Estimation of children's physical characteristics from their voices

Jill Fain Lehman, Rita Singh

Disney Research, Pittsburgh, USA

jill.lehman@disneyresearch.com, rsingh@cs.cmu.edu

Abstract

To date, multiple strategies have been proposed for the estimation of speakers' physical parameters such as height, weight, age, gender etc. from their voices. These employ various types of feature measurements in conjunction with different regression and classification mechanisms. While some are quite effective for adults, they are not so for children's voices. This is presumably because in children, the relationship between voice and physical parameters is relatively more complex. The vocal tracts of adults, and the processes that accompany speech production, are fully mature and do not undergo changes within small age differentials. In children, however, these factors change continuously with age, causing variations in style, content, enunciation, rate and quality of their speech. Strategies for the estimation of children's physical parameters from their voice must take this variability into account. In this paper, using different formant-related measurements as exemplary analysis features generated within articulatory-phonetic guidelines, we demonstrate the nonlinear relationships of children's physical parameters to their voice. We also show how such analysis can help us focus on the specific sounds that relate well to each parameter, which can be useful in obtaining more accurate estimates of the physical parameters.

Index Terms: children's voices, age prediction, height, weight, gender estimation, physical characteristics, voice biometrics

1. Introduction

This paper addresses the problem of estimating children's physical parameters – specifically height, weight, age and gender – from their voices. There is a growing body of scientific literature on such anthropometric deductions from voice, e.g. [1, 2, 3, 4, 5, 6]. However, most of these studies have been almost exclusively carried out on adult voices. Children's voices have only been studied in relation to their age and various defects (such as stuttering) and medical pathologies, but not in relation to such deductions. The reasons may be threefold. The first is that anthropometry [7] itself has had very specialized use thus far in fields such as archeology, forensics etc., wherein it has particular relevance in the reconstruction of the body structure of the individual from partial evidence. Lately, there has been a renewed interest on making anthropometric deductions from voice, particularly in aid of automated applications that could benefit from being able to dynamically adjust to the physical dimensions of different users. The second reason could be the paucity of good data due to legal and societal difficulties in obtaining children's speech samples and their physical measurements for study. The third reason is presumably the inadequacy of small amounts of children's data to represent the full range of voice-related phenomena exhibited within this population. Since children's voice changes very rapidly with age, more data and more equitable distribution of data across ages is required to enable studies that could be regarded as statistically significant or conclusive.

There are other issues with existing studies that restrict their applicability to children's voices. Most prior studies assume or derive *linear* relationships between the speaker parameters and the features derived from speech signals. In cases where they do model the non-linear relationships using parametric or non-parametric models, such as [8, 9], they use aggregate statistics derived from features computed over the entire speech signal, without regard to the types of sounds in the signal [10]. Others use human judgment for analysis [11, 12]. Many studies focus on the correlations of physical parameters of the speaker with certain categories of sounds in speech – such as vowels [13], but these studies do not cover a comprehensive analysis of all the sounds in speech.

In our paper, we analyze the effects of children's physical parameters on *all* the compositional units of speech. We devise a strategy based on articulatory-phonetic principles and non-parametric regression models to study these relations in order to understand the nature of these relationships. Our articulatory-phonetic approach is predicated on the broad assumption that since the vocal tract is part of the extremely complex human biological system, every factor that affects this system also affects the parameters of the vocal tract, influencing the synergy, extent and rate of the physical movements of its articulators.

Our studies are conducted on a phonetically rich database of children's voices collected in-house by Disney Research, where children were asked to repeat sentences spoken by an adult prompter. The recordings were made in a studio environment using high quality far-field microphones, since children do not like to wear close-talking microphones and will often tamper with them, causing recording disturbances. Their biometric parameters were measured on-site prior to their recording sessions. We call this database the *Copycat* database. For our study, we derive formant measurements from speech of children from 4-10 years of age, since they are known to correlate reasonably well in adults with their height, weight, age etc. Our experiments on Copycat bring out interesting relationships between children's height, weight, age and gender, and their voice. The relationships appear to be complex, as expected, often specific to the particular dynamics of the vocal tract as represented by different categories of articulatory-phonetic units that constitute speech. Also, as expected, we see that certain subsets of articulator configurations are indeed affected by different physical characteristics more than others, while others remain justifiably (from computational and acoustic-phonetic speech production viewpoints) unaffected. Moreover, these subsets are different for different types of physical parameters. We determine that even within individual utterances, the correct articulatory-phonetic units are frequently sufficient to predict the child's physical measurements. Furthermore, predictions made by combining evidence from the correct group

of phonemes can be significantly superior to predictions made from aggregate characterizations of the signal that do not consider phonetic distinctions.

2. Analysis strategy

There are five components that play a pivotal role in our strategy. These are 1) the compositional acoustic units of speech, 2) the features that capture their key characteristics and help disambiguate between them, 3) the statistical models that we use to learn these effects quantitatively, 4) the manner in which we measure the features, and 5) the manner in which we use the outcomes of the modeling process. In this section we describe each of these aspects briefly.

2.1. Phonemes and articulatory-phonetic categorizations

In articulatory-phonetics, sounds with consistent spectral patterns are recognized as the basic compositional units of speech, or *phonemes*. Each phoneme is produced by the modulation of sound through a specific set of articulator configurations in the vocal tract [14]. Words in any language are produced by enunciating specific sequences of phonemes.

In English, phonemes are divided into two broad categories – vowels and consonants, and two other categories that involve some intersections of these – semivowels and diphthongs. Based on the manner of articulation and voicing (i.e. whether or not the vocal folds vibrate in the production of the sound), the consonant categories that we consider in this work are 1) Plosives: B D G (voiced) P T K (unvoiced), 2) Affricates: V DH Z ZH (voiced) F TH S SH HH (unvoiced), 3) Fricatives: JH (voiced) CH (unvoiced), 4) Nasals: M N NG (voiced), 4) Liquids: L R (voiced), and 5) Glides: Y W (voiced). For vowels, the categorizations are given more explicitly in Fig. 1(a). The list shown is narrowed to those typical of North American English only. Note that we use a non-conventional notation for phonemes - they are denoted in uppercase or lowercase symbols as required to maximize clarity of presentation. Vowels are termed *high*, *mid* or *low* depending on the position of the jaw, and *front*, *middle* and *back* depending on the location of articulation in the mouth. *Tense* and *lax* vowels differ in the amounts of stress placed on the articulators, while *rounded* vowels involve some rounding of the lips.

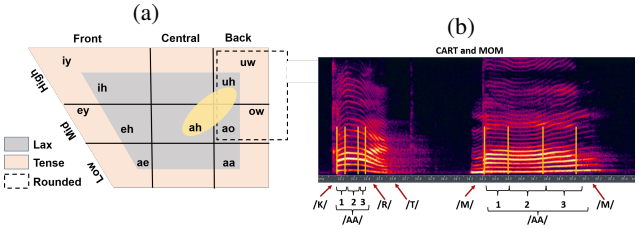

Figure 1: (a) Vowels in North American English, as represented in the CMU Sphinx ASR system. (b) HMM-generated segments of the phoneme AA in two words.

2.2. Formants and formant-related features

In the production of different sounds, the human vocal tract is shaped into different configurations by the movement of the articulators. *Formants* are the resonant frequencies of each of these configurations. In this paper we choose to illustrate our strategy for the estimation of physical parameters through formants and formant-related measurements. Prior studies have shown that these can be extremely useful in understanding the manner in which different physical parameters that affect voices of adult humans, e.g. [15, 16, 17, 18, 19, 20], or even animals [21]. We describe the different formant-related measurements that we use in this study below:

2.2.1. Formant position

This is merely the peak frequency of a formant. The formants are numbered by convention – the formant with the lowest frequency is called $F1$, the second lowest frequency formant is $F2$, the next is $F3$ and so on. Up to five formants ($F1-F5$) are typically observable in the spectrograms of children's speech.

2.2.2. Formant bandwidth

Formant bandwidth is defined as spread of frequencies around any formant within which the spectral energy remains within 3db of the formant's peak energy. While formant bandwidths are not known to play a role in disambiguating phonemes, they carry information about the speaker's vocal tract *composition*, such as the elasticity of the walls, energy dissipation through the glottis etc., and are correlated to the specific vocal tract configurations that produce phonemes [22]. In general, higher formants have greater bandwidths.

2.2.3. Formant-Q

The Q-factor of a filter is defined as the ratio of the peak frequency of the filter to its bandwidth. In the source-filter representation of the vocal tract [23], the formants are considered to be the peak filter frequencies, and the formant-Q is defined as the ratio of a formant frequency to its bandwidth. Formant-Q's are also thought to be dependent on the speaker characteristics [16], since they reflect the frequency dependent characteristics of the speaker's vocal tract.

2.2.4. Formant dispersion

Formant dispersion is defined as the average spacing between the formants. It is thought to be indicative of the vocal tract length of the speaker [24]. The conventional definition of formant dispersion is the arithmetic average of the spacing between phonemes. However, this merely captures the spacing between the highest and lowest formant. In this paper we use a modified version of Formant dispersion, as suggested in [16] $D = \sqrt[n-1]{\prod_i F_i - F_{i-1}}$, which is the *geometric* mean of the formant spacings.

2.3. Extracting accurate feature measurements

The extraction of accurate measurements from within the boundaries of a phoneme can be very difficult in continuous speech, where the boundaries of phonemes are not clear even to humans. In additions, the measurements may not be consistent due to co-articulation effects. According to the widely accepted *locus theory* of co-articulation, each distinct phoneme has a *locus*, which is an ideal configuration of the vocal tract necessary for its correct enunciation by the speaker. In continuous speech, as one phoneme leads into another, the vocal tract changes shape continuously, moving from the locus of one phoneme to another, often not achieving the target loci of successive phonemes at all. A consequence of this continuous variation is that formant patterns at the extremities of any phoneme vary by its adjacent phonemic context, and the degree of variability can be high. This is illustrated in Fig. 1(b). These

context-related variations of formant patterns can confuse analyses, and mask the relations between formant features and the speaker's physical parameters. In order to minimize this confusion, we therefore take all formant related measurements from the *central* segments of each phoneme, since these are relatively less affected by context, and are most representative of the locus of the given phoneme. These segments are automatically generated by a state-of-art automatic speech recognition (ASR) system trained specifically for generating accurate word, phoneme and state-level segmentations. Here the term *state* refers to the states of a Hidden Markov Models (HMMs) used in the ASR system. In our work, we train 3-state Bakis Topology HMMs, and use the segmentations corresponding to the central state only to measure our features. Our feature (formant) measurements are derived from LPC spectral analysis of the speech signal using Burg's method [25].

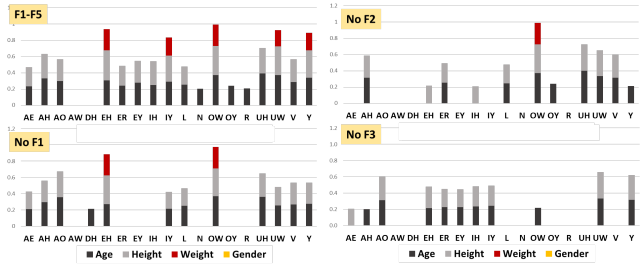

Figure 2: Relationship of formant position measurements to physical parameters.

2.4. Linear vs. Non-parametric models for analysis

We do not expect the children's physical characteristics to be linearly related to acoustic features. Hence linear regression models, and the direct correlations and R^2 values of features, that capture linear relationships between predictor and dependent variables, may be unsuitable for our purpose. We therefore use an alternate strategy to quantify these relationships. For each physical characteristic, we train a non-parametric regression model by phoneme. We quantify the relationship between the acoustic features and the physical parameter through the *correlation* between the predictions made by the model and true value of the parameter. For our experiments we use Random Forest (RF) regression [26], which has been shown to be effective for such studies [27].

2.5. Phoneme selection and statistical significance

The statistical significance of the computed correlations between predicted and actual values is determined using a t-test [28]. Our data occur in groups. When all instances of a phoneme individually predict the same speaker parameter, the predictions cluster together more closely than those for other speakers. In this respect, the data do not exactly correspond to the assumptions made by the t-test, and the P values reported by the t-test may be optimistic. To compensate for this, we use a conservative P value threshold of 0.001 to report results. Thus, all reported correlations, even if low, have high confidence.

3. Experimental results

Our experimental data comprise recordings from 26 children from the 2015 collection of Copycat. The data were hand-transcribed, with all speech and non-speech events, including breaths, coughs etc. marked. The CMU Sphinx ASR system [29], trained on more than 120 hours of children's speech, was used to extract the phonemes. To maximize accuracy, the system was acoustically adapted to each child's speech separately. The resulting phoneme segmentations were manually checked and found to be extremely accurate.

Each experiment evaluated one of the physical parameters: age, height, weight or gender. In each experiment, we trained and tested a 100-tree RF-based parameter predictor using the formant features derived from all examples of a single phoneme. Following this, an 8-fold cross-validation experiment was performed. There was no overlap of speakers between the training and test partitions in each fold. All parameters were evaluated against all phonemes in this manner. The phoneme set comprised 47 phonemes including silence, of which 24 were consonants and 7 were filler sounds such as laughter, breath etc.

3.0.1. The effect of physical parameters on formant positions

The focal point of our strategy is to identify the effect of physical parameters on formant measurements under different articulator configurations. As a baseline, we make our predictions using all formants $F1$-$F5$. To isolate the effect of individual formants, we compare this with the predictions obtained by leaving one formant out. The difference between the two outcomes represents the contribution of the left-out formant. Fig. 2 shows the results from these cases. We do not show results of isolating F4 and F5 since, for children's speech, these are frequently too high to be measured accurately or just absent.

The following phonemes did not show any correlations with any of the physical parameters: AA AY B CH D F G HH JH M NG P S SH T TH W Z ZH. Note that since Copycat is a phonetically rich database, all phonemes have more than 2000 instances in it, some have over 17000 instances. Fig. 2 shows the correlations for the remaining phones. The patterns observed validate several observations in the literature. Some examples are:

1. Among vowel sounds, the ability to predict age reduces or disappears when formants F1-F3 are not considered. This is because when learning to disambiguate sounds with progression of age, the child's emphasis is expected to be on F1-F3, which are well known to play pivotal roles in disambiguating sounds. When we ignore these specific formants, correlations disappear. Similar effects are also seen in a few other sounds that may have similar undiscovered trends.
2. The dimensions of the nasal cavity do not change during articulation, only the opening to its passageway is affected. Opening of the passagway results in anti-resonances which can cancel out some of the formants. We see that nasal sounds other than N are absent from Fig. 2. Interestingly, we found that in the case of nasals, F1-F4 are *jointly* needed to even find a correlation with age.
3. It is known that formants are often not clearly discernible in fricative sounds such as CH, JH, HH etc. due to the turbulent nature of these sounds. They fail to appear as significant in our analysis for relations with physical parameters as well.
4. No phoneme predicts gender, confirming several studies that there is no significant difference between the voices of male and female pre-pubertal children.
5. Formants correlate with height as expected. Taller children have longer vocal tracts, and hence lower formants [30].
6. Plosive sounds do not show up in our plots. This is explained by the fact that their central regions are usually in the low-energy transitions between the stop and release phases that

define the complete plosive, where the formant peaks are weak or nonexistent.

Several other such observations may be derived from the charts above. However, perhaps the most important observation is that individual instances of many of these phonemes, which are often just a fraction of a second long, are able to predict physical parameters of the speakers.

3.0.2. *Relations with formant bandwidths and formant-Q*

Fig. 3(a) and (b) summarizes the results obtained with the formant bandwidths and formant-Q respectively. These were not found to be correlated to weight and gender. In both cases, we find that these measurements are predictive of age and height for only a small selection of *voiced* phonemes, which have high-energy spectral peaks. The physiological interpretation of these measurements is unclear.

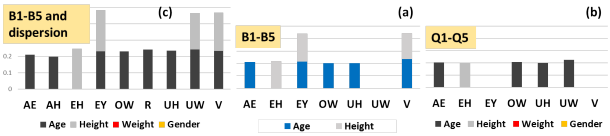

Figure 3: Relation of formant bandwidths, -Q and dispersion (D) to physical parameters (a) Formant bandwidths – when we remove B2 and B3, all correlations vanish (graphs not shown). (b) Formant-Q – correlations vanish again when we remove Q2 and Q3. (c) Formant Bandwidths B1-B5 and dispersion.

3.0.3. *The effect of physical parameters on formant dispersion*

In our study we evaluated the relation of formant dispersion to children's body parameters by adding it as an extra feature along with bandwidths B1-B5. The difference between the performance obtained with B1-B5 alone and B1-B5+dispersion indicates the contribution of dispersion to the prediction of the parameters. This difference can be noted by comparing Fig. 3(c) with Fig. 3(a). As in the case of bandwidths and -Q, formant dispersions are most informative in vowel sounds. For these sounds we note that dispersion carries significant information about physical parameters of children as well. This brings us a long way from some earlier studies, e.g. [24].

3.0.4. *Aggregate and utterance-level predictions*

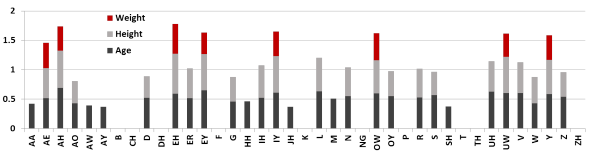

Figure 4: Correlations within aggregates of articulatory-phonetic units across speakers.

The previous results reported statistical correlations between true and predicted parameter values, when predictions were made with *individual* phonetic segments. As such, they allow us to evaluate the information encoded in each phoneme. However, different phonemes occur with different frequencies. Also, the prediction errors made by different instances of a phoneme may exhibit different degrees of correlation among themselves. Thus, a better picture of the information carried by a phoneme is obtained by considering the statistical aggregate across collections of data. Fig. 4 shows the correlation between predicted and true parameter values, when a *single* prediction was obtained for each child (for each phoneme) by averaging the predictions made by the individual instances of the phonemes. All features F1-F5, B1-B5, Q1-Q5 and D were used. Only correlations greater than 0.35 are shown. We note that predictions are much more accurate, and correlations greater than 0.6 and as high as 0.7 are obtained for several phonemes, both for height and age. Weight is, in general less easy to predict, but is nevertheless predictable.

One of our objectives in this paper was to identify the specific articulatory phonetic units that would enable accurate prediction of children's physical parameters. Table 1 shows the mean absolute error in predicting age, height and weight using three classifiers. The first was obtained using a single i-vector feature derived from all the recordings for a child [9], and predicting the feature using the i-vector. Predictions were made using SVM regression. This procedure produces state-of-art results for anthropometric prediction in adults [10]. The second and third predictors were based on formant measurements. Our second predictor averaged predictions obtained from individual phonemes in the recordings for any child. The third predictor used only phonemes which resulted in a correlation greater than 0.4 in fig. 4. We note that segmenting the speech signal into phonemes and averaging the individual predictions from them results in significant improvement in prediction over simply using all the speech. But utilizing only the most informative phonemes results in the best performance, as expected.

	Age (yrs)	Height (cm)	Weight (kg)
Global	1.24	3.57	7.5
All phonemes	1.10	3.44	7.35
Best phonemes	0.95	3.32	7.24

Table 1: Mean average error in the prediction of age, height and weight using different ensemble aggregations.

4. Conclusions

We note that as originally hypothesized, focusing on the specific articulatory-phonetic units that are most influenced by the parameters results in the best performance. Weight is, in general, much harder to predict than age or height. However, we note that these parameters are not independent. It is known that in children, age is correlated with height and weight (and must influence their expression in voice as well). The correlation between age and height in this set was 0.86, between age and weight was 0.67. So any predictor that predicts age well is likely to also predict height well, and vice-versa. Of greater importance is the fact that the results support our original hypothesis – that it is important to consider the individual aspects of the speech production mechanism and the speech signal itself, in predicting physical parameters, and that ensemble characterizations that ignore these distinctions can be less effective. In children, age also correlates with *macro* characteristics within sentences or sentence-level segments of children's speech, such as prosody, cadence, loudness, rate etc. Statistical features such as i-vectors may be useful to capture these effects. Finally, we note that in this paper, we have used formant measurements for illustrative purposes. Several other types of features that capture different signal characteristics at subphonetic levels can be readily used in the same manner.

5. References

[1] T. Ganchev, I. Mporas, and N. Fakotakis, "Automatic height estimation from speech in real-world setup," in *Proc. of the 18th European Signal Processing Conf*, 2010, pp. 800–804.

[2] I. Mporas and T. Ganchev, "Estimation of unknown speaker's height from speech," *International Journal of Speech Technology*, vol. 12, no. 4, pp. 149–160, 2009.

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

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

[5] K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. M. O'Connor, S. Röder, P. W. Andrews, B. Fink, L. M. DeBruine, B. C. Jones, and D. R. Feinberg, "Vocal indicators of body size in men and women: a meta-analysis," *Animal Behaviour*, vol. 95, pp. 89–99, 2014.

[6] K. Pisanski, P. J. Fraccaro, C. C. Tigue, J. J. M. O'Connor, and D. R. Feinberg, "Return to oz: Voice pitch facilitates assessments of men's body size," *Journal of Experimental Psychology: Human Perception and Performance*, June 2014.

[7] "Anthropometry (n.d.)," *Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health*, vol. Seventh Edition, 2003.

[8] M. H. Bahari, M. McLaren, D. A. van Leeuwen *et al.*, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

[9] M. Li, K. J. Han, and S. 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.

[10] T. Ganchev, I. Mporas, and N. Fakotakis, *Audio features selection for automatic height estimation from speech*. Springer, 2010, pp. 81–90.

[11] J. 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.

[12] N. J. Lass, D. T. Kelley, C. M. Cunningham, and K. J. Sheridan, "A comparative study of speaker height and weight identification from voiced and whispered speech," *Journal of Phonetics*, 1980.

[13] E. P.-M. Ma and A. L. Love, "Electroglottographic evaluation of age and gender effects during sustained phonation and connected speech," *Journal of voice*, vol. 24, no. 2, pp. 146–152, 2010.

[14] W. A. Smalley, *Manual of Articulatory Phonetics: Workbook Supplement*. Institute of Education Sciences (ERIC), USA, 1964.

[15] R. Greisbach, "Estimation of speaker height from formant frequencies," *International Journal of Speech Language and the Law*, vol. 6, no. 2, pp. 265–277, 2007.

[16] R. Singh, D. Gencaga, and B. Raj, "Formant manipulations in voice disguise by mimicry," in *4th International Workshop on Biometrics and Forensics (IWBF)*. Limassol, Cyprus: IEEE, 2016.

[17] D. A. Puts, C. L. Apicella, and R. A. Cárdenas, "Masculine voices signal men's threat potential in forager and industrial societies," *Proceedings of the Royal Society of London B: Biological Sciences*, p. rspb20110829, 2011.

[18] R. Greisbach, "Estimation of speaker height from formant frequencies," *International Journal of Speech Language and the Law*, vol. 6, no. 2, pp. 265–277, 2007.

[19] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 133–150, 1994.

[20] M. Iseli, Y.-L. Shue, and A. Alwan, "Age-and gender-dependent analysis of voice source characteristics," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[21] P. 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.

[22] G. Fant, "Formant bandwidth data," *Speech Transmission Laboratory Quarterly Progress and Status Report 2*, vol. 3, pp. 1–3, 1962.

[23] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 2733–2749, 2008.

[24] W. T. 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.

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

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

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

[28] R. Lowry, "Concepts and applications of inferential statistics," *Available at: www.vasserstats.net/textbook/*, 2013.

[29] "The cmu sphinx suite of speech recognition systems," http://cmusphinx.sourceforge.net/, 2013.

[30] R. E. Turner, T. C. Walters, J. J. M. Monaghan, and R. D. Patterson, "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," *The Journal of the Acoustical Society of America*, vol. 125, no. 4, pp. 2374–2386, 2009.