Glottal and vocal tract characteristics of voice impersonators

Talal Bin Amin, Pina Marziliano, James Sneed German

Glottal and Vocal Tract Characteristics of Voice Impersonators

Talal Bin Amin, *Student Member, IEEE,* Pina Marziliano, *Member, IEEE,* and James Sneed German

***Abstract*—Voice impersonators possess a flexible voice which allows them to imitate and create different voice identities. These impersonations present a challenge for forensic analysis and speaker identification systems. To better understand the phenomena underlying successful voice impersonation, we collected a database of synchronous speech and ElectroGlottoGraphic (EGG) signals from three voice impersonators each producing nine distinct voice identities. We analysed glottal and vocal tract measures including F0, speech rate, vowel formant frequencies, and timing characteristics of the vocal folds. Our analysis confirmed that the impersonators modulated all four parameters in producing the voices, and provides a lower bound on the scale of variability that is available to impersonators. Importantly, vowel formant differences across voices were highly dependent on vowel category, showing that such effects cannot be captured by global transformations that ignore the linguistic parse. We address this issue through the development of a no-reference objective metric based on the vowel-dependent variance of the formants associated with each voice. This metric both ranks the impersonators natural voices highly, and correlates strongly with the results of a subjective listening test. Together, these results demonstrate the utility of voice variability data for the development of voice disguise detection and speaker identification applications.**

***Index Terms*—acoustic, disguise, glottal, speech rate, open quotient, formant, vocal tract, voice impersonator, voice identity.**

I. Introduction

VOICE impersonation is an art which involves changing one's voice to sound like another person. It is mostly used for entertainment purposes (e.g., for caricaturization and in media related fields). However, the study of voice impersonators is also important in other fields of research including forensics [1], speaker recognition [2], Text-To-Speech (TTS) synthesis and voice conversion [3]. From the point of view of forensics, for example, one can mask one's identity through voice disguise in order to avoid being identified. It is therefore important to develop methods that allow law enforcement authorities to identify individuals in spite of such modifications. Similar issues apply to security systems based on speaker recognition, which are also vulnerable to circumvention by voice impersonators [4], [5]. Currently, many speech transformation applications such as voice conversion [6], [3] and Text-To-Speech (TTS) synthesis also suffer from a lack of naturalness in the synthesized speech. This suggests that key aspects of speech that contribute to naturalness are being ignored by current speech transformation techniques. There exists a need, therefore, to identify the set of voice parameters that are involved in successful (i.e., natural-sounding) voice modification, and to explore how these can be used to improve voice disguise identification, voice identity detection, as well as how these parameters can be manipulated in a way that preserves naturalness across transformations.

A direct inspiration can be drawn from voice impersonators, who successfully maintain both naturalness and individuality while producing different voice identities. Such impersonations can be natural enough to deceive humans as well as automated speaker identification systems [4], [5].

The analysis of the glottal and vocal tract parameters of impersonated voices can be useful for voice disguise identification, where there is a need to identify a set of parameters that can be used to determine whether a voice is disguised or not. In this regard, the analysis and comparison of an impersonator's natural voice with the impersonated (disguised) voices reveals how various acoustic parameters are manipulated to extend a space of disguised voices around the natural voice. It can also reveal any invariant parameters or systematic relationships between the natural and impersonated voices, either of which may be readily exploited for voice disguise identification. In [3] the need for studying voice impersonators was specifically highlighted in connection with voice conversion applications in order to better understand how the issue of naturalness under identity changes can be better incorporated into speech transformation algorithms.

Not all portions of the range of variation that a speaker is capable of producing will result in natural-sounding speech. At the same time, there are limitations on the range of variation that a given speaker can produce. Speech parameters such as F0 (pitch) range and formant frequency, for example, may be constrained by a speaker's physical traits (esp. vocal cord anatomy and vocal tract length). As a first step, then, it is important to be able to model not only the amount and type of variation within the total parameter space that results in natural-sounding voices, but also to consider which regions of that space are achievable by a single speaker given his or her inherent physical limitations. In that sense, a central goal of our study is to begin to "map out" the space of variation in speech parameters that corresponds to natural-sounding speech, and to do so in a way that takes into account speaker-specific limitations.

Some voice parameters can be important both for speaker identity as well as for the actual linguistic content of the utterances involved. Vowel formant frequencies, for example,

Talal Bin Amin is a Ph.D. student at the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798 Singapore e-mail: talal1@e.ntu.edu.sg.

Pina Marziliano is an Associate Professor in the Division of Information Engineering at the School of Electrical and Electronic Engineering and James Sneed German is an Assistant Professor in the Division of Linguistics and Multilingual Studies at the School of Humanities and Social Sciences, Nanyang Technological University, Singapore.

are influenced by vocal tract length, and therefore serve as an important cue to a speaker's age and gender. Simultaneously, it is the relative differences between vowel formant frequencies that ultimately creates the distinction between different vowel sounds (e.g., the difference between the vowels in the words "bed" and "bad"). Crucially, this suggests that the space of variation cannot be correctly modeled without taking into account linguistic structure. A second goal of our study, therefore, is to explore the extent to which the variation exhibited across different natural-sounding voice identities depends on linguistic structure in a systematic way. In essence, we seek to test whether the shape of this parameter space is influenced or constrained by specific features of the language involved. More generally, we hope to bring to light previously undocumented challenges faced by current approaches to voice disguise, speaker identification, voice conversion, and speech synthesis, and to identify potential solutions to those challenges.

Studies on voice impersonation are limited [4], [7], [8], [9]. The focus of existing studies has been to determine how closely an impersonator can approximate a target speaker, as well as whether the glottal and vocal tract measures exhibit a close correspondence. Different data sets have led to different observations in this regard. For example, in [7], 30-second excerpts of uninterrupted Swedish sentences were analyzed, while in [8], only two short Japanese sentences were used. In [4], different sentences were used for different target voices, and only one word was common to all sentences and therefore useful for comparison. Additionally, the sentences used in [4] were designed to be humorous and therefore lacked emotional neutrality, a fact which may have confounded or masked the effects of voice identity. In [7], it was concluded that the voice impersonator found it difficult to accurately modify vocal tract characteristics towards the target speaker, whereas in [8] the impersonator was able to modify both the prosodic and vocal tract characteristics towards the target speaker. The different outcomes among these studies may be attributed to the fact that the impersonators had different skill sets, and different voice targets to imitate in different languages. In all four previous studies, the goal of the impersonator was to imitate the voices of specific speakers.

This is in contrast to our study, where the impersonators creatively adapted their voices to produce character voices from their own repertoire. While the impersonators gave labels to some of these voices that were indicative of certain identity traits (e.g., "high pitch female"), they were not given instructions to target specific identities, voices, or identity traits. This allowed the impersonators to more fully express the flexibility of their voices, by impersonating a wide range of voice identities that they were comfortable producing. This in turn allowed us to explore the issues of variation and naturalness rather than similarity to a target speaker. In our previous study [10], we analyzed nine different voices from a single voice impersonator using a single sentence. In this paper, we build upon the previous analysis and better generalize our earlier findings by (i) using three impersonators (including one from the previous study [10]), (ii) using a more comprehensive sample of vowel categories for the analysis of the vocal tract characteristics, and (iii) using a total of 486 sentence tokens for analysis (versus 9 in the previous study).

Crucially, none of the previous studies have investigated how vocal fold behavior changes when an impersonator produces different voices. For languages like English, vocal fold behavior (e.g., creakiness or breathiness) is largely unimportant for word or sentence meaning, though it is known to be associated with social identity traits, especially gender [11]. We therefore hypothesize that our voice impersonators will be able to recruit variation in vocal fold parameters in their attempt to create distinct voice identities. If they cannot, then there is evidence for one or more speaker-specific stable parameters that may be useful for speaker identification or voice disguise identification. Here, we make use of the ElectroGlottoGraphic (EGG) signal, which provides a direct representation of the vocal fold vibration patterns and is free from the filtering effects of the vocal tract. The EGG signal has been found to be independent of vowel category [12] and to depend primarily on the anatomical characteristics of a speaker's vocal folds [13]. While a few studies have shown that the EGG signal can be used for speaker identification [13], this is, to our knowledge, the first study to use EGG signals for the analysis of voice impersonations.

In the first part of our study, we seek to determine which vocal and acoustic parameters the impersonators make use of in order to achieve different voice identities, and to a certain extent, the relationship of those parameters to specific identity traits indicated by the associated labels (e.g., age or gender). The parameters we chose to investigate are in fact largely motivated by what is already believed to have implications for voice identity (e.g., F0 or pitch is indicative of gender since female speakers are generally associated with a higher mean F0 and greater temporal variation than males [14], [15], [16]), though identifying such associations was not the central goal of this study since these are mostly well-known. Instead, we seek to explore the issue of how large the space of variation is within the constraints of naturalness, and how this is influenced by both speaker-specific traits as well as linguistic structure (in this case, the structure of the English vowel inventory). In the second part of the paper, we report the results of a subjective test by naive listeners that provides an estimate of how realistic the impersonated voices were and relates this to their natural voice productions. Since our findings on vowel formants reveal an important effect of vowel category dependency, we introduce a no-reference objective measure for voice disguise that accounts for such effects. The resulting scores of this objective test are then compared against those of the listening test. Section II explains the data collection process, Section III describes each part of the analysis in detail, including additional background, results and preliminary conclusions and Section IV concludes the paper.

II. Data Collection

Three professional voice-over artists (one female, two male) served as the impersonators in this study. We refer to them henceforth as impersonator 1F, 2M and 3M respectively. The first and dominant language of all three impersonators is English, with some differences in dialectal features (South

TABLE I
MEAN AND STANDARD DEVIATION OF F0 IN HERTZ.

(a) Impersonator 1F

Voice	Label	μ	σ
V_8	YM	151.43	33.78
V_5	OM	159.39	47.28
V_1	-	196.75	53.81
V_6	OF	212.72	64.15
V_3	-	258.80	88.27
V_4	YM	266.96	63.66
V_9	YF	274.73	67.81
V_7	YF	294.34	70.99
V_2	-	414.87	88.15

(b) Impersonator 2M

Voice	Label	μ	σ
V_1	-	94.52	25.06
V_6	-	100.49	36.59
V_2	-	103.82	44.53
V_5	-	124.45	28.66
V_7	OM	150.53	34.63
V_3	-	155.73	51.39
V_9	YM	170.39	57.79
V_8	OM	204.14	58.27
V_4	YF	235.85	62.81

(c) Impersonator 3M

Voice	Label	μ	σ
V_8	-	113.73	52.98
V_7	OM	123.75	36.36
V_1	-	125.25	32.60
V_6	-	142.54	36.92
V_9	-	149.67	48.30
V_4	-	166.93	39.94
V_2	-	178.53	35.82
V_5	-	185.98	40.43
V_3	-	309.19	70.24

Asian, Southeast Asian and North American for 1F, 2M and 3M, respectively).

Data collection took place inside a sound-attenuated room, and synchronous speech and EGG signals were recorded from the productions. The speech signal was recorded using an AKG C520L head-mounted condenser microphone. The EGG signals were obtained using a EG2-PCX2 Electroglottogram by Glottal Enterprises [17]. This required placing two electrodes, 35mm in diameter, externally over the larynx in order to measure the electrical conductance of the vocal folds during voiced phonation. The analog speech and EGG signals were captured on separate channels using a Zoom H4n recorder, and were digitized in WAV format at a sampling rate of 44.1 kHz with 16-bit resolution. Following the recording, the speech data was segmented at both the word- and phone- [1] level using the Penn Phonetics Lab Forced Aligner [18]. The results of the automated segmentation were then manually corrected by a trained phonetician.

A. Protocol

The impersonators were given no target speakers to imitate and had the freedom to choose the voices they wanted to impersonate. They were instructed to use a consistent regional variety of English across the nine voices being impersonated, but were given freedom to vary any other identity characteristics of the voices including age and gender. Each voice impersonator produced nine distinct voice identities which included eight (fictional) character voices and their natural voices. Thus a total of 27 distinct voice identities were produced by the voice impersonators. All of these voices were natural-sounding and readily distinguished from each other.

B. Speech material

The same speech materials were used for all 27 voice samples, and consisted of nine short sentences. For each impersonated voice, the impersonators produced the nine sentences in a sequence, and then repeated the sequence in the same voice, for a total of 18 sample sentences per voice. Thus a total of 486 sentences were collected for analysis.

Each sentence included two monosyllabic target words containing one of the vowels /æ/, /ʌ/, /ɪ/, /i/, /u/, and /ɛ/. These target vowels were chosen (a) to provide a representative sample of the overall vowel 'space' of English (i.e., the organization of vowels in the F1-F2 plane, explained in more detail in Section III-C), and (b) because they are relatively robust to subtle differences in regional dialect (e.g., the vowel in 'heard' was excluded on this basis, since American speakers tend to produce it with a stronger 'r'-quality than most British speakers). Factors affecting word prominence, such as sentence stress and phrasing, are known to affect vowel formant measures [19]. To maximize consistency across samples, therefore, target words were placed in positions within the sentence that are associated with maximal prominence. Specifically, the sentences were designed so that target words would be produced with a nuclear accent and occur at the right edge of an intonational phrase boundary.

[1]A phone is a unit of speech, or segment, that can be distinguished on the basis of articulatory, acoustic and perceptual properties. In our data, it is roughly equivalent to an allophone.

III. ANALYSIS AND RESULTS

According to the source-filter theory of speech production [20], the fundamental frequency (F0) and speech rate can be viewed as source characteristics, while the formants reflect the filtering effects of various vocal tract parameters, particularly the positioning of the tongue, lips, jaw, etc. All of these features may be readily extracted from the speech signal. In this section, we present the critical analysis and results for these voice parameters.

A. Glottal measures

*1) Fundamental frequency (*F0): Fundamental frequency (F0) is the acoustic correlate of perceived pitch in speech. Since certain F0 characteristics of speech may vary significantly from speaker to speaker, it is important to consider their relevance for voice identity. A number of studies [14], [15], [16] have investigated the role of mean F0 values for distinguishing the voices of men and women. Overall, the mean F0 tends to be inversely correlated with the length and size of the vocal folds, thus men generally have a lower mean F0 compared to women [21], while adults tend to have a lower mean F0 than children. Additionally, women tend to exhibit a higher degree of temporal variation in F0 than men [22], [23], meaning that there are more frequent peaks and valleys in the temporal F0 contour, and the differences between those peaks and valleys tend to be larger. It is therefore important to consider the extent to which the voice impersonators exploit

TABLE II
MEAN AND STANDARD DEVIATION OF THE SPEECH RATE (SYLLABLES PER SECOND).

(a) Impersonator 1F

Voice	Label	μ	σ
V_5	OM	2.33	0.36
V_7	YF	2.52	0.35
V_6	OF	2.53	0.29
V_3	-	3.48	0.47
V_1	-	3.95	0.61
V_2	-	3.81	0.47
V_9	YF	4.08	0.60
V_4	YM	4.31	0.41
V_8	YM	4.48	0.46

(b) Impersonator 2M

Voice	Label	μ	σ
V_8	OM	3.19	0.48
V_5	-	3.25	0.45
V_6	-	3.31	0.51
V_7	OM	3.46	0.47
V_2	-	3.54	0.53
V_3	-	3.94	0.55
V_1	-	4.17	0.48
V_4	YF	4.19	0.57
V_9	YM	4.22	0.55

(c) Impersonator 3M

Voice	Label	μ	σ
V_7	OM	3.42	0.55
V_1	-	3.70	0.52
V_5	-	3.85	0.64
V_9	-	4.03	0.49
V_6	-	4.06	0.51
V_3	-	4.09	0.65
V_4	-	4.20	0.71
V_2	-	4.53	0.69
V_8	-	4.61	0.75

this variability in F0 characteristics in creating the various voice identities.

For the F0 analyses, Praat [24] was used to first obtain F0 samples at 10 ms intervals, using a frequency window of 75-600 Hz. The mean and standard deviation were estimated from all samples occurring within the voiced portions of all 18 sentences for a given voice. Table-I shows the mean and the standard deviation of F0 for all the voices of the three impersonators arranged by mean F0. The various voice identities are represented by V_i, where i refers to the voice number for that speaker, and V_1 is always the natural voice. In some cases, the impersonators provided labels for the voices that were indicative of either age or gender identity. We indicate this using a combination of the labels "Y" (young), "O" (old), "M" (male), and "F" (female). Speakers chose their voices freely, and numbering was assigned arbitrarily, so there is no correspondence between same-numbered voices across impersonators.

A first glance at Table I reveals that all impersonators were flexible with their pitch in creating different voice identities. The mean F0 exhibited a range of at least one octave across the voices for each impersonator. Not surprisingly, the male voices ranked consistently lower than the female voices both in terms of mean F0 and standard deviation. For the female speaker 1F, the two voices with the lowest mean and standard deviation (V_5 and V_8) are both male, while for the male speaker 2M, the voice with the highest mean and standard deviation i.e.V_4 is the only female voice he produced. While the role of age is less apparent, it can be noted that for 1F, the "old female" (V_6) has the lowest mean and standard deviation among the female voices, while the "old male" is very close to the bottom of the range. 2M and 3M show a similar tendency. A one-way ANOVA confirmed that the effect of voice on mean F0 is significant for all three impersonators ($F(8, 51216)$=7606.244, $F(8, 42224)$=4755.150, $F(8, 43253)$=8367.154; p<0.05 for 1F, 2M and 3M, respectively).

These results confirm our assumption that the impersonators would exploit the stereotypical correspondences between F0 and identity in order to achieve different voice identities. It also illustrates the sense in which variability for a given parameter may be limited on a speaker-specific basis. Even when 1F was using a stereotypically male voice, her mean F0 was higher than the lowest voices for the two male impersonators, 2M and 3M. Similarly, neither 2M or 3M exhibited a mean F0 as high as the maximum for 1F, and their standard deviations were remarkably consistent in being lower than those for 1F. Interestingly, the lowest voice for 2M is his natural voice (on both measures), suggesting that he typically speaks near the bottom of his range, and can only increase both the mean and standard deviation of F0 in order to achieve variation in voice identity.

2) Speech Rate: Speech rate has been linked to a number of stylistic factors, though it can also be related to speaker identity features, including gender and age. Men, for example, generally speak faster than women [25], [26], [27] while young adults tend to speak faster than older adults [26], [27], [28]. We therefore explored the extent to which differences in speech rate were exploited by the voice impersonators in creating different voice identities.

The speech rate, in syllables per second, was calculated for each voice by counting the total number of syllables in each sample and then dividing by the overall duration of all non-silent portions of the sample. Table-II shows the average speech rate and standard deviation for the voices of each impersonator arranged by the average speech rate. All impersonators showed differences in speech rate across the voices of at least 32% (for 2M) and as much as 92% (for 1F). Consistent with earlier studies on age effects, the highest and lowest speaking rates for each impersonator were instantiated by "young" and "old" voices respectively. Additionally, "young" and "old" voices tend to cluster at the top and bottom of the range, respectively, for each impersonator. The exception is V_7 of 1F, which impressionistically sounds like a small child speaking deliberately and somewhat effortfully. The role of gender is less clear. The fastest speaking rate for both 1F and 2M was instantiated by a male voice rather than a female one (as predicted), though overall, the effect of age appears to dominate. Since these labels do not represent controlled variables in the proper sense (e.g., a given "young" voice may not correspond to precisely the same age as another "young" voice), it is not possible to clearly isolate the contribution of gender. Nevertheless, our results confirm the prediction that impersonators use speech rate as an important parameter in the creation of distinct voice identities, and they provide an indication of the amount of variability that is achievable for a given speaker within the bounds of naturalness and individual physical traits. A one-way ANOVA revealed that there was a significant effect of voice on the speech rate for all

three impersonators ($F(8,153)$=53.760, $F(8,153)$=12.800, $F(8,153)$=6.603; p<0.05 for 1F, 2M and 3M, respectively).

B. Measures using the Electroglottogram

The ElectroGlottoGraphic (EGG) signal provides an estimate of the vocal fold contact area [29] by measuring the electrical conductance between the vocal folds. It is useful for analyzing the complex three dimensional movements of the vocal folds since it provides an image of the signal generated at the glottis. Compared to the speech signal, then, the EGG signal is generally free from the filtering effects of the vocal tract. Historically, the EGG signal has been used for detecting voice quality [30] as well as for speaker identification [13]. Some studies including [31] have suggested that speakers do not possess as much voluntary control over their vocal fold behavior as compared to their vocal tract characteristics. The rationale behind using the EGG signals is therefore to explore whether and in what ways the voice impersonators actively exploit differences in vocal fold patterns while impersonating different voices. The measure used for our study was the Open Quotient (OQ), which is directly related to the timing characteristics of the vocal folds, and is described in detail below. A number of studies have reported a correspondence between this measure and various identity features, including age and gender [32], [33], [34], as well as voice quality [30]. On that basis, we predicted that OQ would differ across voices for a given speaker, and that these differences would show an approximate correspondence with the identity labels provided by the impersonators.

For voiced phonation, the vocal folds vibrate in a periodic manner, moving in and out of contact with each other. Thus, the EGG signal also varies periodically as a function of the contact area between the vocal folds. Now consider a vocal fold vibratory cycle in which the vocal folds are initially not in contact, resulting in the electrical conductance being minimum. As the vocal folds begin to move in contact, the electrical conductance starts to increase. The time instant at which the glottis becomes closed is called the Glottal Closing Instant (GCI). The glottal closing is generally abrupt and appears as a steep slope in the EGG signal as shown in Figure 1. It is widely accepted that the GCI appears as a sharp positive peak in the Differentiated ElectroGlottoGraphic (DEGG) signal [29], [35]. The glottis then remains closed for a short period of time before the vocal folds start separating again, causing the measured electrical conductance to decrease. The time instant at which the glottis becomes opened is called the Glottal Opening Instant (GOI). The GOI appears as a low amplitude peak in the DEGG signal with a polarity opposite to that of the GCI peak [35]. The EGG and DEGG signals corresponding to a voiced segment of speech together with the labeled GCIs and GOIs are shown in Figure 1. Using the GCIs and GOIs as two distinct landmarks in the DEGG signal, we can now define some of the glottal parameters, i.e. the open and close phase as shown in Figure 1. The period of time for which the glottis remains closed over a glottal cycle is called the Closed Phase (CP). For the k^{th} glottal period

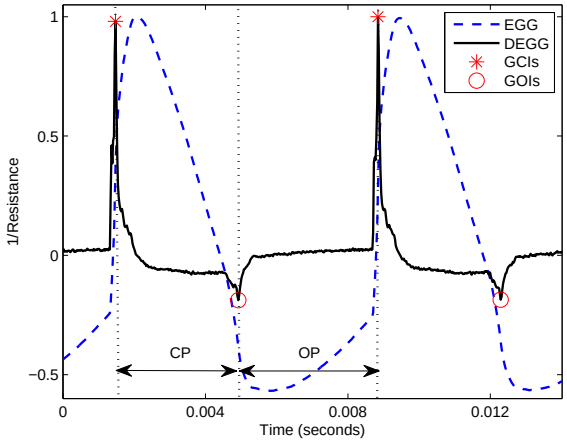

Fig. 1. The EGG and DEGG signals corresponding to a voiced speech segment with the labeled GCIs and GOIs.

T(k), the CP(k) and OP(k) are defined as

$$\text{CP}(k) = \text{GOI}(k) - \text{GCI}(k). \tag{1}$$

The period of time for which the glottis remains opened over a glottal cycle is called the Open Phase (OP). For the k^{th} glottal period T(k), the OP(k) is given as

$$\text{OP}(k) = \text{GCI}(k+1) - \text{GOI}(k). \tag{2}$$

The time period of the k^{th} glottal cycle is then defined as

$$\text{T}(k) = \text{CP}(k) + \text{OP}(k). \tag{3}$$

Once the CP and OP are obtained, we can define the open quotient.

1) Open Quotient (OQ)*:* The OQ represents the percentage of time for which the glottis remains opened over a glottal period. For the k^{th} glottal period, the OQ(k) is defined as

$$\text{OQ}(k) = \text{OP}(k)/\text{T}(k). \tag{4}$$

Various studies have investigated the relationship between the OQ and the perceived age and gender of the speaker. The authors in [32], [33] report that the OQ decreases with increasing age for females, while in [32] the OQ increased with increasing age for males. Since the OQ measure has been found to be mostly independent of the vowel category [12], in our study, all target vowels for a given voice were used to obtain the final estimate of its mean and standard deviation. Table-III shows the mean OQ and its standard deviation for the voices of all three impersonators arranged by mean OQ. The method in [36] was used to estimate the GCIs and GOIs from the DEGG signals.

From Table-III, we find that overall, the "young" voices consistently showed higher OQ means than the "old" voices (in the case of 1F and 2M). Consistent with earlier studies showing an interaction between age and gender, within the young voices, female voices had higher OQ means than male voices, while there was no difference between genders for the two old voices produced by 1F. There was a significant effect of voice on the

TABLE III
MEAN AND STANDARD DEVIATION OF OPEN QUOTIENT

(a) Impersonator 1F

Voice	Label	μ	σ
V_5	OM	0.60	0.11
V_6	OF	0.60	0.20
V_8	YM	0.61	0.13
V_3	-	0.63	0.22
V_4	YM	0.70	0.13
V_7	YF	0.72	0.12
V_1	-	0.72	0.12
V_9	YF	0.73	0.09
V_2	-	0.78	0.09

(b) Impersonator 2M

Voice	Label	μ	σ
V_1	-	0.54	0.14
V_6	-	0.57	0.16
V_8	OM	0.59	0.17
V_7	OM	0.61	0.13
V_5	-	0.65	0.15
V_9	YM	0.75	0.08
V_2	-	0.78	0.12
V_3	-	0.79	0.07
V_4	YF	0.79	0.07

(c) Impersonator 3M

Voice	Label	μ	σ
V_9	-	0.43	0.09
V_4	-	0.46	0.07
V_5	-	0.48	0.05
V_2	-	0.48	0.06
V_1	-	0.49	0.09
V_7	OM	0.52	0.17
V_6	-	0.55	0.14
V_8	-	0.58	0.23
V_3	-	0.83	0.07

mean OQ for all three impersonators ($F(8, 7286)$=203.998, $F(8, 5257)$=319.113, $F(8, 6332)$=1806.637; p<0.05 for 1F, 2M and 3M, respectively). Together, these results show that the impersonators not only have significant voluntary control over their vocal fold patterns, but that they actively manipulated these patterns in order to achieve distinct voice identities.

C. Vocal tract measures

Formant frequencies are identified by the peaks in the spectral envelope of the speech signal, and are determined by the natural resonances of the vocal tract. For a given speaker, changes in formant frequencies depend primarily on changes in the shape and position of the articulators (tongue, lips, jaw, etc.) during speech production. For linguistic purposes, the first two formant frequencies, F1 and F2, are the principle acoustic correlates of perceptual differences among vowel categories, and are also responsible for subtle differences between tokens (spoken instances) of the same vowel. A useful way to visualize relationships between the vowels and formant frequencies is through a two-dimensional Cartesian plot of F1 versus F2, otherwise known as the "vowel space" (see Figure 2 for example). Each data point in the vowel space represents a token of a particular vowel category, and different vowel categories will tend to have distributions of tokens that occupy different regions of the space. The vowel category /i/, for example, tends to occupy a region corresponding to a low F1 and a high F2, while /a/ tends to have a high F1 and a low F2. It is this differentiation that makes it possible for vowel sounds, which are distributed over a continuous space, to be perceived and represented in discrete, categorical terms. Nevertheless, there is typically some overlap between the distributions of neighboring vowels, and the arrangement and positioning of vowels in the vowel space can be sensitive to dialectal [25], stylistic [37], prosodic [19], [25], and importantly for our study, speaker-specific factors (esp. the length and proportioning of the vocal tract) [38].

One important influence of the speaker has to do with the fact that the overall range of formant values depends inversely on vocal tract length. In general, men have a vocal tract that is approximately 20 cm longer than females [39], so it is expected that men have lower overall formant frequencies than females when producing the same vowel [40]. Since vocal tract length increases as children grow, adults generally have lower overall formant frequencies than children. Additionally, speakers may vary in how spread out their vowels are from one another in the vowel space (otherwise known as "dispersion"). For American English speakers, this difference can be recruited as a marker of identity, with female and gay male speakers generally showing higher levels of dispersion than other groups [41], [42]. In short, speakers tend to exhibit substantial variation in the overall positioning of vowels in the vowel space, though the *relative* positions of the vowels tend to be constant for a given language. Given that formant frequencies are an important cue to differences between speakers, they are predicted to be an important source of variability for voice identity construction [43]. We therefore analyzed the key formants (F1 and F2) of six vowel categories in the inventory of English, in order to explore whether the voice impersonators systematically manipulated aspects of the range, positioning, and distribution of the vowels in their attempts to create distinct voice identities.

The Burg method in Praat [24] was used to obtain formant measures by estimating the value of F1 and F2 at the temporal center of each target vowel. A frequency window of 0-5.5 kHz was used with an analysis window length of 25 ms, and the number of poles set at 12. Following extraction, a small number of tokens were identified as having potentially erroneous formant estimates based on what is typical for each vowel. Visual inspection of the Short Time Fourier Transform (STFT) time-frequency distribution (spectrogram) was then used to determine whether each such measurement was indeed erroneous, and to obtain a manual reading using Praat's built-in formant tracking.

Figure 2 shows the vowel space of impersonator 1F for the six target vowels for the nine voices together with the 75% confidence regions for each of the target vowels. A confidence region is a two-dimensional generalization of a confidence interval and is represented as an ellipse placed around the point of central tendency of a distribution. For vowels, a confidence region is useful for visualizing the location and spread of a vowel category, both relative to other vowel categories, and under different conditions. The ellipses in Figure 2 represent the 75% confidence region for each vowel category. Points represent individual vowel tokens, and are color-coded according to vowel category. Different voices are represented by the shape of the points. This plot shows, first of all, that different vowels are subject to different kinds of variability. The vowel /i/ varies mostly in F2, for example, and

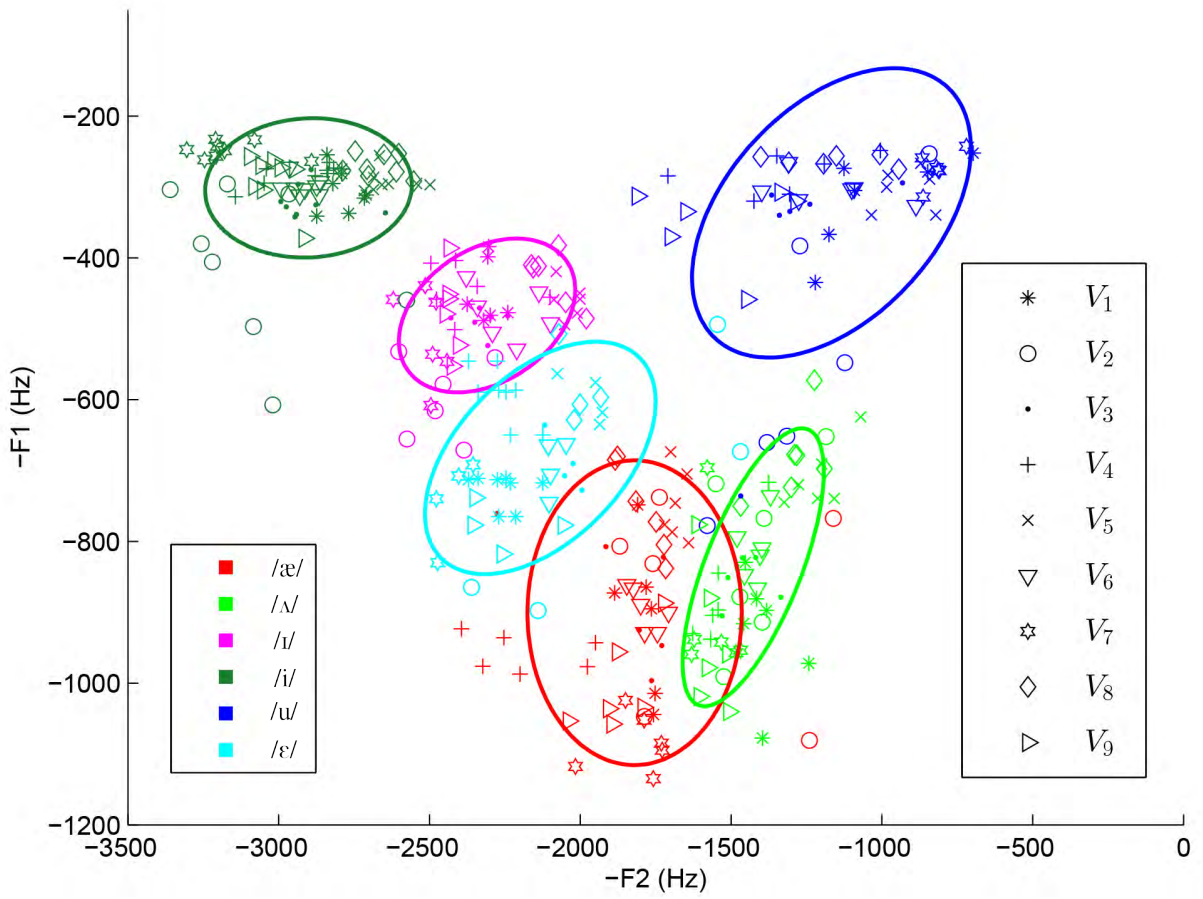

Fig. 2. The vowel space for the nine impersonated voices of impersonator 1F. The vowel ellipses represent the 75% confidence region.

/æ/ varies mostly in F1, while /ɛ/ varies in both dimensions. Some clustering by voice is evident in Figure 2, suggesting that at least some of the within-vowel variation is due to the effects of voice.

To highlight this relationship, Figure 3 shows the same plot with some of the voices omitted for each vowel, and with the multiple tokens for each voice-vowel combination replaced by the corresponding centroid. An arrow points from the centroid for the natural voice to that for each impersonated voice. Here, it can be seen that specific voices are driving much of the variation for specific vowels. The variation in /æ/, for example, is largely driven by V_5 and V_7. Moreover, the same voice may affect the formants in different ways for different vowels. Voice V_7, for example, is clustered in the high end of the F1 range for /i/ and /ɪ/, but in the low end of the F1 range for /u/. Vowel tokens from the voice impersonator's natural voice V_1 generally fall close to the center of each distribution. Figure 4 and Figure 5 similarly show the natural voice for 2M and 3M, respectively, along with a selection of two other voices. Again we find that the natural voice V_1 of 2M and 3M is roughly in the middle of each ellipse, and that the impersonated voices tend to deviate from the center in specific ways. The hypothesis that voices have a significant effect on the vowel formant values was confirmed by a two-factor (voice by vowel) MANOVA analysis (following [42]) which showed that there was a significant voice and vowel interaction effect for F1 ($F(40, 270)$=4.882, partial η^2=0.420; $F(40, 270)$=3.654, partial η^2=0.351 and $F(40, 270)$=5.322, partial η^2=0.441; p<0.05 for 1F, 2M and 3M, respectively) and F2 ($F(40, 270)$=5.902, partial η^2=0.466; $F(40, 270)$=3.925, partial η^2=0.368 and $F(40, 270)$=2.846, partial η^2=0.297; p<0.05 for 1F, 2M and 3M, respectively). Crucially, the effect of each voice cannot be characterized in a general way for all vowels. It is not the case, for example, that the differences among the voices can be characterized in terms of a wholesale shift in F1 or F2 across all vowels. Nor can the differences be characterized in terms of vowel space dispersion (i.e., expansion/contraction relative to the center of the space). If that were the case, then the effect of a specific voice would be to shift all vowels universally either towards or away from the center of the space relative to the natural voice, which is clearly not the case for 3M. Instead, there is an interaction between voice and vowel category, such that the effect of a particular voice on the average formant values depends on the vowel category in question. In other words, the impersonators are making adjustments to formants

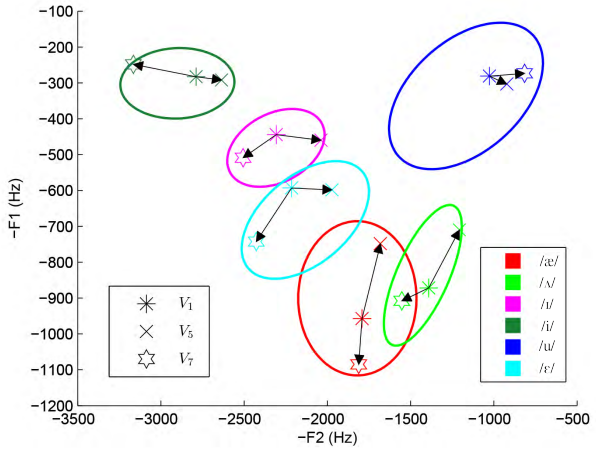

Fig. 3. The vowel ellipses for the target vowels for some voices of impersonator 1F. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

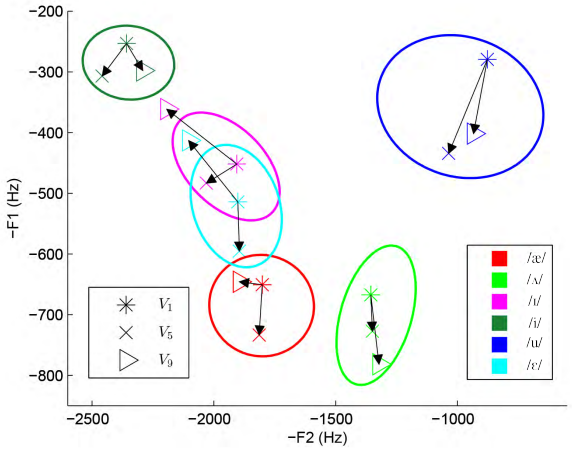

Fig. 5. The vowel ellipses for the target vowels for some voices of impersonator 3M. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

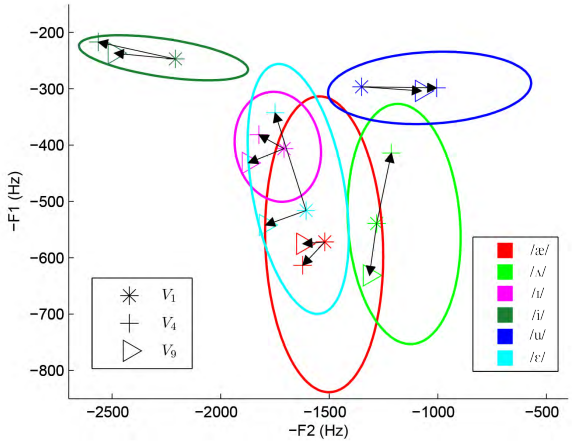

Fig. 4. The vowel ellipses for the target vowels for some voices of impersonator 2M. These voices are driving the variance for the different vowel categories indicating the relationship between voice identity and formants.

on a vowel-by-vowel basis, suggesting that any successful model of either naturalness or disguise identification will at a minimum need to treat formant measures in conjunction with a linguistic parse that includes vowel category.

A visual inspection of Figure 2 suggests that the natural voice tokens of each vowel are associated with lower variance (i.e., greater clustering) than those from the fictional voices. This was partly confirmed by the results of the MANOVA, which showed that the average variances in F1 across all vowels were among the lowest for the natural voices of each speaker (third, first and second lowest out of nine total voices for 1F, 2M and 3M, respectively). For F2, the average variances for the natural voices were somewhat higher (fourth, fourth and fifth out of nine total voices for 1F, 2M and 3M, respectively). Given the tilted orientation of the confidence regions for each vowel, however, we speculated that the apparent low variance of the natural voices could be captured more adequately by a measure that first decorrelates the formant measurements in the vowel space. In Section III-D2, we show that by addressing the issue of the orientation of vowel variance through Principal Component Analysis, it is possible to both capture the intuitive sense in which natural voices are associated with lower variance in the vowel distribution, and that these distributions provide a useful metric for voice disguise identification.

Together, these results show that the high degree of variability in formants is a major resource that the impersonators exploit to achieve different voice identities. In order to capture such differences in any model, however, it is essential to take account of the implicit linguistic structure that underlies the organization of the overall vowel space. In our study, the impersonators were able to rely on their own implicit linguistic knowledge of which combinations of vowel formant frequencies are both permissible (in terms of perceptual distinctness, e.g.) and also natural-sounding. Any automated system would need to model these aspects of human linguistic competence explicitly.

D. Voice impersonation as a case of voice disguise

The field of automatic speaker recognition has seen significant improvements in recent years, though it still suffers from several limitations. In the context of forensics, recognition is still done manually by phoneticians using aural-spectrographic inspection, a method which is both highly labor intensive and subject to bias. This is largely due to the fact that automated systems still do not adequately address the problem of voice disguise [2], [44]. Most forensic cases of speaker identification in fact involve a criminal disguising his or her identity through voice disguise. A disguise identification system can therefore provide a front end to a speaker recognition system by giving a probability estimate that the voice is disguised before actually attempting to determine the speaker's identity. Such a system has the potential to both conserve resources and facilitate automation of speaker recognition more generally. All current methods of automated voice disguise detection [45], [46] that

we are aware of use machine learning methods and require large amounts of training data to achieve reasonable process. However, in real forensic scenarios the amount of available data is usually very limited. Thus there exists a need for an objective metric that evaluates voice disguise without requiring any training data or a reference.

In order to address these issues, we first conducted a test based on subjective human evaluations of voice disguise. This test provides an estimate of how successful our impersonators were in providing natural-sounding voices, and serves as a baseline for comparison with objective measures. Then, taking inspiration from the results presented in Section III-C, we propose a new no-reference objective metric that relies on the distributions of individual vowel categories to provide a disguise rating for a given voice. The results of the two sets of ratings are compared in order to evaluate the potential for the new objective metric in an automated voice disguise recognition system.

1) Subjective evaluation: The subjective test obtained judgements from naive human listeners regarding whether or not a given voice sample was disguised. For this test, the entire database was divided into nine different lists, with equal numbers of samples from each impersonator, voice, and sentence. Each sentence appeared three times in a list, but never more than once by the same impersonator in the same list. The order of samples was pseudorandomized so that the minimum average distance between samples from the same impersonator was 2.0.

A total of 18 listeners participated in the study (ages 22-45, balanced bilinguals in English and one other language). Samples were presented one at a time through headphones using Psychtoolbox [47]. The listeners were asked the following question: *Is the voice disguised or not?*. They responded by clicking "Yes" or "No" on a computer screen. Table-IV shows the percentage of subjects who rated each voice as non-disguised. The voices for all three impersonators are included together, so subscripts are used to indicate the impersonator (right-hand numeral) and voice number (left-hand numeral). V_{32}, for example, refers to the third voice of impersonator 2M. The voices are arranged by their objective rating scores (described below), which are presented in parallel. Overall, the table confirms the prediction that the natural voices (in bold) would receive very high scores. V_{11} and V_{12} are tied with three other voices for the highest score (94.4%), while somewhat surprisingly, V_{13} was judged to be natural only 77.8% of the time. For the impersonated voices, listeners correctly judged these as disguised only 56% of the time, which is only somewhat better than chance. It is important to note that impersonators were not specifically instructed to avoid disguise detection, thus this test provides an estimate of the lower bound on the ability of impersonators to deceive human listeners.

2) Objective evaluation: The objective metric of voice disguise that we present here is motivated by the results of Section III-C which showed that there exists an interaction between voice and vowel category. Specifically, the impersonators made changes to the first two formant frequencies on a vowel-by-vowel basis, suggesting that any successful automated system needs to be sensitive to the linguistic parse. Our results suggested that the vowels associated with the impersonators' natural voices tended to exhibit less variability than the artificial (or disguised) voices, thus we speculated that higher variability might be an important feature associated with voice disguise. This idea is supported by a number of findings in the linguistics literature showing that variability in speech production is closely tied to routinization and practice [48], [49]. In [50], this prediction is generated directly from general facts about the organization of the phonological grammar. In short, the impersonators are less practiced with the vowel patterns associated with their artificial voices, and should therefore exhibit more variability in those patterns. We therefore developed an objective metric based on the first two vowel formants that assesses variability across the vowels associated with a voice, but in a way that does not depend on the same *type* of variability occurring in different vowels (e.g., a systematic increase in F1, or a systematic movement towards or away from the center of the vowel space). To accomplish this, the metric makes reference to the linguistic (phonemic) parse, and in doing so remains robust to the voice-by-vowel interaction observed in our earlier findings.

TABLE IV
SUBJECTIVE AND OBJECTIVE RATINGS OF THE VOICES OF THE THREE IMPERSONATORS.

Voice	γ	Subjective Rating (%)
V_{63}	0.8551	83.3
$\boldsymbol{V_{11}}$	**0.8215**	**94.4**
$\boldsymbol{V_{12}}$	**0.8204**	**94.4**
V_{32}	0.8122	94.4
$\boldsymbol{V_{13}}$	**0.8101**	**77.8**
V_{81}	0.8040	72.2
V_{52}	0.7991	16.7
V_{22}	0.7889	33.3
V_{23}	0.7887	38.9
V_{43}	0.7868	61.1
V_{92}	0.7778	94.4
V_{62}	0.7740	5.6
V_{41}	0.7685	66.7
V_{83}	0.7648	66.7
V_{91}	0.7631	94.4
V_{31}	0.7601	22.2
V_{33}	0.7523	0.0
V_{53}	0.7451	22.2
V_{61}	0.7449	44.4
V_{82}	0.7349	5.6
V_{71}	0.7294	27.8
V_{51}	0.7283	22.2
V_{21}	0.7253	22.2
V_{72}	0.7237	16.7
V_{42}	0.6945	5.6
V_{93}	0.6908	55.6
V_{73}	0.6793	27.8

The metric we propose is based directly on the distribution of vowels for a single voice in the F1-F2 plane and is calculated as follows:

Let $\mathbf{F}$ be a matrix which contains the formant values associated with a voice for a vowel v. It is defined as

$$\mathbf{F} = \begin{bmatrix} f_{1,1} & f_{2,1} \\ f_{1,2} & f_{2,2} \\ \vdots & \vdots \\ f_{1,n} & f_{2,n} \end{bmatrix} \tag{5}$$

where the columns of $\mathbf{F}$ represent the F1 and F2 values in hertz respectively and the rows represent tokens (samples). Each column of $\mathbf{F}$ is zero mean centered. Then by applying Principal Component Analysis (PCA) we find the reconstruction of $\mathbf{F}$ in the principal component space denoted as $\hat{\mathbf{F}}$. Let σ_1 and σ_2 denote the standard deviation along the two columns of $\hat{\mathbf{F}}$, where $\sigma_1 > \sigma_2$. The deviation factor α for the vowel v is then defined as

$$\alpha_v = \frac{\sigma_1}{\sigma_1 + \sigma_2} \tag{6}$$

where $0.5 < \alpha_v < 1$.

To obtain the overall disguise score for a voice, the deviation factor α_v is first obtained for N vowels. The disguise score γ for a voice is then defined as the average value of α_v over the N vowels

$$\gamma = \sum_{i=1}^{N} \frac{\alpha_{v_i}}{N} \tag{7}$$

Our analysis uses the same set of vowels from Section III-C, namely, $v = \{$/æ/, /ʌ/, /ɪ/, /i/, /u/, /ɛ/$\}$.

The disguise score γ provides an estimate of how much a voice varies along the first principal component compared to the total variation along both the principal components. A value of γ near 1.0 is predicted for undisguised (i.e., natural) voices, while a value of γ near 0.5 is predicted for poorly disguised voices. Table-IV shows the objective rating score for each voice ordered by the value of γ, alongside the subjective ratings from the previous section. As predicted, the natural voices (in bold) are highly ranked. Interestingly, the natural voice for 3M is ranked more highly by γ than by the human raters, suggesting a possible human bias to which it is immune. The table also suggests a good correlation between the objective metric and the subjective ratings. The Spearman rank order correlation coefficient [51] was calculated to determine the relationship between γ and the subjective ratings of disguise. This statistic measures the strength of a monotonic relationship between two ordinal variables. This test revealed a "strong", positive correlation between γ and the subjective ratings, which was also statistically significant ($r_s(25)$=$0.6542, p$=8.0144×10^{-4}).

Overall, these results indicate that γ is a highly useful metric for automated voice disguise identification applications. This metric makes use of vowel formants in combination with a linguistic parse, two factors which are generally ignored by automated speaker recognition systems, but which are essential for phoneticians in their manual analyses. In doing so, it provides an assessment of disguise that is highly comparable to that of human listeners, and may even outperform them in certain cases.

IV. Conclusion

In this study, we analysed the speech of three voice impersonators producing a total of 27 different voice identities. These analyses confirmed that the impersonators were able to exploit differences in mean F0, speech rate, vocal fold patterns (Open Quotient), and vowel formant distributions in order to create the various voice identities. This is the first study we know of that shows the ability of voice impersonators to exploit differences in vocal fold patterns, and the only study that considers a comprehensive set of speech parameters in a single study based on voice impersonation. Our study also sought to explore the space of variability that is possible for the various speech parameters given the impersonators' sensitivity to naturalness constraints and their inherent physiological limitations. On the one hand, by eliciting a wide range of voice identities from the impersonators' repertoire, our study provides an estimate of the size of the range that is possible for each parameter for a given speaker. Additionally, our study showed speaker-by-speaker limitations, since, for example, 1F was unable to achieve F0 means comparable to the two male speakers even when impersonating males.

Our analysis of the effects of voice identity on vowel formant measures revealed that while impersonators exploit variation in vowel formants, they do so in a way that is sensitive to linguistic structure. Specifically, they make changes to formant distributions on a vowel-by-vowel basis, rather than by systematically shifting the entire vowel space along some dimension, or by expanding or contracting the vowel space. We noted that this has important consequences for automated disguise detection systems, since it suggests that such systems cannot do without a linguistic parse. This is the first study we know of to show that the modification of vowel formants by impersonators is sensitive to, and constrained by, the specific structure of the linguistic system (i.e., the language) involved.

In our study, we attempted to isolate the effects of voice identity on vowel formants by observing target words under consistently prominent prosodic conditions (nuclear accented, intonation phrase-final). An anonymous reviewer points out that vowel reduction effects based on differences in prosodic prominence are likely to contribute substantial variability to vowel formant distributions, thus complicating the analysis. Since prosodic factors tend to affect vowel formants more systematically by way of expansion and contraction of the overall vowel space [19], we speculate that the ultimate solution will need to treat these two factors independently.

We presented an objective metric for detecting voice disguise. This metric not only rated the impersonators' natural voices very highly, but it exhibited a strong correspondence with the subjective ratings obtained from human listeners (even outperforming them in one instance). The disguise estimates provided by this metric can be exploited for making current speaker recognition systems more robust to attacks of voice disguise, especially when there is no a priori information available as to the type of disguise. In future research, it will be important to evaluate the improvement in performance of a speaker recognition systems by adding such a front end voice disguise system based on the disguise metric γ. In our study, we selected a subset of vowels and treated the contribution of all vowel categories equivalently. Another important issue for future research to explore, therefore, is how the number, type, and weighting of vowels used in calculating γ affects its performance in discrimination tasks. The application of our findings on vowel formant variability to the development of a disguise metric is highly suggestive of how the findings may be further utilized. Ultimately, an automated disguise detection or speaker identification system that considers variability across

both glottal and vocal tract parameters is likely to achieve additional gains in reliability.

V. Acknowledgements

The authors would like to thank the voice-over artists: Noella Menon, Marc X Grigoroff and Rishi Budhrani; Charmaine Hon for helping with the manual segmentation of the data and Ng Chen Yi for the programming of the subjective experiment.

References

[1] H. Hollien, *Forensic voice identification*. Academic Press, 2002.

[2] J. Bonastre, F. Bimbot, L. Boë, J. Campbell, D. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 33–36.

[3] Y. Stylianou, "Voice transformation: A survey," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3585–3588.

[4] E. Zetterholm, "Same speaker–different voices. a study of one impersonator and some of his different imitations," in *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, 2006, pp. 70–75.

[5] M. Farrús, M. Wagner, D. Erro, and J. Hernando, "Automatic speaker recognition as a measurement of voice imitation and conversion," *The International Journal of Speech, Language and Law*, vol. 17, no. 1, pp. 119–142, 2010.

[6] A. Machado and M. Queiroz, "Voice conversion: A critical survey," *Proc. Sound and Music Computing (SMC)*, 2010.

[7] A. Eriksson and P. Wretling, "How flexible is the human voice?-a case study of mimicry," in *Fifth European Conference on Speech Communication and Technology*, 1997, pp. 1043–1046.

[8] T. Kitamura, "Acoustic analysis of imitated voice produced by a professional impersonator," in *Proc. Interspeech*, 2008, pp. 813–816.

[9] E. Zetterholm, "Impersonation: a phonetic case study of the imitation of a voice," *Lund Working Papers in Linguistics*, vol. 46, pp. 269–287, 2009.

[10] T. B. Amin, P. Marziliano, and J. S. German, "Nine Voices, One Artist: Linguistic and Acoustic Analysis," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2012, pp. 450–454.

[11] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, no. 3, pp. e111–e116, 2012.

[12] M. Epstein, "Voice quality and prosody in English," Ph.D. dissertation, University of California, 2002.

[13] W. Campbell, T. Quatieri, J. Campbell, and C. Weinstein, "Multimodal speaker authentication using nonacoustic sensors," in *Workshop Multimodal User Authentication*, 2003, pp. 215–222.

[14] J. Hillenbrand and M. Clark, "The role of f_0 and formant frequencies in distinguishing the voices of men and women," *Attention, Perception, & Psychophysics*, vol. 71, no. 5, pp. 1150–1166, 2009.

[15] D. Honorof and D. Whalen, "Identification of speaker sex from one vowel across a range of fundamental frequencies," *The Journal of the Acoustical Society of America*, vol. 128, no. 5, pp. 3095–3103, 2010.

[16] M. Sambur, "Selection of acoustic features for speaker identification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 2, pp. 176–182, 1975.

[17] M. Rothenberg, "A multichannel electroglottograph," *Journal of Voice*, vol. 6, no. 1, pp. 36–43, 1992.

[18] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics*, 2008, pp. 5687–5690.

[19] E. Lee and J. Cole, "Acoustic effects of prosodic boundary on vowels in American English," in *Proceedings of the* 42^{nd} *Annual Meeting of the Chicago Linguistic Society*, Illiniois, Chicago, 2006.

[20] G. Fant, *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*. Mouton & Co. N.V., The Hague, 1970.

[21] M. Latinus and P. Belin, "Human voice perception," *Current Biology*, vol. 21, no. 4, pp. R143–R145, 2011.

[22] J. van Rie and R. van Bezooijen, "Perceptual characteristics of voice quality in dutch males and females from 9 to 85 years," in *Proceedings of the XIIIth International Congress of Phonetic Sciences 2*, 1995, pp. 290–293.

[23] R. Brend, "Male-female intonation patterns in american english," *Language and sex: Difference and dominance*, vol. 86, 1975.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 5.3," Retrieved October 21, 2011, from http://www.praat.org/.

[25] D. Byrd, "Relations of sex and dialect to reduction," *Speech Communication*, vol. 15, no. 1-2, pp. 39–54, 1994.

[26] J. Yuan, M. Liberman, and C. Cieri, "Towards an integrated understanding of speaking rate in conversation," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 541–544.

[27] E. Jacewicz, R. Fox, C. O'Neill, and J. Salmons, "Articulation rate across dialect, age, and gender," *Language variation and change*, vol. 21, no. 02, pp. 233–256, 2009.

[28] B. Smith, J. Wasowicz, and J. Preston, "Temporal characteristics of the speech of normal elderly adults," *Journal of Speech and Hearing Research*, vol. 30, no. 4, pp. 522–529, 1987.

[29] D. Childers and A. Krishnamurthy, "A critical review of electroglottography." *Critical reviews in biomedical engineering*, vol. 12, no. 2, p. 131, 1985.

[30] A. Fourcin, "Voice quality and electrolaryngography," in *Voice Quality Measurement*, R. Kent and M. Ball, Eds. San Diego: Singular Publishing Group, 2000.

[31] A. Neocleous and P. Naylor, "Voice source parameters for speaker verification," in *Proc. Eur. Signal Process. Conf*, 1998, pp. 697–700.

[32] M. Higgins and J. Saxman, "A comparison of selected phonatory behaviors of healthy aged and young adults," *Journal of Speech, Language and Hearing Research*, vol. 34, no. 5, pp. 1000–1010, 1991.

[33] R. Winkler and W. Sendlmeier, "Open quotient (EGG) measurements of young and elderly voices: Results of a production and perception study," *ZAS Papers in Linguistics*, vol. 40, pp. 213–225, 2005.

[34] E.-M. Ma and A. 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.

[35] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *The Journal of the Acoustical Society of America*, vol. 115, p. 1321, 2004.

[36] T. B. Amin and P. Marziliano, "Glottal activity detection from differentiated electroglottographic signals using finite rate of innovation methods," submitted for publication.

[37] S.-J. Moon and B. Lindblom, "Interaction between duration, context, and speaking style in English stressed vowels," *The Journal of the Acoustical Society of America*, vol. 96, p. 40, 1994.

[38] K. Stevens, *Acoustic phonetics*. The MIT press, 2000, vol. 30.

[39] G. Fant, "A note on vocal tract size factors and non-uniform f-pattern scalings," *Speech Transmission Laboratory Quarterly Progress and Status Report*, vol. 1, pp. 22–30, 1966.

[40] G. Peterson and H. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[41] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, no. 3, pp. 255–272, 1996.

[42] J. B. Pierrehumbert, T. Bent, B. Munson, A. R. Bradlow, and J. M. Bailey, "The influence of sexual orientation on vowel production (l)," *The Journal of the Acoustical Society of America*, vol. 116, p. 1905, 2004.

[43] R. Coleman, "A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice," *Journal of Speech and Hearing Research*, vol. 19, no. 1, pp. 168–180, 1976.

[44] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[45] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automatic detection: review and perspectives," in *Progress in nonlinear speech processing*. Springer, 2007, pp. 101–117.

[46] L. Mary, K. A. Babu, and A. Joseph, "Analysis and detection of mimicked speech based on prosodic features," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 407–417, 2012.

[47] D. H. Brainard, "The Psychophysics Toolbox," *Spatial vision*, vol. 10, no. 4, pp. 433–436, 1997.

[48] J. Edwards, M. E. Beckman, and B. Munson, "The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition," *Journal of Speech, Language and Hearing Research*, vol. 47, no. 2, p. 421, 2004.

[49] B. Munson, "Phonological pattern frequency and speech production in adults and children," *Journal of Speech, Language and Hearing Research*, vol. 44, no. 4, p. 778, 2001.
[50] J. S. German, K. Carlson, and J. B. Pierrehumbert, "Reassignment of consonant allophones in rapid dialect acquisition," *Journal of Phonetics*, vol. 41, no. 3, pp. 228–248, 2013.
[51] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

Talal Bin Amin received the B.Sc. degree in Communication Systems Engineering from the Institute of Space Technology, Pakistan in 2008 and the M.Sc. degree in Signal Processing from the Nanyang Technological University, Singapore in 2009. In 2010, he joined the Data Mining Department at the Institute for InfoComm Research (I2R), Singapore as a Computer Programmer before joining the Laboratory of Audio and Visual Perception at the Duke NUS Grduate Medical School, Singapore as a Research Assistant. Since August 2010, he is pursuing the Ph.D. degree in the Division of Information Engineering at the Nanyang Technological University, Singapore.

His research interests include forensic speaker recognition, voice impersonation and voice disguise detection.

Pina Marziliano obtained a B.Sc. Applied Mathematics in 1994 and the M.Sc. Computer Science (Operations Research) in 1997, both, from the Université de Montréal, Canada. In 2001 she completed her Ph.D degree in Communication Systems from the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. She then joined a start-up company, Genimedia SA in Lausanne, Switzerland and developed perceptual quality metrics for multimedia applications. In 2003, she became an Assistant Professor in the School of Electrical and Electronic Engineering at the Nanyang Technological University in Singapore, and was tenured and promoted to Associate Professor in August 2012. In October 2012, she co-founded and is a Director of PABensen Pte. Ltd. Where Art and Science Design®, a Singapore based design company engaged in developing innovative artistic objects and functional products that bridge the fields of art, science and technology.

In 2007, she received the 2006 Best Paper Award from the IEEE Signal Processing Society Awards Board for the paper "Sampling Signals with Finite Rate of Innovation" co-authored with Prof. M. Vetterli and Dr. T. Blu which appeared in IEEE Trans. Signal Processing, Vol. 50, June 2002. A patent for her work related to sampling was granted in May 2006 and then acquired by Qualcomm Inc. USA in December 2007.

She has been an Associate Editor for IEEE Signal Processing Letters and IEEE Transactions on Signal Processing since January 2010 and February 2013, respectively. She is also a Member of the IEEE Signal Processing Society Signal Processing Theory and Methods Technical Committee since January 2012 and has served on the Technical Program Committee of numerous international conferences.

She currently teaches undergraduate and graduate courses in Digital Signal Processing and her research interests include sampling theory and applications in communications and biomedical engineering, information security and perceptual quality metrics for multimedia.

James Sneed German received the Ph.D. in linguistics from Northwestern University in 2009. He then spent two years as a Postdoctoral Researcher at the Laboratoire Parole et Langage (CNRS) in Aix-en-Provence, France. Since 2010, he has been an Assistant Professor in the Division of Linguistics and Multilingual Studies at Nanyang Technological University, Singapore.

His research interests cover the cognitive architecture of linguistic sound systems, as well as prosody and the role it plays in signalling both literal and non-literal meaning.