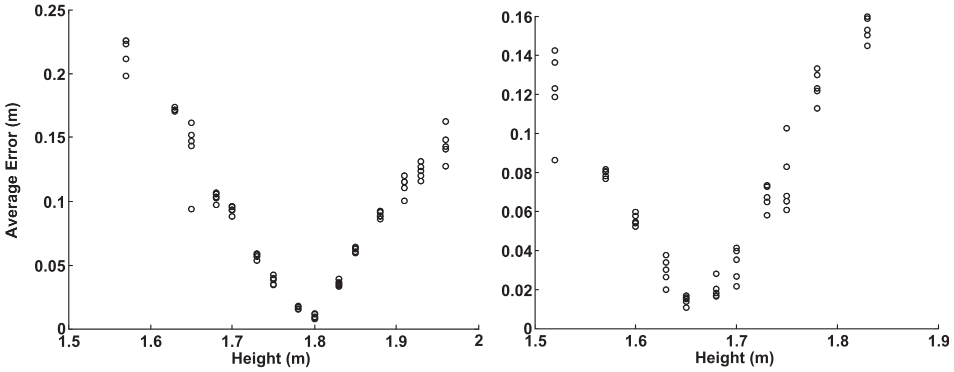
FIG. 12. Histogram of error for MFLTR method with five different training sets; left/right, coefficients for males/females, respectively.

in Table VIII as well as the estimated height resulting from the individual MFLTR and the GMM-HDBC methods. The error for both methods is included in this table as well. For MFLTR, the error is calculated by subtracting the actual height from the estimated height. For the GMM-HDBC method, if the actual height is within the height bin class, then the label "IN" is presented together with the height class. If it is outside the height class bin, then the actual height is subtracted from the median of the height class and shown.

Using open YouTube audio data, both MFLTR and GMM-HDBC methods perform well and achieve similar results when compared to the previous TIMIT open test results. For male actors, the MFLTR method produced individual errors ranging from 1.58 to 6.53 cm. The highest error occurred with the tallest speaker. For females, the MFLTR performed similarly with individual errors ranging from 1.7 to 8.29 cm. The highest errors generally also occurred for the tallest females as well. The error for all speakers is consistent across the height ranges unlike the TIMIT test results, which had greater error in the height distribution tails. However, the TIMIT test results did have many speakers with error under 1 cm (17.69% of the male speakers, 16.94% of the female speakers). For the GMM-HDBC method, two male speakers were classified correctly with another four male speakers actual heights being close to one of the boundaries. The female set had four speakers correctly classified with another three having an actual height close to one of the boundaries. The same pattern is seen with the TIMIT data; this was one of the reasons that the combined fusion method helped improve MFLTR results because the closest boundary to the MFLTR result is used to calculate the final height. Overall, the height estimation system trained with TIMIT data can meaningfully estimate a speaker's height even with channel mismatch as seen in open YouTube audio sets.

X. CONCLUSION

In this study, the problem of accurate height estimation from speech was investigated. Two alternative solutions were developed for engaging in automatic speaker height estimation as well as a fusion of the two individual methods. The first method, MFLTR, obtains a point estimate of height for each speaker but requires an occurrence of at least one of four specific vowels in the test sample. The proposed GMM-HDBC statistical method is text independent, but rather than exact height, it assigns a height bin class representing a range of heights. This classification method also produces a complementary confidence measure. To utilize the complementary information produced by the two methods, a fusion system was also developed. The fusion system produces a single height estimate per speaker and improves the accuracy of the MFLTR regression method by utilizing the additional height bin class information and confidence score. An error analysis in the MFLTR, GMM-HDBC, and fusion systems was performed to provide better understanding of the respective performances. All of solutions showed higher error for subjects in the short and tall height range (i.e., the tails of the height distribution for males and females). Evaluations were performed using multiple alternate training sets and open



a) Males

b) Females

FIG. 13. Average error vs height for MFLTR method with five different training sets; left/right, coefficients for males/females, respectively.

TABLE VI. MFLTR height estimation results across speaker sessions in MARP Corpus; three sessions per speaker acquired throughout a 3-yr period.

	MFLTR [Height (m)]		
	Male		
Session	1	2	3
AAA	1.8526	1.8688	1.8897
AAL	1.8566	1.8741	1.8404
AAN	1.8290	1.8384	1.8367
	Female		
Session	1	2	3
AAH	1.6890	1.7036	1.6911
ABA	1.6985	1.6779	1.6504
ACC	1.6257	1.6379	1.6328

TABLE VIII. Corpus of actors speech—open test set results for MFLTR and GMM-HDBC methods.

	Actual height (m)	MFLTR (m) [Error (cm)]	GMM-HDBC (m) [Error from median (cm)]
		Male	
Adam Baldwin	1.93	1.8647 (−6.53)	1.765–1.79 (−15)
Christian Kane	1.78	1.7396 (−4.04)	1.9–1.96 (+15.5)
David Boreanaz	1.85	1.8174 (−3.26)	1.765–1.79 (−7)
Jim Parsons	1.86	1.7955 (−6.45)	1.765–1.79 (−8)
Johnny Galecki	1.65	1.6480 (−0.2)	1.57–1.715 (IN)
Nicholas Brendan	1.80	1.8413 (+4.13)	1.765–1.79 (−2)
Seth Green	1.63	1.6458 (+1.58)	1.57–1.715 (IN)
Simon Helburg	1.70	1.6769 (−2.31)	1.765–1.79 (+8)
		Female	
Alyson Hannigan	1.68	1.6420 (−3.80)	1.665–1.69 (IN)
Amy Acker	1.73	1.7129 (−1.72)	1.74–1.83 (+6)
Gina Torres	1.78	1.8286 (+4.86)	1.74–1.83 (IN)
Kaley Cuoco	1.65	1.6247 (−2.53)	1.45–1.585 (−14)
Mayim Bialik	1.63	1.6493 (+1.93)	1.45–1.585 (−12)
Melissa Rauch	1.52	1.5504 (+3.04)	1.45–1.585 (IN)
Sara Gilbert	1.60	1.5830 (−1.70)	1.45–1.585 (−9)
Sarah Rafferty	1.75	1.6671 (−8.29)	1.74–1.83 (IN)

session-to-session test speakers and open public speech data from TV/movie actors. The MFLTR and GMM-HDBC systems demonstrated their consistency and accuracy through open testing as well as across session-to-session speaker with recordings over extended time periods. Compared to previous investigations on height estimation, these systems are at least equal or in most cases outperform previous methods in terms of MAE. The MFLTR method achieved an MAE of 4.93 and 4.76 cm, and the fusion method achieved an MAE of 4.89 and 4.55 cm for males and females from the TIMIT database, respectively. It should be noted that the partitioning of the TIMIT data in our study differs from the reference studies and also differs among the reference studies themselves (e.g., Ganchev *et al.*, 2010a versus Arsikere *et al.*, 2013). In this sense, the obtained results cannot be compared in absolute terms. However, it is assumed the task difficulty is comparable as the training and test samples were all drawn from the same TIMIT speaker population.

A majority of the previous studies that used MAE as a performance measure achieved errors exceeding 5 cm (Ganchev *et al.*, 2010a; Mporas and Ganchev, 2009; Arsikere *et al.*, 2012, 2013; Williams and Hansen, 2013). An exception can be found in Ganchev *et al.* (2010b), where a MAE of 4.1 cm was reached for one of the open environment test sets (an indoor set) with a Gaussian process based regression scheme. The same scheme yielded a MAE of 5.3 cm on a TIMIT test set. It is noted that the test set followed the traditional training/test set partitioning introduced by the TIMIT authors for automatic speech recognition experiments, and hence the MAE cannot be directly compared to the results in our study in absolute terms.

TABLE VII. GMM-HDBC height interval results across speaker sessions in MARP corpus; three sessions per speaker acquired throughout three year period.

	GMM-HDBC [Height range (m)]		
	Male		
Session	1	2	3
AAA	1.790-1.815	1.790-1.815	1.790-1.815
AAL	1.790-1.815	1.840-1.865	1.790-1.815
AAN	1.765-1.790	1.840-1.864	1.790-1.815
	Female		
Session	1	2	3
AAH	1.615-1.640	1.640-1.665	1.615-1.640
ABA	1.450-1.585	1.450-1.585	1.450-1.585
ACC	1.665-1.690	1.450-1.585	1.640-1.665

Overall the MFLTR and GMM-HDBC methods have been shown to provide a reasonably accurate height estimation when utilized separately as well as in combination. The MFLTR method relies on the occurrence of a particular set of phones; however, as shown with the open set testing, it is not a requirement that all designed vowels be present. The algorithms presented in this study are not limited to merely performing height estimation as a speaker trait. As future directions, height estimations could be employed to help improve speaker identification systems as well as vocal-tract length normalization (VTLN) algorithms for normalizing speaker differences in speaker independent speech recognition. For speaker identification systems, height estimation as a speaker trait could be used as confidence measures or help cluster in-set/out-of-set speaker models (Angkititrakul and Hansen, 2007) in the selection of speaker identity. For VTLN, the work here could be used to help calculate an improved warping factor to increase ASR system performance. Height estimation could be employed in the selection of cohort speakers in audio lineups for forensic speaker analysis. Overall height estimation could be used in various capacities to help improve speech processing or language technology systems with advances not being limited to merely improving algorithm accuracy but providing further knowledge of subjects for human-computer interaction.

ACKNOWLEDGMENT

This project was funded by AFRL under Contract No. FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L.H.

[1] Information on cumulative percent distribution of population by height and sex is available at http://www.allcountries.org/uscensus/230_cumulative_percent_distribution_of_population_by.html (Last viewed 05/10/2014).

Angkititrakul, P., and Hansen, J. H. L. (**2007**). "Discriminative in-set/out-of-set speaker recognition," IEEE Trans. Audio Speech Lang. Process. **15**, 498–508.

Arsikere, H., Leung, G., Lulich, S., and Alwan, A. (**2012**). "Automatic height estimation using the second subglottal resonance," in *IEEE International Conference on Acoustics, Speech and Signal Processing 2012 (ICASSP)*, pp. 3989–3992.

Arsikere, H., Leung, G. K., Lulich, S. M., and Alwan, A. (**2013**). "Automatic estimation of the first three subglottal resonances from adults speech signals with application to speaker height estimation," Speech Commun. **55**, 51–70.

Brestoff, J., Perry, I., and Van der Broeck, J. (**2011**). "Challenging the role of social norms regarding body weight as an explanation for weight, height, and BMI misreporting biases: Development and application of a new approach to examining misreporting and misclassification bias in surveys," BMC Public Health **11**, 331–341.

Crosmer, J., and Barnwell, T. P., I. (**1985**). "A low bit rate segment vocoder based on line spectrum pairs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing 1985 (ICASSP)*, Vol. 10, pp. 240–243.

CRSS (**2015**). "Training and test evaluation lists for height estimation," http://www.utdallas.edu/~hxb076000/HeightEstimation (Last viewed 05/29/2015).

Dusan, S. (**2005**). "Estimation of speakers height and vocal tract length from speech signal," in *INTERSPEECH (ISCA)*, pp. 1989–1992.

Eide, E., and Gish, H. (**1996**). "A parametric approach to vocal tract length normalization," in *Proceedings of ICASSP 1996* (IEEE Computer Society, Los Alamitos, CA), Vol. 1, pp. 346–348.

Eyben, F., Wollmer, M., and Schuller, B. (**2009**). "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009 (ACII 2009)*, pp. 1–6.

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

Ganchev, T., Mporas, I., and Fakotakis, N. (**2010a**). "Audio features selection for automatic height estimation from speech," in *Lecture Notes in Computer Science. Artificial Intelligence: Theories, Models and Applications*, edited by S. Konstantopoulos, S. Perantonis, V. Karkaletsis, C. Spyropoulos, and G. Vouros (Springer, Berlin), Vol. 6040, pp. 81–90.

Ganchev, T., Mporas, I., and Fakotakis, N. (**2010b**). "Automatic height estimation from speech in real-world setup," in *Proceedings of EUSIPCO 2010*, Aalborg, Denmark, pp. 800–804.

Godin, K. W., and Hansen, J. H. L. (**2010**). "Session variability contrasts in the MARP corpus," in *INTERSPEECH (ISCA)*, pp. 298–301.

Greenwood, A. R., and Goodyear, C. C. (**1994**). "A polynomial approximation to the acoustic-to-articulatory mapping," in *IEE Colloquium on Techniques for Speech Processing and their Application*, pp. 8/1–8/6.

Greisbach, R. (**1999**). "Estimation of speaker height from formant frequencies," Forensic Ling. **6**, 265–277.

Hansen, J. H. L. (**1988**). "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA.

Hasan, T., Sadjadi, O., Gang, L., Shokouhi, N., Bořil, H., and Hansen, J. H. L. (**2013**). "CRSS systems for 2012 NIST speaker recognition evaluation," in *IEEE ICASSP 2013*, Vancouver, Canada, pp. 6783–6787.

Itakura, F. (**1975**). "Line spectrum representation of linear predictor coefficients of speech signals," J. Acoust. Soc. Am. **57**, S35.

Jain, A. K., Dass, S. C., and Nandakumar, K. (**2004**). "Can soft biometric traits assist user recognition?," in *SPIE—Biometric Technology for Human Identification*, Vol. 5404, pp. 561–572.

Kent, R. D., and Read, C. (**1992**). *The Acoustic Analysis of Speech* (Whurr Publishers, San Diego), p. 22.

Kinnunen, T., and Li, H. (**2010**). "An overview of text-independent speaker recognition: From features to supervectors," Speech Commun. **52**, 12–40.

Künzel, H. J. (**1989**). "How well does average fundamental frequency correlate with speaker height and weight?," Phonetica **46**, 117–125.

Lamel, L. F., and luc Gauvain, J. (**1995**). "A phone-based approach to non-linguistic speech feature identification," Comput. Speech Lang. **9**, 87–103.

Lass, N. J., and Brown, W. S. (**1978**). "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies," J. Acoust. Soc. Am. **63**, 1218–1220.

Lee, L., and Rose, R. (**1996**). "Speaker normalization using efficient frequency warping procedures," in *Proeedings. of ICASSP'96* (IEEE Computer Society, Los Alamitos, CA), Vol. 1, pp. 353–356.

Mporas, I., and Ganchev, T. (**2009**). "Estimation of unknown speakers height from speech," Int. J. Speech Technol. **12**, 149–160.

National Institute of Standards and Technology (**1988**). *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database* (Gaithersburg, MD).

Pellom, B. L., and Hansen, J. H. L. (**1997**). "Voice analysis in adverse conditions: The Centennial Olympic Park Bombing 911 call," in *Proceedings of the 40th Midwest Symposium on Circuits and Systems 1997*, Vol. 2, pp. 873–876.

Rabiner, L., and Schafer, R. (**2011**). "Algorithms for estimating speech parameters," in *Theory and Applications of Digital Speech Processing*, 1st ed. (Pearson Higher Education, Upper Saddle River, NJ), pp. 548–645.

Rendall, D., Kollias, S., Ney, C., and Lloyd, P. (**2005**). "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. Am. **117**, 944–955.

Reynolds, D. A. (**1995**). "Speaker identification and verification using Gaussian mixture speaker models," Speech Commun. **17**, 91–108.

Schwarz, P. (**2009**). "Phoneme recognition based on long temporal context," Ph.D. thesis, Brno University of Technology, Czech Republic.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.

van Dommelen, W. A., and Moxness, B. H. (**1995**). "Acoustic parameters in speaker height and weight identification: Sex-specific behaviour," Lang. Speech **38**, 267–287.

Williams, K., and Hansen, J. (**2013**). "Speaker height estimation combining GMM and linear regression sub-systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing 2013 (ICASSP)*, pp. 7552–7556.