FORMANT MANIPULATIONS IN VOICE DISGUISE BY MIMICRY

Rita Singh [†], *Deniz Gencaga* [‡] *and Bhiksha Raj* [†]

[†] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[‡] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA

ABSTRACT

The human voice can be disguised in many ways. The purpose of disguise could either be to impersonate another person, or to conceal the identity of the original speaker, or both. On the other hand, the goal of any biometric analysis on disguised voices could also be twofold: either to find out if the originator of the disguised voice is a given speaker, or to know how a speaker's voice can be manipulated so that the extent and type of disguise that the speaker can perform can be guessed *a-priori*. Any analysis toward the former goal must rely on the knowledge of what characteristics of a person's voice are least affected or unaffected by attempted disguise. Analysis towards the latter goal must use the knowledge of what sounds are typically most amenable to voluntary variation by the speaker, so that the extent to which given speakers can successfully disguise their voice can be estimated. Our paper attempts to establish a simple methodology for analysis of voice for both goals. We study the voice impersonations performed by an expert mimic, focusing specifically on formants and formant-related measurements, to find out the extent and type of formant manipulations that are performed by the expert at the level of individual phonemes. Expert mimicry is an extreme form of attempted disguise. Our study is presented with the expectation that non-expert attempts at voice disguise by mimicry will fall within the gold standard of manipulation patterns set by an expert mimic, and that it is therefore useful to establish this gold standard.

Index Terms— Voice disguise, Mimicry, Impersonation, Voice biometrics, Voice forensics, Formant analysis

1. INTRODUCTION

While the human vocal tract is restricted in its ability to produce the larger fraction of non-speech sounds encountered in the real world, the fact that one human being *can* successfully impersonate another's voice is not contested. Voice mimicry is in fact an established subarea of the performing arts, and is taught as a skill in many schools of Drama and performing arts across the world. Until voice-only communications became commonplace, voice impersonation was largely of value only for entertainment purposes. In the current day, however, nefarious motives have begun to be associated with this. There are many situations in which a person may try to impersonate another. When communicating content of a misleading nature, a person may deliberately try to sound like another person. Successful impersonation is of special benefit to people who try to bypass or confuse voice biometric systems, such as voice-password (or voice-fingerprint) based applications. Hoax callers who attempt to mobilize emergency and security services for malicious purposes often attempt to disguise their voices by impersonation (e.g. a man trying to sound like another, or like a woman). There are increasing incidences of *swatting* – wherein people call law enforcement agencies with false reports of dangerous activities at a location (e.g. report a neighbor concocting amphetamines in their basement or holding an illegal dog fight at their location). Swatting can have tragic and disastrous consequences [1]. It has in fact become a serious problem for law enforcement agencies. In these, and many other types of ill-intentioned voice communications, the use of voice impersonation is commonly observed.

At the outset we note that impersonation in this context need not always be of a known (living or dead) person's voice. Impersonation may often merely be that of a hypothetical person's voice to convey the impression that the speaker is *not* the person who is the real originator of the voice. Impersonation is also the most misleading kind of voice disguise. When voice is disguised by physical masking [2] or other methods that "scramble" voice resulting in unnatural sounding voice with intelligible content [3], it is often easy to detect it as a disguised voice. This information alone could be used to block certain categories of activities such as socially harmful communication, influence decisions to respond (or not) to hoax calls, etc. When the disguise is in the form of impersonation, such information may not be easy to derive. It becomes necessary to perform a finer analysis of voice to confirm or rule out the possibility of disguise. Furthermore, if the perpetrator habitually engages in voice disguise using multiple impersonations to further thwart any identification, it may be necessary to find ways to identify the common originator of all or a subset of such voice samples at hand.

To achieve these goals, it is necessary to understand clearly what aspects of speech are variant or invariant during impersonation. The goal of this paper is more directed towards deducing *invariant* units of sounds during mimicry, so that these can be reliably used for further forensic investigations to uncover the identity of the speaker. We perform an extremely controlled study of formants and formant-related measurements, namely formant bandwidths, dispersion and spacings, of impersonated voices by a professional mimic who has been publicly acclaimed to be one of the best impersonators in the world today. Note that while formants are clearly not the only parameters that are active in any successful or attempted impersonation, and the mimic often makes heavy use of facial expressions, speaking styles and gestures, they are nevertheless extremely significant in defining and preserving intelligibility and discriminability of individual speech sounds. They are also important in impersonations that are entirely done within voice-only communications.

1.1. Formants and their relation to body parameters

We use formants for deducing sounds that are invariant to impersonation, rather than the more easily derived voice fundamental frequency, often referred to as $F0$, or pitch. The reason is that pitch does not appear to contain information that could be useful in tracking invariant characteristics of the same speaker (and therefore their speech) in different situations. This is evidenced by the fact that strong relationships of $F0$ to speaker characteristics have never been

978-1-4673-9448-2/16/$31.00 © 2016 IEEE

found, despite the fact that more studies have been performed with $F0$ than with formants. An interesting example is that of body size. Historically, as far back as 1871 [4], it was believed that pitch relates to body size. This view was still controversial after a century, where there are contradictory studies reporting $F0$ to be correlated to body size [5], and not so [6, 7, 8, 9].

Formants are related to F0, but only weakly so. Formants are the resonant frequencies of the vocal tract that are activated in response to the vocal tract configuration, and also to any residual voicing or excitation of the vocal tract. While $F0$ can remain constant on the average through a specific set of sounds, formants differ for different phonemes, since the vocal tract shape is different while enunciating each phoneme. Formants are important in the discrimination of different sounds by the human ear. Pitch is not important in this respect. Formants and formant-related measurements have also been consistently found to correlate to body parameters in many studies with significantly more success. Studies associate them positively to body size [10], height [11], vocal tract shape [12], age [13] etc. Other studies have reported their correlation to body size in humans by association to our evolutionary relatives, e.g. the study in [6] was the first to present strong evidence that *formant dispersion*, defined as the mean difference between successive formant frequencies, correlated strongly with vocal tract length and body size. Later studies have confirmed this.

These, and other similar studies, indicate that formant measurements in disguised or impersonated voices could present an interesting and potentially valuable study case. In a collection of impersonations by the same speaker, neither the absolute vocal tract length nor the body parameters change, especially if the impersonations are done within the same short time duration (minutes or even hours). If the impersonator is successful at giving the impression of a different person's body parameters (implicitly), it could be interesting to find out if there is any subset of sounds nevertheless that could still lead to the identity of the impersonator. Our study is precisely directed towards finding such a subset.

The rest of this paper is organized as follows. In Section 2 we explain formant and formant-related measurements and discuss their role in voice disguise. In Section 3 we describe the key methodologies used to derive the formants, and the statistical measures we propose to use in this study. In Section 4 we describe our experimental setup and also present our results. In Section 5 we present our conclusions.

2. FORMANT CHARACTERIZATIONS AND MIMICRY

2.1. Formants

Formants are frequencies at which the spectral energy peaks in the speech signal. The human vocal tract may be viewed as an acoustic tube. In the act of producing different speech sounds, the speaker modifies the configuration of the vocal tract to result in partially separated segments or cavities of different lengths. Formants are produced by the resonances of these chambers.

Formants generally appear as peaks in spectrographic displays of the speech signal. The top panel in Figure 1 shows the spectrogram for the word "DARK". The spectrogram is seen to exhibit several distinct horizontal bands of energy. Each of these bands is 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 spectrogram. While resonant frequencies corresponding to formants $F6$ and higher do exist, they are generally hard to discern.

Formants, being the primary characteristic of the sounds produced by any given configuration of the vocal tract, are key to disambiguating phonemes. Typically, the first three formants $F1-F3$ are sufficient to disambiguate vowel sounds. Other sounds too have formant-arrangement signatures. For example the liquid /l/ generally exhibits a formant at about 1500Hz. Nasal consonants often have their third formant canceled out by anti-resonances resulting from the opening of the nasal passageway. The liquid /r/ is distinguished by a third formant that dips below 2000Hz. Formants are often not clearly discernible in fricative sounds such as /ch/, /jh/, /hh/ etc. due to the noisy nature of these sounds; they are nevertheless present. Even stop sounds such as /t/, /p/, /k/ etc., which are associated with a complete closure of the vocal tract and consequent cessation of energy in the output signal, are nevertheless produced by vocal tract configurations with resonances and often have observable formant frequencies. The bottom panel in Figure 1 shows the spectrogram of the word "TOOTHPASTE" spoken by a young female. We observe that formants are discernible for almost all phonemes, regardless of whether they are voiced or not.

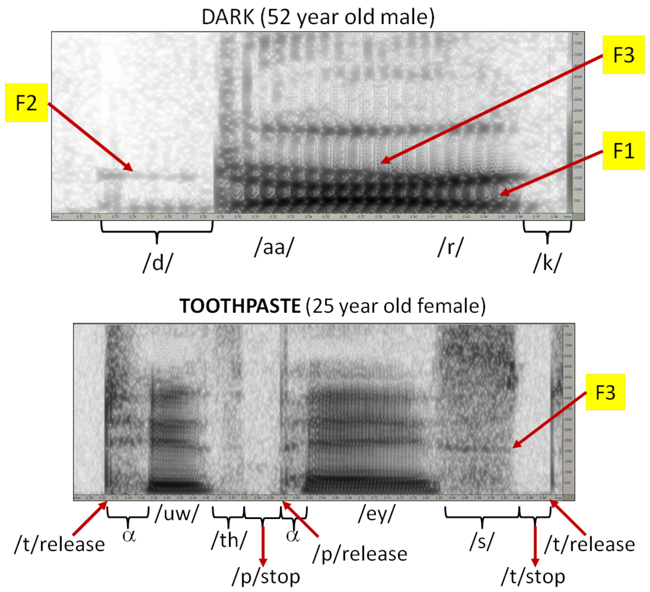

Fig. 1. **Top:** Spectrogram of the word "DARK" spoken by a 52 year old male. The formants F1-F5 are clearly observed, as are their bandwidths. **Bottom:** Spectrogram of the word TOOTHPASTE showing the variation of formants between phonemes. The symbol α marks the voice onset time. Note that even the unvoiced phonemes such as /t/, /p/ and /s/ have formants since the vocal tract continues to resonate through their production period. This is *not* the case for the voice pitch, which manifests only for voiced sounds, i.e. when the vocal cords vibrate.

While the formant frequencies are thus characteristic of the underlying phoneme, they are also characteristic of the *vocal tract* itself; in fact vocal tract length is known to be the second largest contributor to formant frequency variation after phoneme identity [14]. Longer vocal tracts may be expected to lower resonances and hence lower the formants. In principle, for the open sound /ah/, which is produced by a relaxed and fully open vocal tract, a loss-less tube model gives us an estimate of the vocal tract length derived from the n^{th} formant as $L = \frac{F_n}{2n-1}$. Under more realistic assumptions formulaic estimators relating measured formant frequency values to vocal tract length can nevertheless be found [15].

Not surprisingly, therefore, formant frequencies are also characteristic of the *speaker*, since the underlying physical system, namely the vocal tract, is an invariant characteristic of the speaker. From our perspective, it is this characteristic that we wish to key in on when characterizing speech mimicry, since the mimic's vocal tract is an inherent characteristic and its effect cannot entirely be eliminated from his/her speech.

2.2. Formant bandwidth

The spectral energy that peaks at formant frequencies also rolls off as the frequency moves away from the formant. The formant *bandwidth* is defined as the spread of frequencies around the formant within which the energy remains within 3dB of the formant energy. Figure 1 also shows the bandwidths of the formants.

Formant bandwidth generally has a smaller effect on the identity of the sound, but is nevertheless related to formant frequency: higher formants have greater bandwidths in general. However formant bandwidth is also dependent on vocal tract *composition*, such as the elasticity of the walls, energy dissipation through the glottis etc., in addition to other configuration-related characteristics [16].

Not surprisingly then, formant bandwidth too has speaker-specific characteristics. In particular, damping effects due to the nature of the tissue of the vocal tract, and damping due to energy dissipation from the glottis and nasal pathways are related to the size and coupling of these passages with the main vocal cavity. From our perspective, these effects may be expected to manifest directly in a mimic's voice in a manner which he/she cannot entirely control.

2.3. Formant Q

A vocal tract configuration that generates a specific formant may be viewed as a filter with a particular resonant frequency. The *Q*-factor of any filter is defined as the ratio of the peak frequency of the filter to its bandwidth. In the case of formants, the formant *Q* is the ratio of the formant frequency to the formant bandwidth.

There is an inherent relationship between the formant frequency and the formant bandwidth. In general, as the formant frequency increases, so will its bandwidth. However, the actual Q of any formant depends not only on the frequency of the formant, but also on the frequency-dependent characteristics of the vocal tract of the speaker. Once again, given the invariance of the vocal tract of the speaker, one may expect un-maskable speaker-specific characteristics to manifest in formant Qs of mimicked speech.

2.4. Formant Dispersion

Formant dispersion is defined as the average spacing between formants. Formant dispersion, being dependent only on the spacing between formants and not on the absolute position of the formants themselves, has been suggested as being more characteristic of the speaker's vocal tract length than the formant positions themselves [6]. Once again, as in the other measurements, this translates to speaker dependence of formant dispersion. Consequently, we may expect the distribution of formant dispersion values to have speaker-specific characteristics that a mimic may not have complete control over.

We note here that the common definition of formant dispersion, given by $D = \frac{(F_2-F_1)+(F_3-F_2)+\cdots+(F_n-F_{n-1})}{n-1}$, simply collapses to $\frac{F_n-F_1}{n-1}$, *i.e.* the average distance between formants, which loses the distinction between the individual spacings. In order to better reflect the effect of individual formant spacings, we redefine dispersion in this paper as the *geometric* mean of the formant spacings, *i.e.* $D = \sqrt[n-1]{\prod_i F_i - F_{i-1}}$.

2.5. Measuring Formant Features

In general, the estimation of formant frequencies and bandwidths from a speech signal is a challenging task even when the speech recording is clean, and particularly so when it is noisy. A number of algorithms have been proposed in the literature for this purpose. The most common approach is to analyze segments of a speech signal using an auto-regressive (AR) model, and to track the positions and bandwidths of the poles in the model [17].

In this paper we have employed the Burg algorithm for AR analysis as implemented in PRAAT [18], a popular open-source tool that is probably the most popular tool for these purposes. The speech signal is segmented into analysis "frames" of 25ms each, where adjacent frames overlap by 15ms, leading to an analysis frame rate of 100 frames per second. Formant measurements are obtained from each analysis frame. Measurements from adjacent frames are smoothed in order to minimize the occurrence of random variations and outliers in the measurements. Q- and dispersion values were derived from the computed formant and formant-bandwidth values.

3. STATISTICAL ANALYSIS OF IMPERSONATION

There is, in general, a paucity of studies on what makes a speech mimic successful. Much of the literature on speech mimicry deals with unintended mimicry due to various convergence effects in conversation [19], or in the context of voice spoofing to beat speaker verification or identification systems [20, 21]. Relatively less information is available from studies of professional mimics.

However, a few studies are, indeed available. Eriksson [22] reports that professional mimics are generally good at mimicking timing and prosodic cues of the target, as one might expect intuitively, and further confirmed in [23], but less so at mimicking a target's formants. Zetterholm [24] reports that mimics also capture the intonation and articulation of the target. This study reports that the mimic varies his formants to match the target, and while he may not actually achieve the target's formants, he does succeed in making them significantly different from his own. However, the study only reports aggregate statistics, and it is not at all clear that this effect will persist for *every* phoneme. Intuitively, one may expect that such change is more likely in vowels, where formant frequencies are well perceived and recognized, and less so in consonants, fricatives, and diphthongized sounds where the precise location of instantaneous formant frequency is less critical to the recognition of the sound. Ashour and Gath [25] show that mimics also track the *pitch* of the target speaker, a hypothesis also confirmed by Majewski [26]. They also study a small number of vowel triplets that appears to confirm our hypothesis that mimics will vary their formants towards that of the target in vowels, although they may not actually achieve the target's values.

We accept the hypothesis that mimics do a very good job of capturing their target's timing, articulation and pitch cues – this is not only evident from the literature, but also from even a casual perusal of performances of mimicry uploaded to YouTube. We also accept the hypothesis that they vary their formants, although they may not necessarily achieve the actual formants of their targets.

Instead, we focus on their *failures* – the phonemes they *fail* to vary, and yet nevertheless manage to successfully imitate the target. We hypothesize that the mimic does not actually modify *all* phonemes – since the intent is to project the vocal image of the target speaker, or perhaps even a caricature of the target, this does not require perfect mimicry or even modification of all phonemes. Rather, there may be a group of one or more phonemes that the mimic does not, or possibly *can*not vary significantly during mimicry, and that the characteristics of formants for these phonemes retain the statistical characteristics of the speaker himself.

3.1. Statistical tests for the similarity of phonemes

We base our analyses on statistical tests conducted on formant measurements of speech from the mimic. We present the details of the data later in Section 4; for now we only outline the testing procedure.

Our tests are by phoneme. The objective is to establish, statistically, if it can be stated with high confidence that the mimic *did* vary the formants of his speech when pronouncing the phoneme when imitating different targets.

We obtain several instances of each phoneme when the mimic is imitating each of several target speakers. We obtain formant trajectories for each phoneme. Since formant trajectories are constantly changing, and we wish to evaluate, as closely as possible, the canonical expression of the phoneme by the mimic, we only use the formants from the central third of each phoneme segment. In general, not all formants can be measured in any segment; however, where formants *can* be estimated, $F1-F3$ can almost always be measured. $F4$ and $F5$ are more difficult to detect and track. Our analysis focuses on $F1-F3$, using $F4$ and $F5$ where available, and treating them as missing at other times.

We treat the analysis as a hypothesis testing problem. Our null hypothesis is that formant measurements from all instances of a phoneme, obtained when the mimic imitates different speakers, are drawn from the same distribution. The alternate hypothesis is that they have different distributions. The statistical test attempts to establish a two-sided probability that the observed distributions of the measurement could have been obtained from a common distribution. If this probability is high, the alternate hypothesis must be discarded.

Our measures are generally multivariate – since we will be considering groups of measured variables at a time. Moreover, within any instance of a phoneme we obtain multiple measurements, one from each analysis frame of the signal. These measurements are correlated, since formant variation is generally not rapid and formant trajectories are smooth. Additional smoothness is also imposed by the smoothing of formant trajectories during measurement (although in practice we find that smoothing is infrequently invoked and does not affect our analyses significantly).

Our statistical analysis must therefore be able to handle repeated multi-variate measurements. We use *repeated measures* ANOVA, which we will refer to as R-ANOVA, for this purpose[27]. Repeated measures ANOVA is a variant of multi-variate ANOVA that accounts for correlations between temporally adjacent measurements by subtracting out the effect of correlations between the adjacent measurements when computing within-class variability. We refer the reader to [28] for a detailed explanation of the test.

ANOVA and its variants are, in general, not ideally candidates for this test. ANOVA implicitly assumes a Gaussian distribution for the groups and further that the groups are homoskedastic. However, it is known to be robust to variations in class variance, and is found to be widely applicable even in situations where the theoretical requirements of the test do not apply exactly. In our situation, our tests incorporate many groups, and provided we arrange our variables appropriately for the test to be applicable, the test remains sufficiently informative for our purposes, although the exact P-values (i.e, the probability that the measurements for all speakers were drawn from a common distribution) provided by the test may not be accurate.

4. EXPERIMENTS AND RESULTS

Our study was performed on recordings of impersonations by Jim Meskimen, a renowned voice mimic. The data included impersonations of 50 different celebrities (actors, scientists, famous political figures etc.), rendered by Jim Meskimen at the same time while reading text from Shakespeare. The audio recordings were 16kHz mono recordings. The total speech used was a little over 0.5 hours. While the amount of data for each of these impersonations was limited to less than a minute each, and it would be hard to find the same words across the impersonations, this amount of data is nevertheless sufficient to obtain multiple examples of each phoneme from each impersonation. Formant readings were then obtained at 0.01 second intervals within each phoneme, resulting in a total of about 190,000 readings spread over 42 phonemes, excluding silences and other non-speech sounds.

The reason for selecting this particular set of impersonations is that the voice artist renders these impersonations in a rapid sequence, in one sitting, while reading a piece from some literature so that the subject and theme are constant. This is an important factor in normalizing sentiment, since sentiment (and the expression thereof) can cause significant changes in the the underlying formant patters. Some examples of renderings in our collected database include EG6 [29], Celebrity impressions Alphabet [30], Don't do drugs do impressions [31], What is in a human voice (with music removed) [32], Sonnet 130 by William Shakespeare [33], T'was the night before Christmas from Saint Nicholas by Clement Clark Moore [34] etc. Note that in all of these, the artist speaks in his own voice as well, usually at the beginning and the end of each performance. Our analysis used the artist's voice from these segments as reference (the advantages of doing so are many, including the normalization of sentiment and mood). In some recordings, other people spoke as well, e.g. the host of the show introducing the artist. These non-impersonations were removed from the analysis.

Phoneme segmentation for accurate analysis: The database was transcribed manually to obtain accurate transcriptions with accurate speech and non-speech events, and disfluencies clearly marked. The CMU Sphinx-3 automatic speech recognition system [35], trained with approximately 5000 hours of clean speech for this experiment, was then adapted to the recordings in a supervised fashion to obtain the best possible acoustic segmentation for the phonemes. The phoneme boundaries that were automatically obtained by the speech recognition system were thereafter manually spot-checked and found to be extremely accurate.

Choosing successful mimicry: For each of the mimicry targets we attempted to establish the success of the mimicry. To do so, we ran the following test, comparing actual spoken utterances by the target of the mimicry to sentences spoken by the mimic: we played a randomly selected pair of sentences to listeners. Each of the sentences was spoken either by the mimic (attempting to sound like the target), or by the target himself/herself. The listener was asked to determine if the two sentences were spoken by the same person or not. For each mimicry target we ran the above test using 18 subjects. The null hypothesis was that the listeners would be unable to distinguish between the mimic and the target. The alternate hypothesis was that they could. A two-sided test was conducted to test the hypothesis. If the null hypothesis could not be rejected at a generous confidence level of $p = 0.20$, the mimicry was deemed to be perfect. Otherwise, the attempt at mimicry was considered imperfect. Of all the imitated subjects, only 8 were rejected as imperfect mimicry. The successfully mimicked targets included a female target.

Results: Figure 2 shows the P-values for formants, formant bandwidths, Q-values and formant dispersion as defined by us in Section 2. We note that the phonemes that are clearly unaffected by all impersonation attempts by the artist include /jh/, /ch/ and /y/ in all cases, and other phonemes as well. The absence of distinction between different target-specific renditions of these phonemes could either imply a failure on the part of the mimic to make these distinctions, or that these phonemes themselves do not vary across speakers. To test the latter hypothesis we evaluate the usefulness of these phonemes in making biometric predictions. If they do provide biometric information, then it can be assumed that they do in fact vary across speakers, but the mimic was unable to replicate this variation.

We studied the predictive potential of these phonemes using the TIMIT acoustic-phonetic corpus [36]. This corpus comprises 630 speakers representing eight major dialects of American English. The recordings contain 16kHz sampled speech recordings of ten phonetically rich sentences that are read by each speaker. The corpus is phonetically balanced. All phonemes are well represented within each speaker and across the database. The biometric parameters of each speaker were recorded at the time of corpus collection. From amongst these, we chose to demonstrate the usefulness of the phonemes in question by investigating their predictive potential for the speaker's gender and height. As a contrast, we must also use instances of these phonemes from the mimic's voice to determine if they are predictive of *his* gender or height, or those of the targets. However, since all of his targets chosen were male, and we were unaware of their heights, this test could not be run. The gender classification detected all instances as male.

In our experiment we first obtained accurate phoneme segmentations for the TIMIT database using the clean speech acoustic models mentioned above, and built a Gaussian-Mixture-based gender and height classifier individually from examples of each phoneme. For each experiment, the training set for each phoneme comprised all examples of the phoneme from the 462 speakers in the corpus-designated training set. The test set comprised phonemes from 56 male and 56 female speakers selected from the corpus-designated TIMIT test set. The results are shown for a selected set of phonemes in Table 1. Some phonemes including /jh/, /b/, and /th/ were excluded for logistic reasons (a few speakers in the test set did not have enough examples of these). We note from Table 1 that height estimation is not as good as state-of-art but since the inference of biometric parameters is not the focus of this paper, we believe these results are sufficient to show the biometric potential of these phonemes.

Phoneme	CH	Y	DH	M	OY	P	HH	SH
Gender	83	92	72	80	83	69	81	88
Height	57	66	52	51	58	57	57	63

Table 1. Gender and height recognition accuracies (in rounded percent figures) for some impersonation-invariant phonemes, demonstrated on the TIMIT corpus. Height is determined in inches with a range of +- 2 inches.

Variation of vowels in impersonations: Our study found that all vowels were maximally varied by the impersonator. Figure 3 shows the $F1-F2$ and $F2-F3$ variations of vowels for 39 of the best impersonated speakers.

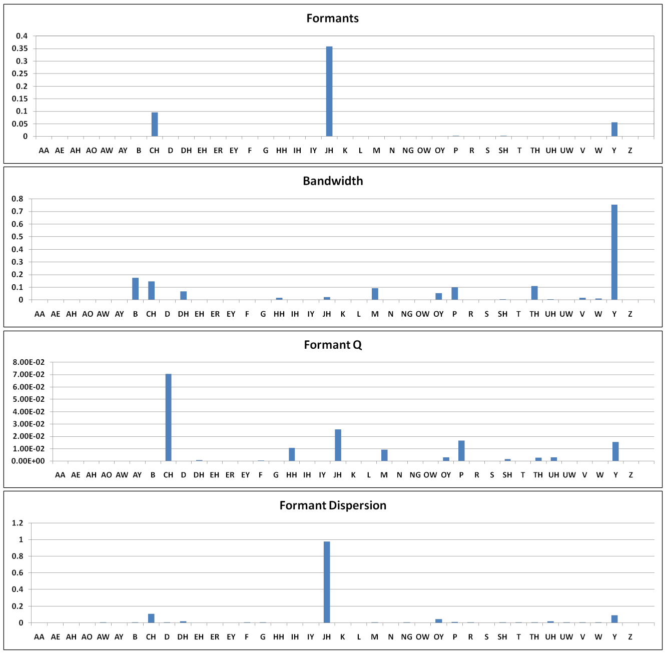

Fig. 2. P-values for various formant measurements.

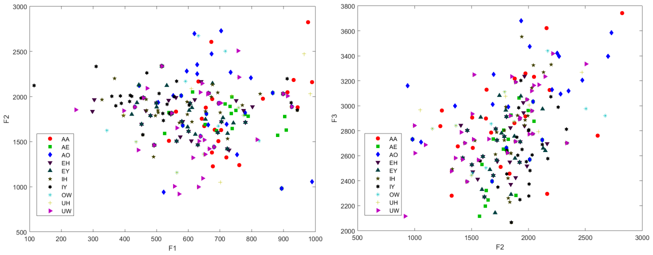

Fig. 3. Average vowel positions on the $F1-F2$ and $F2-F3$ planes for a few impersonations. No systematic patterns are evident so far in these variations.

5. CONCLUSIONS

We observe that expert impersonations have complex formant variations for some phonemes, and almost none for others. We present a methodology that can be followed in a full analysis for formant patterns in such situations. Formant variations in vowels seem to be patternless. However, patterns could potentially be uncovered by more complex models that possibly take into account the voice characteristics of *target* speakers that the artist impersonates. There are several follow-ups to this work that are clearly needed. First, the target speakers who are impersonated must be correlated to the impersonated ones. Second, the biometrics of the *impersonator* must be matched with those generated or deduced from impersonation-resistant phonemes to evaluate their value in biometric deductions from disguised voices.

More generally, our works suggests that the study of mimicry to identify mimicry- or spoofing-invariant features may be a valuable tool for biometric analysis, particularly of disguised voices.

6. REFERENCES

[1] USA Federal Bureau of Investigation, "The Crime of 'Swatting'," https://www.fbi.gov/news/stories/2013/september/the-crime-of-swatting-fake-9-1-1-calls-have-real-consequences.

[2] Mel Goldberg, "Face mask with voice modifying capability," July 1987, US Patent 4,683,588.

[3] Cuiling Zhang and Tiejun Tan, "Voice disguise and automatic speaker recognition," *Forensic Science International*, vol. 175, no. 2, pp. 118–122, 2008.

[4] Charles Darwin, "The descent of man and selection in relation to sex," *London: Murray*, 1871.

[5] Sarah Evans, Nick Neave, and Delia Wakelin, "Relationships between vocal characteristics and body size and shape in human males: an evolutionary explanation for a deep male voice," *Biological Psychology*, vol. 72, no. 2, pp. 160–163, 2006.

[6] 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.

[7] Norman J Lass and William S Brown, "Correlational study of speakers heights, weights, body surface areas, and speaking fundamental frequencies," *The Journal of the Acoustical Society of America*, vol. 63, no. 4, pp. 1218–1220, 1978.

[8] Harry Hollien, Rachel Green, and Karen Massey, "Longitudinal research on adolescent voice change in males," *The Journal of the Acoustical Society of America*, vol. 96, no. 5, pp. 2646–2654, 1994.

[9] 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," *The Journal of the Acoustical Society of America*, vol. 117, pp. 944, 2005.

[10] David A Puts, Coren L Apicella, and Rodrigo 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.

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

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

[13] Markus Iseli, Yen-Liang Shue, and Abeer Alwan, "Age-and gender-dependent analysis of voice source characteristics," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2006, vol. 1, pp. I–I.

[14] Richard E Turner, Thomas C Walters, Jessica J M Monaghan, and Roy 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.

[15] A. C. Lammert and S. S. Narayanan, "On short-time estimation of vocal tract length from formant frequencies," *PLoS One*, vol. 10, no. 7, 2015.

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

[17] Roy C Snell and Fausto Milinazzo, "Formant location from LPC analysis data," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 2, pp. 129–134, 1993.

[18] Paul Boersma and David Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[19] Elizabeth Couper-Kui-Ilen, "The prosody of repetition: On quoting and mimicry," *Prosody in Conversation: Interactional Studies*, , no. 12, pp. 366, 1996.

[20] Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi, "Spoofing and countermeasures for automatic speaker verification.," in *INTERSPEECH*, 2013, pp. 925–929.

[21] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," *TARGET*, vol. 4000, pp. 5000, 2014.

[22] Anders Eriksson and Pär Wretling, "How flexible is the human voice?–a case study of mimicry.," *Target*, vol. 30, no. 43.20, pp. 29–90, 1997.

[23] Pär Wretling and Anders Eriksson, "Is articulatory timing speaker specific?–evidence from imitated voices," in *Proc. FONETIK*, 1998, vol. 98, pp. 48–52.

[24] Elisabeth Zetterholm, "Intonation pattern and duration differences in imitated speech," in *International Conference on Speech Prosody*, 2002.

[25] Gal Ashour and Isak Gath, "Characterization of speech during imitation.," in *EUROSPEECH*, 1999.

[26] Wojciech Majewski and Piotr Staroniewicz, "Imitation of target speakers by different types of impersonators," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pp. 104–112. Springer, 2011.

[27] Kevin P Weinfurt, *Repeated measures analysis: ANOVA, MANOVA, and HLM.*, American Psychological Association, 2000.

[28] LUND Research, "Repeated measures ANOVA," 2013.

[29] Jim Meskimen, "EG6," `https://www.youtube.com/watch?v=3WgI7xa7ETM`.

[30] Jim Meskimen, "Celebrity impressions Alphabet," `https://www.youtube.com/watch?v=Ao64ONUG5Uw`.

[31] Jim Meskimen, "Don't do drugs do impressions," `https://www.youtube.com/watch?v=CtOLQWnM2GQ`.

[32] Jim Meskimen, "What is in a human voice," `https://www.youtube.com/watch?v=MtJdd-BS8MM`.

[33] Jim Meskimen, "Sonnet 130 by William Shakespeare," `https://www.youtube.com/watch?v=AHnHDq5S9cQ`.

[34] Jim Meskimen, "Twas the night before Christmas from Saint Nicholas by Clemet Clark Moore," `https://www.youtube.com/watch?v=iobTXPgETOY`.

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

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