SHORT-TERM ANALYSIS FOR ESTIMATING PHYSICAL PARAMETERS OF SPEAKERS

Rita Singh, Bhiksha Raj, James Baker

Carnegie Mellon University, Pittsburgh, PA, USA
{rsingh, bhiksha, jkbaker}@cs.cmu.edu

ABSTRACT

Conventional approaches to estimating speakers' physiometric parameters such as height, age, weight etc. from their voice analyze the speech signal at relatively coarse time resolutions, typically with analysis windows of 25ms or longer. At these resolutions the analysis effectively captures the structure of the *supra-glottal* vocal tract. In this paper we hypothesize that by analyzing the signal at a finer temporal resolution that is lower than a pitch period, it may be possible to analyze segments of the speech signal that are obtained entirely when the glottis is open, and thereby capture some of the subglottal structure that may be represented in the voice. To explore this hypothesis we propose an analysis approach that combines signal analysis techniques suited to fine-temporal-resolution analysis and well-known regression models. We test it on the prediction of heights and ages of speakers from a standard speech database. Our findings show that the higher-resolution analysis does provide benefits over conventional analysis for estimating speaker height, although it is less useful in predicting age.

Index Terms— Physiometric measurements, Voice biometrics, Voice forensics, Height, Age, Short-time analysis

1. INTRODUCTION

In its capacity as a biometric signal, the human voice has been the subject of scientific investigations in a large number of fields. Amongst other things, researchers have sought to investigate its correlations to various physiometric parameters of the speaker, which includes the set of physiological parameters such as age [1, 2], and physical parameters such as height [3, 4], weight [5, 6], body size [7, 8, 9, 10] etc. The underlying expectation is that the dimensions and health of various parts of the human speech production apparatus including the lungs, the larynx, the length of the vocal tract, the oral cavities etc. depend on the size and age of the person. The characteristics of the speech production apparatus, in turn, affect the nature of the speech produced by it. Ergo, correlations must exist between the speech signal and and the physiometric parameters of the speaker.

From a forensic perspective, these correlations may be exploited to make prognostications about the speaker from samples of their speech, a fact that has not escaped the observation of researchers. Most attention has been paid to estimating the height of the speaker. Griesbach, Ganchev *et al.* explore a number of utterance-level characteristics of the speech signal and a number of regression strategies in order to estimate the speaker's height [3, 11, 4]. Williams and Hansen suggest the fusion of multiple regression strategies for the same problem based on spectral characterizations of the signal in [12, 13]. Poorjam and co-authors attempt to predict speaker height from factors of aggregate statistical characterizations of the signal [14]. Arsikere *et al.* [15, 16] and others suggest the prediction of speaker height through a preliminary estimate of the speaker's subglottal resonances.

In [17] Schotz attempts to predict speaker *age* through a CART tree applied to speech measurements. Muller and Burkhardt [18] attempt to predict age from cepstral and pitch measurements. Li *et al.* [19] consider prosodic cues for the same problem. Bahari *et al.* [20] do so based on characterizations of aggregate statistics. In general, though, prediction of age from voice is considerably less successful than the prediction of height.

In all cases, the analyses of the signal utilized to derive features that may be predictive of the feature being estimated are temporally coarse-grained. The speech signal is generally segmented into frames of about 25 milliseconds, an analysis that is no different than that applied to other tasks such as speech recognition. Approaches such as [14] further do not resolve the frequency or temporal structure of the signal, and represent it through aggregate statistical characterizations computed over ensembles of cepstral or spectral vectors derived from it. While [16] and [13] consider more physiologically motivated features such as formants and subglottal-resonances, these too are generally derived from temporally-coarse characterizations of the signal and face the loss of resolution it entails.

The relatively broad analysis window of 25ms is a good compromise by being not too short for effective spectral analysis, and not too broad for the spectral characteristics of the signal to change considerably within the window, and is an excellent choice for pattern recognition tasks such as speech or speaker recognition. However, it is not clear that this is the best choice of analysis window for physiometric characterization of the speaker. As we argue in Section 2, this may, in fact, be too wide for effective characterization of *lower* portions of the vocal tract below the glottis, and not wide enough to characterize the *upper* regions above the glottis well.

We investigate this hypothesis by analyzing the speech signal at various temporal resolutions. At higher temporal resolutions conventional spectral analyses do not provide useful characterizations. We also propose alternate mechanisms to compute features for these analysis. While we cannot directly determine whether these analyses do indeed explicitly capture subglottal structure, we can evaluate them through secondary effects – our ability to predict physiometric parameters through these measurements. We investigate the utility of the analyses at these various temporal scales for the problems of estimation of speaker height and speaker age.

Our results are mixed. While we find that analyses of the speech signal with greater temporal resolution does provide gains when predicting height, compared to conventional analysis, the improvement is minor. However the higher resolution analyses provide complementary information that can be combined with conventional analyses for improved predictions. Observed improvements in age prediction are noted to be statistically insignificant when compared to a default baseline predictor. We also note in passing that much reported literature also lists prediction accuracies that are comparable

978-1-4673-9448-2/16/$31.00

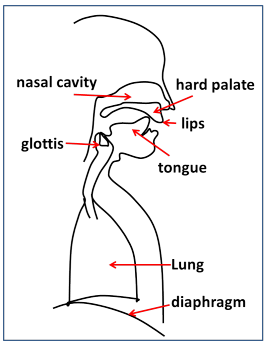

Fig. 1. The human vocal apparatus.

to or worse than the baseline predictors, at least on the dataset we evaluate predictions on.

The rest of our paper is as follows. In Section 2 we briefly outline the motivation behind our hypotheses. In Section 3 we describe the signal processing procedures required to derive features at higher temporal resolution. In Section 4 we describe our overall analysis and prediction setup. In Section 5 we describe our experiments and finally, in Section 6 we present our conclusions.

2. TEMPORAL RESOLUTION OF SPEECH PRODUCTION

The speech production system is well known. Figure 1 shows a summary illustration of the system. The entire structure includes the lungs, the glottis, the epiglottal opening, the pharynx, and the oral and nasal cavities. The glottis and the regions below the glottis including the trachea and the lungs are the "sub-glottal" regions of the vocal apparatus. The regions above the glottis including the pharynx and the oral and nasal cavities are the "supra-glottal" regions. During speech production air pressure generated by the lungs makes its way past the glottis and excites the vocal tract. The speech we hear is the response of the vocal tract to this excitation. Variations of sound are obtained by manipulating the structures of the vocal cavity, creating different resonance chambers and varying their resonant frequencies.

A key aspect of the entire process is the *excitation* produced by the lungs. The passage from the lungs to the rest of the vocal tract is "gated" by the vocal folds (2), which control the glottal opening between the lungs and the rest of the vocal tract. During *unvoiced* sounds the vocal folds do not vibrate and the glottis remains open. The airflow through the glottis is continuous, resulting in noise-like sounds such as the fricated /s/, /sh/, /hh/, /f/ etc.

During *phonated* speech, including vowel and other voiced sounds, on the other hand, the vocal folds vibrate, resulting in periodic complete or partial closure of the glottis. This results in a pulsed airflow through the glottis, which gives voiced sounds their periodic nature. Figure 3 shows a typical pattern of glottal opening for a voiced sound, the resulting glottal airflow and the periodic speech signal that results. The *pitch* period of the signal is the spacing between concurrent pulses in the glottal waveform. Typical voiced speech has a pitch ranging from about 80Hz to 400Hz, corresponding to a pitch period ranging from 2.5ms to about 12.5 ms.

As can be inferred from the above discussion, the speech signal provides cues for physiometric characterization. The speaker's height is clearly related to the length of the vocal tract – taller people may be expected to have longer vocal tracts. Hence, the vocal tract

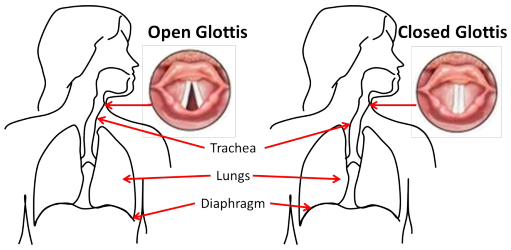

Fig. 2. Open and closed glottis.

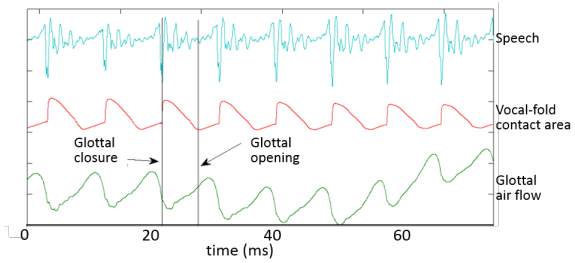

Fig. 3. The upper plot shows the speech waveform. The middle plot is an electroglottograph measurement showing vocal-fold contact area. The lower curve is the estimated glottal airflow.

resonances present in the speech signal may be expected to inform us of the speaker's height. Information about the speaker's height is also present in the regions of the vocal apparatus that are below the glottis. The length of the trachea, the diaphragm, and the size of the lungs relate to the speaker's body size. Consequently, we may also expect to derive information about the speaker's body size from the effect of sub-glottal structures on the speech signal. Indeed, at least one of these characteristics, namely sub-glottal resonances, have been demonstrated to be good predictors of peoples' heights [16]. Sub-glottal structures are also affected by age; hence we may also expect to derive information about the talker's age from characterizing sub-glottal phenomena in the speech signal.

This brings us to the key point we aim to make with the above discussion. Conventional analysis of speech signals typically analyzes the signal in windows of about 25ms. In voiced sounds a single window will hence include multiple pitch periods, even for speech with the lowest pitch. Thus each analysis window includes regions of both *closed* and *open* glottis. A somewhat paradoxical phenomenon that results from the acoustics of closed tubes is that the energy in the speech signal is, in fact, considerably higher when the glottis is *closed* than when it is open. This is also illustrated by Figure 3, where we notice that the amplitude of the speech signal is much higher when the vocal-fold contact area is largest, *i.e.* when the glottis is closed. The signal is seen to be much weaker when the glottis is actually open. In the glottis-closed phase the speech signal primarily characterizes the acoustic properties of the vocal tract above the glottis, since the region below does not contribute to the signal. Consequently, in any analysis window of 25ms or longer, the spectral energy in any analysis window largely reflects the properties of the *supra*-glottal vocal tract, which overwhelms the acoustic signatures of the sub-glottal regions.

Unvoiced sounds other than /hh/ are usually associated with a constriction of the vocal tract and the spectrum primarily reflects the acoustic properties of the supra-glottal vocal tract ahead of the

constriction. We do not, in general, expect to get strong signatures of sub-glottal phenomena from unvoiced sounds.

We propose that as a remedy shorter analysis windows, such as those that fall entirely within a single open period of the glottis, will enable us to effectively "look into" the sub-glottal regions of the vocal tract by analyzing segments of the signal that are obtained entirely when the glottis is open during voiced portions of the signal. In effect, by taking snapshots in the brief periods when the glottis is open, signal analysis may be able to capture some information about what lies beyond it. As Figure 3 shows, glottal opening lasts considerably less than a pitch period. Pitch periods can vary from 12.5ms to 2.5ms (or even go lower or higher on occasion), with the higher range generally holding for females. This would argue for analysis windows that may be as low as 1ms.

On the other hand, the conventional analysis windows of 25ms typically only consider one or two glottal closure periods. This may be insufficient to capture resonances of the vocal tract with longer time constants. This argues for *longer* analysis windows than the ordinary.

We will therefore investigate the extraction of features from both extremely short analysis windows, and analysis windows that are much longer than conventional analysis windows as the basis for predicting height. For the purposes of this study we will restrict ourselves to spectral characterizations, and not consider the more detailed analyses, such as of formants etc., that are commonly used in this context.

3. ESTIMATING THE SPECTRUM

Estimation of spectral characteristics from extremely small windows however comes with predictable time-frequency resolution tradeoffs. The speech signal is typically sampled at 16000 samples per second, a sampling rate that is adequate to capture information upto 8000Hz. However, in a 1ms window this translates to a mere 16 samples, and thus to a discrete Fourier transform with only 9 magnitude spectral values spanning all frequencies. This is clearly inadequate to obtain a well-resolved spectrum that can identify the spectral detail we wish to capture.

In order to increase the number of samples in any analysis window, we upsample the signal to 256000 samples per second. Simply upsampling the signal to obtain a larger number of samples in the analysis window does not solve the problem. A Fourier spectrum obtained from the upsampled signal is simply an interpolation of the spectrum obtained with the lower sampling frequency. We illustrate this in the upper panels of Figure 4.

Instead we use autoregressive (AR) spectral analysis to estimate the spectrum of the signal [21]. AR analysis models the signal as the output of an autoregressive process, such that the n^{th} sample in the signal is obtained as $s[n] = \sum_{k=1}^{K} a_k s[n-k] + e[n]$, where K is the order of the autoregression, a_k are the regression parameters, and $e[n]$ is the innovation that drives the process, and is assumed to be white. This is equivalent to modelling the signal as the output of an all-pole filter $H(z) = \frac{g}{1-\sum_k a_k z^{-k}}$ excited by white noise. The filter can also equivalently be expressed as $H(z) = \frac{g}{\prod_{i=1}^{K}(1-p_i z^{-1})}$, where p_is are the *poles* of the filter that represent its resonant frequencies, and can be derived from the AR parameters a_k. The spectrum of the signal at any frequency f can be read off from $H(z)$ as $H(e^{j2\pi f})$.

Traditional AR analysis applies low-order (low values of K) analysis to estimate the spectrum of a signal – typically using an order of 10-20 for an analysis window of 250-400 samples. For our problem, however, we must consider the fact that the signal has been upsampled. In theory, AR models are only applicable when the spectrum has no zeros. The upsampled signal does not actually have zeros at any frequency, and spectral artifacts corresponding to low-pass filtered aliased copies of the lower-frequency components persist; however this also means that many of the poles of the filter end up modelling the unnecessary high-frequency components of the signal. In order to compensate for this, we must therefore use a high-order AR model to capture the spectrum. We have found a high AR order of 128, corresponding to half the number of samples in a 1ms analysis window, to result in the best estimates of the spectrum in the 0-8000Hz band. A number of different techniques have been proposed in the literature to estimate AR parameters. We employ Burg's maximum entropy method [22], which has several theoretical guarantees, and was also observed to result in the most reliable spectral estimates.

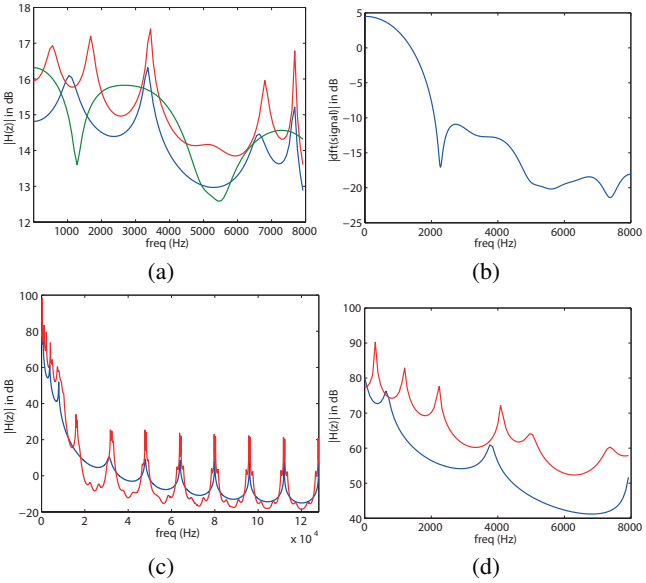

Fig. 4. **(a)** The spectrum of a 1ms segment of a sum of three cosines at 800,1600 and 3200Hz. The signal was upsampled from 16000 samples/sec to 256k samples/sec. The green curve shows the discrete Fourier transform spectrum, the blue curve is a 20th order Burg estimate, and the red curve is a 128th order Burg estimate. **(b)** The discrete Fourier transform spectrum of 1ms of voiced speech after it has been upsampled to 256000 samples/second from 16000 samples/second. Only the frequency range 0-8000Hz is shown. **(c)** The complete spectrum (from 0-128kHz) obtained through 128-order Burg analysis. The blue curve shows the 20th order Burg spectrum, and the red curve is the 128th order Burg spectrum. Both estimators are seen to use several poles to model low-energy high-frequency components. **(d)** A zoom-in of the Burg spectrum between 0 and 8kHz obtained from the upsampled signal. The 128th order Burg spectrum captures more structure.

Figure 4 shows spectra obtained with 20 and 128-order AR models. In the first example (a) we attempt to estimate the spectrum of a mixture of three sinusoids embedded in low levels of white noise, which has been upsampled from 16000Hz to 256 kHz. We observe that the 128-order model is able to capture the low frequency peaks accurately while the low (20) order models fails completely. The second example (b-d) shows spectra obtained for a 1ms window of

upsampled speech using the two models. Again, we observe the 128-order model to result in significantly more detail in the spectrum.

Thus, in summary, we employ the following analysis: for short analysis windows (of less than 5ms), we compute an 128-order AR model using the Burg method. We finally compute a 64-point log spectrum spanning a uniformly spaced range of frequencies from 20Hz-6400Hz from the obtained AR model. For longer analysis windows we compute conventional Mel-frequency cepstra using 64 analysis filters. All representations is subsequently reduced to 20 dimensions via principal component analysis.

4. PREDICTING PHYIOMETRIC PARAMETERS

We utilize the proposed spectral characterizations to predict the age and the height of the individual. We use a simple bag-of-words representation described in [23] together with random forest regression [24] to make our predictions.

We assume a corpus of training recordings, for which the age and height of the speaker are known.

The individual utterances that we work with are of different lengths and hence result in differing numbers of feature vectors. In a first step, we convert these variable-length recordings into a fixed-length feature representation that we can use to perform regressions. To do so, we train a 1024-component Gaussian mixture model with the collection of feature vectors obtained from the entire training data. Subsequently, in order to convert any recording to a fixed-length representation, we compute a soft assignment of each of the feature vectors in the recording to the 1024 Gaussians. Each vector x in a recording X thus contributes a *count* $P(k|x)$ (where k represents the Gaussian index) to each of the K Gaussians in the distribution. The recording is then converted to a 1024-dimensional representation, where the k^{th} component of this representation is given by $\sum_{x \in X} P(k|x)$. Every recording in the training corpus is converted to a fixed 1024-dimensional representation in this manner, as is each test recording.

The collection of training recordings, and the corresponding labels (age or height) are then used to train a random forest regression. We employed forests with 100 trees. Increasing the number of trees did not affect prediction. The dependent parameter (age or height) of the test recordings are subsequently computed using the trained model.

When multiple predictions, obtained for instance from different features, were combined, we performed the combination by simple averaging of the predicted values.

5. EXPERIMENTS

We conducted a series of experiment on the TIMIT database [25]. The TIMIT database comprises recordings from 630 speakers, each of who has spoken ten phonetically balanced sentences. Of these 462 speakers have been designated as the training set of speakers, and 168 have been designated as the test set. The training set, in turn, comprises 136 female speakers and 326 male speakers. The test set includes 112 male speakers and 56 female speakers. Thus, the training set comprises a total of 4620 recordings, 3260 by male speakers and 1360 by female speakers. The test set comprises 1680 utterances, 1120 by male speakers and 560 by female speakers.

The age and height of each of the speakers has been recorded. The age and height ranges of the data are provided in Table 1. Also given are the mean and standard deviation of each set.

In our experiments we consider a "default" predictor as one that predicts a test set parameter (age or height) as the average value of that parameter over the training set. The prediction of the default predictor is the *a priori* estimate of the parameter in the absence of any speech. A predictor is useless unless it produces significantly lower error than this default. Table 2 shows the root mean squared error (RMSE) and the mean absolute error (MAE) obtained with the default predictor on both the male and female subsets of the data. All predictions for male speakers are based on the male component of the training set, while predictions for female speakers are based on the female portion of the training set.

	Male				Female			
	Min	Max	μ	σ	Min	Max	μ	σ
age								
Train	21	76	31.0	7.4	21	86	30.7	9.7
Test	23	65	31.5	8.1	23	69	31.2	9.1
ht								
Train	157	198	180	7.1	145	183	165	6.8
Test	163	203	179	7.0	152	180	167	6.4

Table 1. Age and height statistics for the training and test set of TIMIT. Units for height are cm.

	Age		Height	
	RMSE	MAE	RMSE	MAE
Male	8.1	5.7	7.0	5.3
Female	9.1	6.2	6.5	5.2

Table 2. Root mean squared error and mean absolute error on the test set using the default predictor.

The challenges in predicting physiometric parameters from voice can be gauged from the results in the literature. In predicting height, the MAE of the default classifiers are, in fact, statistically indistinguishable, and sometimes *better* than the *best* prediction error reported by Mporas [11], Ganchev [3], and Williams [12]among others, and all but the best results reported in [15] and Hansen [13]. In general, these results could have been improved or matched simply by using a default predictor that predicts the mean height of the population for all subjects.

There are no published results that beat the default predictor for age on this data set to the best of our knowledge.

It is this rather discouraging state of affairs that we compare our techniques to.

In our first experiment we attempted to predict the height of the speaker from measurements derived from their speech. Predictions were made for each utterance by the speaker. We evaluated a number of different analysis window sizes. To evaluate analysis window sizes that were comparable to pitch periods, window sizes of 1ms, 2ms and 4ms were tested. The conventional 25ms window was also tested. Finally, an analysis window size of 100ms was also tested to evaluate the hypothesis that using *longer* analysis windows may also be useful.

Table 3 shows the results obtained using these analysis windows. Separate predictions were made for male and female subjects, since it was assumed that the gender of a speaker was known, or could otherwise be determined *a priori* accurately. The table shows the results obtained with the different analysis windows. It also shows results obtained by *fusing* the best four results in each case.

We note at the outset that the best results reported in Table 3 are better than a baseline classifier. However, we must also note that a paired t-test found this difference to be significant only at the

	Male		Female	
	RMSE	MAE	RMSE	MAE
1ms	6.8	5.2	6.2	5.0
2ms	6.7	5.1	6.4	5.1
4ms	6.7	5.1	6.5	5.1
25ms	6.9	5.2	6.4	5.1
100ms	6.9	5.2	6.3	5.1
Fusion	6.7	5.0	6.1	5.0

Table 3. Prediction of height

$p = 0.1$ level for both genders. This is nevertheless better than several reported results on this set, where the reported results are often worse than a baseline classifier.

Specifically, in the context of our hypothesis that smaller analysis windows may provide us a benefit by being able to "look" beyond the speaker's glottis, we find we cannot discard this hypothesis. For both male and female speakers the results obtained with smaller analysis windows are comparable to those obtained with the larger windows, and are in fact marginally better, although the difference will not hold up to rigorous statistical significance tests. We may infer that in the worst case, the additional frequency resolution from the longer analysis windows of 25ms does not provide any significant benefit for detecting speaker heights, at least within our framework.

A somewhat different conclusion is drawn when we consider the *fused* results. Fusion of results from low and high-resolution analysis windows results in a significant improvement in prediction accuracy (at the $p = 0.1$ level) compared to both the conventional analysis window of 25ms or a fusion of the results from the 25ms and 100ms windows. We infer from this that the smaller analysis windows do provide information that is *complementary* to conventional analyses.

In a second experiment we tried to predict the *age* of the speaker using the same analysis windows. Once again, we report results with analysis windows of 1,2,4, 25 and 100ms, as well as the fusion of the best four. These results are reported in Table 4.

	Male		Female	
	RMSE	MAE	RMSE	MAE
1ms	7.7	5.4	9.1	6.8
2ms	7.8	5.5	9.2	7.0
4ms	7.9	5.6	9.1	6.8
25ms	8.0	5.7	8.8	6.1
100ms	8.1	6.2	8.6	6.0
Fusion	7.8	5.5	8.9	6.5

Table 4. Prediction of age

Unfortunately, prediction of age remains a difficult task and our results do not significantly support the idea that this may be achievable using simple spectral characterizations of the speech signal. While we *do* observe some improvements over a default predictor, the results do not actually hold up to statistical significance tests, even at the $p = 0.1$ level. Nevertheless, we venture to make some statements about the *patterns* observed. Different trends are observed for male and female subjects. The prediction error for male subjects is observed to decrease with decreasing analysis window size. The best results are obtained with windows of 1ms (and, statistically speaking, the most significant differences from the baseline classifier were obtained with the 1ms analysis windows). In the case of the female subjects the trend was reversed. The best prediction was obtained with the largest analysis window. We speculate that this difference may be because male pitch is lower, and the smallest analysis windows do in fact permit us to "peer" into the sub-glottal vocal tract. The pitch periods for females are too small for useful measurements of age-related measurements from the sub-glottal region – the extra information to be derived by registering signatures from the sub-glottal regions is cancelled out by the loss of frequency resolution due to the very small analysis windows.

6. CONCLUSIONS

We can conclude from the above experiments that high-temporal-resolution spectral analysis using short analysis windows that are short enough to analyze open-glottis regions of the spectrum do appear to provide useful input, at least for the prediction of height. However, our analysis remains incomplete. While the mechanisms for deriving spectral features from larger analysis windows are well studied, it is unclear how best to effectively derive spectral detail from the smaller windows. In any case, we are eventually limited by time-frequency uncertainty. Our analysis windows are fixed in size in each experiment. Both, the results and our own explanations seem to make a case for *pitch-synchronous* analysis with analysis windows which track pitch period; this has not been evaluated.

Also, the literature clarifies that better prediction may be obtained by varying the regression models that make the actual predictions; this direction remains to be explored.

The prediction of age remains a more challenging problem. It is unclear that the trends we observe vis-a-vis male vs. female subjects are meaningful in any way. In general, our results are discouraging.

We also note that our results are borderline in terms of the statistical significance of the improvements over a baseline population-mean predictor. In the case of age no improvements are observed whatsoever. Curiously, we also do note that a significant amount of literature has reported similar or even worse results, without actually performing the comparison to the baseline classifier in what may be an instance of collective oversight.

It is our hope, and indeed expectation that the situation will improve and better results with greater statistical significance may be obtained with larger training and test data. Age and height information are available for several large publicly available corpora. We are currently investigating this avenue.

7. REFERENCES

[1] Steve An Xue and Grace Jianping Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study," *Journal of Speech, Language, and Hearing Research*, vol. 46, no. 3, pp. 689–701, 2003.

[2] Peter B Mueller, "The aging voice.," in *Seminars in speech and language*, 1997, vol. 18:2, pp. 159–68.

[3] Todor Ganchev, Iosif Mporas, and Nikos Fakotakis, "Audio features selection for automatic height estimation from speech," in *Artificial Intelligence: Theories, Models and Applications*, pp. 81–90. Springer, 2010.

[4] Iosif Mporas and Todor Ganchev, "Estimation of unknown speaker's height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.

[5] Julio Gonzalez, "Estimation of speakers'weight and height from speech: A re-analysis of data from multiple studies by lass and colleagues," *Perceptual and motor skills*, vol. 96, no. 1, pp. 297–304, 2003.

[6] Wim A van Dommelen and Bente H Moxness, "Acoustic parameters in speaker height and weight identification: sex-specific behaviour," *Language and speech*, vol. 38, no. 3, pp. 267–287, 1995.

[7] Robert M. Krauss, Robin Freyberg, and Ezequiel Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, pp. 618–625, 2002.

[8] Peter Lloyd, "Pitch (f0) and formant profiles of human vowels and vowel-like baboon grunts: the role of vocalizer body size and voice-acoustic allometry," *Journal of the Acoustic Society of America*, vol. 117:2, pp. 944–955, 2005.

[9] W Tecumseh Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in Rhesus macaques," *The Journal of the Acoustical Society of America*, vol. 102, no. 2, pp. 1213–1222, 1997.

[10] Drew Rendall, John R Vokey, and Christie Nemeth, "Lifting the curtain on the Wizard of Oz: biased voice-based impressions of speaker size," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 5, pp. 1208, 2007.

[11] Iosif Mporas and Todor Ganchev, "Estimation of unknown speakers height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.

[12] Keri A Williams and John H L Hansen, "Speaker height estimation combining GMM and linear regression subsystems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 7552–7556.

[13] John H L Hansen, Keri Williams, and Hynek Bořil, "Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 1052–1067, 2015.

[14] Amir Hossein Poorjam, Mohamad Hasan Bahari, and Vasileios Vasilakakis et. al., "Height estimation from speech signals using i-vectors and least-squares support vector regression," *Proceedings TSP 2014*, pp. 1–5, 2014.

[15] Harish Arsikere, Steven M Lulich, and Abeer Alwan, "Estimating speaker height and subglottal resonances using MFCCs and GMMs," *Signal Processing Letters, IEEE*, vol. 21, no. 2, pp. 159–162, 2014.

[16] Harish Arsikere, Gary K F Leung, Steven M Lulich, and Abeer Alwan, "Automatic height estimation using the second subglottal resonance," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3989–3992.

[17] Susanne Schötz, "Automatic prediction of speaker age using CART," *Working Papers, Lund University, Dept. of Linguistics and Phonetics*, vol. 51, 2005.

[18] Christian A Müller and Felix Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *INTERSPEECH*, 2007, pp. 2277–2280.

[19] Ming Li, Kyu J Han, and Shrikanth Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, 2013.

[20] Mohamad Hasan Bahari, ML McLaren, DA van Leeuwen, et al., "Age estimation from telephone speech using i-vectors," in *Proc. Interspeech*. 2012, Portland.

[21] Steven M Kay and Stanley Lawrence Marple Jr, "Spectrum analysis – a modern perspective," *Proceedings of the IEEE*, vol. 69, no. 11, pp. 1380–1419, 1981.

[22] John Parker Burg, "A new analysis technique for time series data," *NATO advanced study institute on signal processing with emphasis on underwater acoustics*, vol. 1, 1968.

[23] Anurag Kumar, Rita Singh, and Bhiksha Raj, "Detecting sound objects in audio recordings," in *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*. IEEE, 2014, pp. 905–909.

[24] Andy Liaw and Matthew Wiener, "Classification and regression by random forest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[25] Linguistic Data Consortium, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," `https://catalog.ldc.upenn.edu/LDC93S1`, 1993.