Forensic anthropometry from voice: an articulatory-phonetic approach

Rita Singh [†], Bhiksha Raj [‡], Deniz Gencaga [†‡]
[†] Computer Science Department, Carnegie Mellon University, Qatar
[†‡] Robotics Institute and [‡]Language Technologies Institute, Carnegie Mellon University, USA

***Abstract*—This paper addresses a problem that is of paramount importance in solving crimes wherein voice may be key evidence, or the *only* evidence: that of describing the perpetrator. The term *Forensic anthropometry from voice* refers to the deduction of the speaker's physical dimensions from voice. There are multiple studies in the literature that approach this problem in different ways, many of which depend on the availability of sufficient volumes of speech for analysis. However, in the case of many voice-based crimes, the voice evidence available may be limited. In such cases it is especially advantageous to regard the recorded signal as comprising multiple pieces of evidence. In this paper, we show how this can be done. We explain why, for any anthropometric measurement from speech, it makes sense to consider the contributions of each articulatory-phonetic unit independently of others, and to aggregate the deductions from them only in the aftermath. This approach is based on the hypothesis that the *relative* evidence given by different compositional units of speech can be more indicative of the anthropometric factor being deduced, than the evidence derived from the aggregate voice signal. We explain the applicability of this approach through experiments on standard speech databases.**

I. Introduction

Voice-based crimes comprise a significant fraction of crimes committed in the world today. Such crimes include those in which voice may be *key* evidence (such as a security camera video footage of a store robbery where the perpetrator may be masked but may demand something of the victims), and those in which voice is the *only* evidence. Examples of the latter include crimes committed over phone or internet, such as harassment, blackmail, threats, ransom demands in kidnappings, impersonation with intention to defraud in banking and other scenarios, voice-based phishing, hoax emergency calls, false reporting such as bomb threats in public areas, "swatting" calls to the police etc. All of these and more included in this category of crimes are often faced with the investigative challenge of finding the perpetrator(s) through the analysis of the voice evidence.

Forensic analysis of voice for this purpose benefits from a multitude of studies in different areas of science that study voice. These have demonstrated that the human voice carries a wealth of information about the speaker, including the speaker's physical characteristics such as height, weight, physiological characteristics and age, physical and mental state of health, social status, geographical origins etc., and a plethora of other information including that about their immediate physical surroundings. However, deriving such information is currently a goal that is challenged by many scientific problems. It is hinged on the understanding of the *signatures* of all of the speaker's personal characteristics and environmental parameters (at the time of recording) that are embedded in the speech signal, and using these to measure different characteristics of the speaker.

In this context, it is important to find out what must be done to identify the signatures alluded to above, e.g. what feature representations might best capture different signatures, what techniques might help identify them etc. Where our paper becomes relevant is that in addition to all of the above, it is also important to know *where* in the signal to look for such signatures, i.e. which parts of the signal are informative and which are not in the *expression* of the parameter in question and its signature. This may be tied to the actual feature representation(s) being used, but a framework is nevertheless needed to specify these informative locations in the signal. The goal of this paper is to provide such a framework.

In the forensic context, terms such as anthropometry (the measurement of body parameters), psychometry (the measurement of psychological parameters or state of mind), sociometry (the measurement of social parameters) etc. refer to the processes of deducing different categories of speaker characteristics that may help generate a reasonable description of the speaker and may thereby help locate him/her. To deduce these person-descriptive parameters, our framework comprises an approach based on considerations of the human speech production mechanism. In this paper we primarily present the reasoning behind, and evidence in support of, this *articulatory-phonetic* approach to anthropometry.

Although some studies have used information derived from phonemes for biometric applications, such as speaker matching [1], the choice of specific phonemes has largely been based on heuristic decisions. In contrast, the framework we present outlines a generic methodology based on well-established articulatory-phonetic guidelines, for the deduction of any person-descriptive parameter from voice. The key elements of our approach also involve a novel method for the sub-phonetic segmentation of speech in order to derive features that are compatible with this approach, and the demonstration of useful ways to visualize, interpret and utilize the information derived from the articulatory-phonetic categories. We build our arguments in favor of this approach through a brief review of the manner in which the human speech production process relates to the speaker's biometric parameters.

A. The speech production process and biometric parameters

The human vocal tract can be viewed as a system of dynamically configurable resonance chambers. *Voice* is the acoustic signal we hear when the mechanical vibrations of the vocal folds transform the aerodynamic energy of the air expelled from the lungs into acoustic energy in the form of sound waves. This *excitation* signal is further modulated into the sound patterns characteristic of speech by the physical movements of the vocal tract. The movements change the shape and dimensions of the various resonant chambers of the vocal tract, causing time-varying resonance patterns in the acoustic signal. This sequence of resonance patterns in the acoustic signal is perceived as (often) intelligible speech by the listener. Each distinct pattern, supported by the articulatory configuration of the vocal tract that produces it, is considered to be a unique compositional unit of speech, or a phoneme.

In continuous intelligible speech, the articulators are required to move continuously as the speaker forms words and sentences. During the production of continuous speech, the vocal tract attempts to "flow" from the canonical configuration for one phoneme to that of the next. The resonant characteristics of the phoneme-specific configurations are governed by the dynamics of the movement between different configurations, the degree to which the articulators achieve the canonical configuration for any phoneme, the excitation of the vocal tract, and all of the other articulatory and acoustic phenomena that affect the production of the phonemes.

All of these factors are known to be influenced by the speaker's physical and mental (biological) factors. Anthropometric characteristics such as skeletal proportions, race, height, body size etc. largely influence the voice by playing a role in the placement of the glottis, length of vocal cords, relative sizes and proportions of the resonance chambers in the vocal tract etc. When a speaker enunciates different phonemes, all of these structures act in concert, and the final speech signal produced carries the signatures of the specific vocal tract proportions, movements and configurations that the speaker is able to produce for each phoneme. Each phoneme therefore carries some evidence of all of these characteristics, except that the evidence is reasonably expected to be expressed differently for each phoneme.

The advantage of this reasoning is that it can be easily extended to apply to other categories of speaker characteristics, such as the speaker's mental state. Factors that relate to a person's mental state affect the *movement* and *locus* of the articulator configurations. This relationship is evident from several older studies that show that different mental states affect the body's muscle agility and response times, including that of the facial muscles, and by direct association, that of the articulators e.g [2], [3]. In one of his early expositions, Charles Darwin noted the relationship between emotion and specific patterns of muscle activity, particularly in the face [4]. Currently there is a large body of literature on skeletal muscle activity associated with psychological illnesses. Examples include muscle agility changes with anxiety and depression [5], with personality traits [6], etc. All of these effects are expected to carry over to the articulators. Following the same reasoning that we apply to a speaker's physical state, we expect different phonemes to also carry the signatures of the speaker's psychological state, and to express them differently from other phonemes.

Based on this reasoning we expect that estimates of a speaker's person-specific parameters may presumably be recovered more reliably from appropriate phoneme-specific analysis of the individual phonemes.

The rest of this paper is arranged as follows: In Section II we discuss some basic categorizations of speech from an articulatory-phonetic perspective. With this in context, in Section III we describe our approach for deriving anthropometric evidence from speech recordings. In Section IV we present experimental results in support of the proposed methodology. This is followed by conclusions in Section V.

II. A REVIEW OF PHONEME CATEGORIZATIONS

Based on the commonalities and differences between the articulator motions and configurations that produce them, articulatory phonetics differentiates speech into *phonemes*, its constituent compositional units, and further into several categories grouped by specific articulator locations and vocal fold activity. At the broadest level, phonemes are divided into *consonants*, which include some kind of airflow obstruction in the vocal tract, and *vowels*, which do not. These are briefly described below.

A. Articulatory-phonetic categorization of consonants

Depending on the voicing, place and manner of articulation, consonants are divided into several categories. These are named based on the key articulators involved. Fig. 4 (which also doubles as a template for representing results in the experimental section of this paper) lists these categories. Articulators that are considered in this categorization include the teeth, lips, hard palate, soft palate (velum), alveolar ridge, tongue (front, back or middle/sides, i.e. apex, dorsum and laminus respectively), uvula, glottis and pharynx. The list of phonemes in Fig. 4 is limited to those found in North American English, and also confined to the set of phonemes we analyze for the work presented in this paper. Consonants are further divided into two broad categories (not shown in the table). These are the *Obstruents*, which include all Stops, Affricates and Fricatives and are characterized by a significant constriction or obstruction in the vocal tract; and the *Sonorants or Approximants* which include the rest of the consonants, and are characterized by a slight constriction of the vocal tract. The key characteristics of the divisions named in Fig. 4 are described below.

Phonemes that involve the active vibration of the vocal cords are called *Voiced* phonemes, while those in which the vocal cords do not vibrate are termed as *Unvoiced*. For the purpose of this study, we focus on five categories of consonants based on the manner of articulation: namely *Plosives, Fricatives, Affricates, Nasals, Liquids and Glides*. The key characteristics

associated with each of these are: *Plosives*: complete stoppage of airflow, followed by sudden release of air; *Fricatives*: creation of turbulent airstream; *Affricates*: contain the characteristics of both plosives and fricatives; *Nasals*: complete airflow obstruction and release through the nose; *Liquids*: airflow along the sides or top of the tongue; *Glides*: stricture between the roof of the mouth and the tongue. If the stricture occurs such that air flows along the sides of the tongue, the glide is called a *Lateral* glide. If the sound is more "r"-like, the glide is called a *Rhotic* glide.

Depending on where in the vocal tract these key characteristics are generated (e.g the location of the airflow obstruction for a Plosive), the five categories above are further divided into the following subcategories. The articulators that are involved are indicated in parentheses: *Bilabial* (both lips), *Labiodental* (Lips and teeth), *Interdental* (upper and lower teeth), *Alveolar* (alveolar ridge), *Palatal* (hard and soft palate), *Velar* (Velum) and *Glottal* (glottis).

B. Articulatory-phonetic categorization of vowels

The primary difference between vowels and consonants is that in vowels there is no constriction of airflow in the vocal tract, whereas in consonants there is some degree of constriction somewhere in the vocal tract. Vowels are categorized based on their height (how high the tongue is), backness (which part of the vocal tract is pivotal in its production), laxness (rigidness or relaxedness of configuration) and roundedness (whether or not the lips are rounded). Fig. 5 in the experimental section of this paper shows the standard vowel categorization for North American English.

III. Proposed approach to anthropometry

Our methodology for the recovery of anthropometric attributes from voice is based on training a bank of phoneme-based predictors for each attribute measured, selecting a subset of them based on statistical criteria, and combining *their* decisions for a final prediction. Such "combination-of-predictors" approaches for the prediction of attributes are fairly standard in many different contexts in the machine learning literature, including multimedia processing [7] and audio processing [8]. What is novel about our approach is the utilization of articulatory-phonetic criteria to create the predictors in the mixture, and the specific mechanism for locating the right segments of speech for feature extraction. We describe this mechanism below.

A. A case for sub-phonetic *features*

Over the course of an utterance, spectral patterns vary continuously as the vocal tract transitions from the configuration for one phoneme to that for the next, often blurring the boundaries between them, resulting in a continuous, highly variable signal with complex, never-exactly-repeated spectral patterns. However, since the biophysical parameters of the speaker are manifested in *every* part of the speech signal, every section of these complex, never-repeated patterns is affected by the speaker's current biophysical state. Due to the complex nature of the speech signal itself, therefore, it is often difficult to distinguish between a *biophysically-affected* pattern seen in one phonetic context, and a naturally occurring pattern in another. As a result, signal measurements based on overall characterizations of the signal will often show weak, or no statistical relation to the speaker's state, although these relations may be locally evident in different portions of the signal.

In order to effectively characterize the expression of biophysical parameters on the speech signal, it therefore becomes necessary to focus on relatively *stable* sound structures that so typify the underlying phoneme that their absence or modification may change the perceived phoneme itself. Since the nature or state of the speaker also affects the articulation of these structures, the effect of the speaker's biophysical state/parameters on their expression can be isolated with relatively lower ambiguity than from other, more variable parts of the speech.

Such stable structures are generally phoneme-internal or *sub-phonetic* features. Candidate sub-phonetic features include voicing-onset time (VOT) [9], voicing offsets [10], onset of pitch, phonetic loci, etc. Indeed, each of these features is affected by different biophysical factors. For instance it is well known that VOT is affected by neurological disorders [11] and age [10], anomalies in onset of pitch are characteristic of vocal cord paralysis, formant positions in loci are related to age, body parameters [12], vocal tract shape [13] etc. The challenge, however, is that not all sub-phonetic features are affected by all biophysical factors. Possibly the most "universally" affected feature is the phonetic locus, which we briefly describe in the following subsection.

B. An HMM-based midstate-sharing technique using entropic regularizers for deriving stable sub-phonetic measures

As noted above, phoneme expression tends to be affected both by adjacent phonetic context [14] and longer-term prosodic and expressive trends in the speech signal. In order to isolate our measurements of the phonemes from the variability introduced by these contextual and longer-term effects, we must identify regions of each phoneme that are most invariant to context or longer-term trends.

The locus theory of phonemes states that every phoneme has a "locus", corresponding to a canonical arrangement of the vocal tract for that phoneme, and that the articulators move towards it in the production of the phoneme [15]. In continuous speech, the loci of phonemes may not be fully reached as the articulators move continuously from one set of configurations to another. Fig. 1 shows an example. While the locus theory does not explain all the variations observed in different instantiations of a phoneme, a key, valid insight that may be derived from it is that the interior regions of the phoneme that are representative of the locus are much more invariant to contextual and longer-term effects than the boundaries. The exact interior region that represents the locus may, however, vary with the instance of the phoneme. It need

not be at the center of the phoneme, and its position must be carefully estimated.

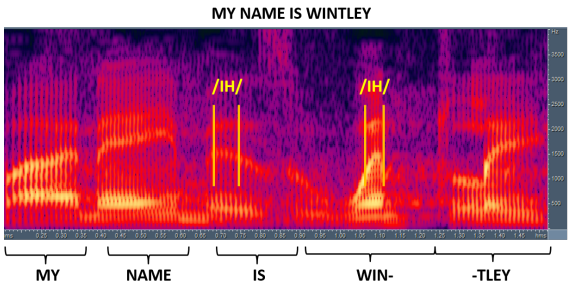

Fig. 1. The sentence "'My name is Wintley"' spoken by an adult male. Note the long-term spectral patterns that flow continuously across the entire utterance while moving towards different phoneme loci in the process. The phoneme IH is marked . It coccurs in two different phonemic contexts, and the influences can be seen at the beginning and end of the phoneme in each instance.

Since we wish to extract context-invariant features from the phoneme, we must identify this central "locus" segment and extract features from it. However, since the actual position and duration of this segment can vary significantly within each instance of a phoneme, actually identifying this segment is not a trivial task, and requires a sophisticated procedure.

We employ an HMM-based automatic speech recognition system trained using a modified version of the Baum-Welch algorithm for this segmentation. HMM-based large vocabulary continuous speech recognition systems model speech through context-dependent phonetic units called *triphones*. Each of these is modeled by an HMM with multiple states. As is well known, triphones are phonemes in context, many of which may be similar across multiple triphones. In conventional HMM-based speech recognition systems, the states of all HMMs are therefore *tied*, i.e. the state-output probability distributions of the HMMs are shared among the triphones corresponding to any phoneme.

The HMMs must be trained on speech data for which word-level, although not necessarily phonetic-level transcriptions are available. The phoneme-level (if not provided) and state-level segmentations are automatically derived during the training process. However, in the absence of other constraints, there is no assurance that any of the derived states will capture the locus regions of the phonemes; additional constraints are required to achieve consistent locus-region segmentations.

To achieve effective segmentation of locus regions, we incorporate additional constraints into our model. We model each triphone using a 3-state HMM. The "boundary" states of these models are intended to capture context-dependent variations of the phoneme, while the central state is intended to model the locus segments. Since the locus segments of different instances of a phoneme are expected to be very similar in structure (independently of context), this conformity is enforced by making the central states of all triphones of any phoneme share a common distribution, eliminating context-dependencies in the model for this state. We call this a *midstate-sharing* technique.

In order to minimize the variance of the distribution of this central "locus" state, we train the HMMs with a modified Baum-Welch algorithm that incorporates an entropic regularizer [16], which attempts to minimize the entropy of the distribution representing the central state. The effect of this regularization is to maximize the statistical similarity between the feature vectors assigned to the central state of every triphone. The details of the regularized training algorithm are omitted here, and largely follow the developments in [16]. The CMU-sphinx speech recognition system was modified for our experiments to include this. Figure 2 shows typical segmentations that are achieved by the algorithm. Note the similarity in the structure of the data in this state across the different instances of the phoneme.

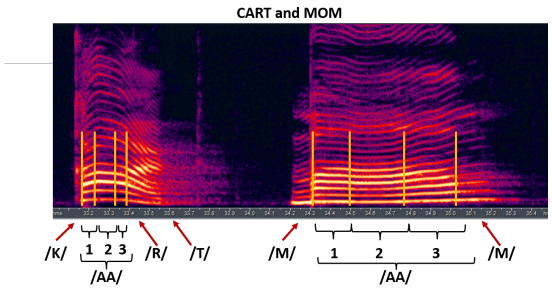

Fig. 2. State level segmentations for the phoneme AA in the American English pronunciation of the word CART (K AA R T) and MOM (M AA M).

IV. Experiments and Results

To evaluate the usefulness of the proposed approach, we apply it the deduction of height of speakers within the widely used and publicly available TIMIT continuous speech database [17], and to the estimation of age from the TIDigits database [18]. Note that our experiments are only for illustrative purposes, and we did not optimize the components used in them, such as feature types and models, to obtain the best possible performance. Nevertheless, we point out at the outset that the results we obtain for height are the best reported for the TIMIT database so far. The results we obtain from TIDigits have not been reported in the literature, and we have no points of comparison. We therefore only state them to the extent that they are illustrative of our procedure.

The first step in both cases was to segment the databases into their phonemic units. For this, we used the technique discussed in Section III-B. We used 3-state left-to-right Bakis topology HMMs for segmentation. These were trained on 5000 hours of clean speech data from a collection of standard speech databases in English available from the Linguistic Data Consortium. The training databases were parametrized using high-temporal-resolution MFCC vectors [19]. These were computed over 20ms analysis frames 200 times a second, to achieve a temporal resolution of 5ms. The trained HMMs were finally used to derive phoneme- and state-level segmentations of the recordings. The region of the middle state of each phoneme HMM was then taken to represent the locus region from which features for the deduction of height and age could be derived.

In this context, we note that the features we use for segmentation are not necessarily the same as the features we use to derive the anthropometric parameters of the speakers. Once the state-level segmentations are obtained, we extract the latter features from the central locus segment, as explained earlier in Section III-B. For the TIMIT data, our features were spectral features that capture resonance and fine variations in the envelope of the spectrum, but do not capture voluntary pitch-level detail. In this representation, we computed a 64-point *liftered* periodogram using Welch's method [20] over the locus segment. We used a 20-point lifter; this retains envelope-level detail in the spectrum, while smoothing out pitch-level structure. For the TIDigits database, we extracted the first five formants using the Burg's method [21] from each relevant segment.

Subsequent to the extraction of features, we followed two strategies. In the case of TIMIT, we conducted a 10-fold cross-validation experiment, partitioned across speakers. In each fold, all instances of a phoneme from 90% of the speakers were used to train a regression, which was used to predict heights from each instance of the phoneme spoken by the remaining 10% of speakers. The overall phoneme-specific predicted height for any speaker was the average of the predictions made by the individual instances of the phoneme by that speaker.

In the case of TIDigits, the phoneme-specific predictors were random forest regressions [22] that predict age, gender and accent from the formant features derived from the TIDigits sub-phonetic segments. Our experiments employed forests with 32 trees. Increasing this number had no significant effect on accuracy.

As described in Section II, a given speaker characteristic may heavily influence the sub-phonetic features of some phonemes, while not affecting others to the same degree. The specific phonemes affected may also vary with the feature. Accordingly, in both cases the next step was to identify the phonemes that resulted in the most statistically significant ($P - value < 0.0001$ using a t-test) predictions.

To demonstrate that the articulatory-phonetic conjectures we present earlier in this paper indeed hold, the predictions from the set of the most predictive phonemes were then combined using simple weighted interpolation. The weights for any predictor were the inverse training error standard deviations. The weighted combination resulted in a single aggregate prediction for the speaker. Note that in practice, for estimation of speaker parameters from new data, the knowledge of *which* phonemes are most predictive of the parameter (obtained from a different database, such as TIMIT in this paper) can still be effectively used in the same manner.

Figure 3 shows the Root Mean Squared Error (RMSE) of the height predictions obtained from each phoneme. The error ranges from 2.75 inches for the phoneme IY to 3.5 inches for ZH. We observe that all phonemes are not equivalent in their ability to predict height. Except for ZH (159 instances) and UH (762) instances, all phonemes are well represented in the set. The results are hence largely reflective of the innate

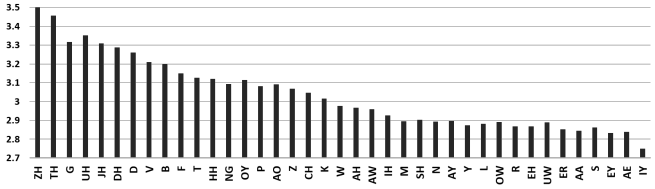

Fig. 3. RMS Error of the predicted values of height for each phoneme.

information about speaker height encoded in the phoneme. While not plotted here, the correlation between the predicted height and the actual height follows the trend shown by prediction accuracies and ranges between 0.1 (for ZH) to 0.7 (for IY). All correlation values are statistically significant with P-values of 0.001 or better. The combined prediction from all the phonemes with P-values less than 0.0001 achieves an overall RMSE of 2.54 inches.

To test our original hypothesis that the predictive ability of the phonemes varies, and reflects the articulator configuration of the phoneme, we evaluate the R^2 value of the individual phonemes, which represents how much of the variance in the dependent variable (height) can be predicted from the variations in the features for the phoneme. Phonemes with greater R^2 are more predictive of height. Fig. 5 shows the ten vowels with the greatest ability to predict height, color coded to show the rank ordering. Fig. 4 shows the top ten consonants.

We note that the phonemes that predict height best cluster closely in terms of articulator configurations. In particular, front vowels, and velar, alveolar and palatal consonants, all of which capture the acoustics of the *back* of the vocal tract, which relates more to body size, are seen to have the highest ability to predict height. Combining the predictions of just these phonemes results in an overall RMSE of 2.48 inches, which is among the best reported results for this data set [10].

Manner of articulation	Voicing	Place of articulation						
		Bilabial	Labio-dental	Inter-dental	Alveolar	Palatal	Velar	Glottal
Stop	Voiced	b			d		g	
	Unvoiced	p			t		k	
Fricative	Voiced		v	dh	z	zh		
	Unvoiced		f	th	s	sh		hh
Affricate	Voiced					jh		
	Unvoiced					ch		
Nasal	Voiced	m			n		ng	
Liquid(L)	Voiced				l			
Liquid(R)	Voiced					r		
Glide	Voiced	w				y	w	

Fig. 4. Classification of consonants based on the involvement of different parts of the vocal tract (place of articulation) and the manner of their movements (presence of absence of voicing, and the manner or articulation). This figure also shows consonants with the highest R^2 value for height. Shades of decreasing color intensities (black to light grey) depict decreasing R^2.

Note that the final combination was done only in the case of height estimation with TIMIT. We present the TIDigits case for illustration of the same technique for age prediction. Fig. 6 shows the capacity of the different phoneme to predict age, accent and gender from the TIDigits database. Of these, we

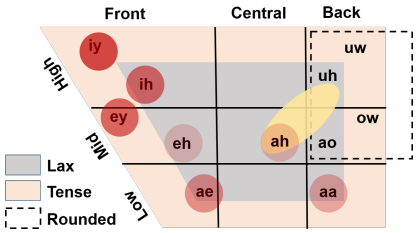

Fig. 5. Vowels in American English, modified from [23] to represent the specific vowels modeled by by the CMU Sphinx ASR system [24] used in our experiments. There are four additional phonemes used in our ASR system. These are the three *Diphthongs*: *ay* as in *side*, *aw* as in *how* and *oy* as in *toy*, and the semivowel *er* as in *surround*. This figure also shows the vowels with the highest R^2 for height. The semivowel ER is not shown here but is also as highly correlated to height as IY, which exhibits the greatest R^2. The figure shows decreasing R^2 values as red circles of decreasing color intensity.

only wish to highlight the case of age. We find that age is highly correlated to only a few phonemes, all of which turn out to be vowels.

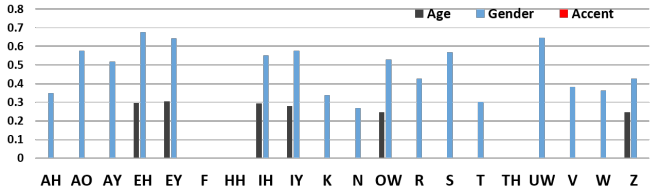

Fig. 6. The predictability of age, accent and gender from different phonemes in the TIDigits database, computed using sub-phonetic formant features over the entire database. Accent is not predictable from individual phonemes at all. This makes sense since prior studies have shown that accent is encoded *jointly* in phonemes, and entire formant charts are needed to identify them.

V. Conclusions

The results clearly demonstrate the validity of an articulatory-phonetic-based approach to forensic analysis of voice. The proposed methodology is of particular use in cases where the available voice sample may be of short duration, e.g. in the word *Mayday* from a hoax call, comprising just a few phonemes. Estimating speaker parameters based on only the most appropriate phonemes can provide useful results in these scenarios.

The articulatory-phonetic approach presented in this paper is exemplified in the context of deducing height and age, but can be applied to any other anthropometric, psychometric, sociometric and suchlike measurements from the voice. While we have not addressed robustness issues and analysis of noise-corrupted recordings in this paper, we have found in practice that once we are able to generate accurate sub-phonetic segmentations in these cases (using appropriate robustness techniques and specially modified state-of-art automatic speech recognition systems), our methodology applies well to the estimation of both anthropometric and psychometric parameters. We are in the process of publishing these results.

Acknowledgment

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI002-07, via the Command, Control and Interoperability Center for Advanced Data Analysis (CCICADA).

References

[1] D. Mendes and A. Ferreira, "Speaker identification using phonetic segmentation and normalized relative delays of source harmonics," in *Proc. 46th Audio Engineering Society Conference on Audio Forensics: Recording, Recovery, Analysis and Interpretation*, Denver, Colorado, USA, 2012, pp. 215–222.

[2] G. D. Burrows, "Skeletal and facial muscle psychophysiology," *in Handbook of studies on depression*, 2013.

[3] R. D. Kent and J. C. Rosenbek, "Acoustic patterns of apraxia of speech," *Journal of Speech, Language, and Hearing Research*, vol. 26, no. 2, pp. 231–249, 1983.

[4] P. Ekman, *Darwin and facial expression: A century of research in review*. Ishk, 2006.

[5] I. B. Goldstein, "The relationship of muscle tension and autonomic activity to psychiatric disorders." *Psychosomatic Medicine*, vol. 27, no. 1, pp. 39–52, 1965.

[6] ——, "Role of muscle tension in personality theory," *Psychological Bulletin*, vol. 61, no. 6, p. 413, 1964.

[7] S. Gutta, J. R. J. Huang, P. Jonathon, and H. Wechsler, "Mixture of experts for classification of gender, ethnic origin, and pose of human faces," *IEEE Transactions on Neural Networks*, vol. 11, no. 4, pp. 948–960, 2000.

[8] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 8, pp. 1177–1193, 2012.

[9] M. Sonderegger and J. Keshet, "Automatic discriminative measurement of voice onset time." in *INTERSPEECH*, 2010, pp. 2242–2245.

[10] R. Singh, J. Keshet, D. Gencaga, and B. Raj, "The relationship of voice onset time and voice offset time to physical age," in *Proc. ICASSP*, 2016.

[11] P. Auzou, C. Ozsancak, R. J. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: a review," *Clinical Linguistics & Phonetics*, vol. 14, no. 2, pp. 131–150, 2000.

[12] 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.

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

[14] C. T. Ferrand, *Speech Science: An Integrated Approach to Theory and Clinical Practice*. Allyn & Bacon, 2006.

[15] P. Delattre, "Coarticulation and the locus theory," *Studia Linguistica*, vol. 23, no. 1, pp. 1–26, 1969.

[16] M. Brand, "Structure learning in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1182, 1999.

[17] L. D. Consortium, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S1, 1993.

[18] ——, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S10, 1993.

[19] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *Proc. International Workshop on Biometrics and Forensics*. Limassol, Cyprus: IEEE, March 2016.

[20] P. D. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE TAE*, vol. 15, no. 2, pp. 70–73, 1967.

[21] 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.

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

[23] T. L. S. Project, "Sounds of standard American English," *University of Arizona*, 2001.

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