Voice disguise by mimicry: deriving statistical articulometric evidence to evaluate claimed impersonation

Rita Singh[1,*], Abelino Jiménez[2], Anders Øland[3]

[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, USA
[3]Computer Science Department, Carnegie Mellon University, Pittsburgh, USA
[*]rsingh@cs.cmu.edu

Abstract: Voice disguise by impersonation is often used in voice-based crimes by perpetrators who try to evade identification while sounding genuine. Voice evidence from these crimes is analyzed to both detect impersonation, and match the impersonated voice to the natural voice of the speaker to prove its correct ownership. There are interesting situations, however, where a speaker might be confronted with voice evidence that perceptually sounds like their natural voice but may deny ownership of it, claiming instead that it is the production of an expert impersonator. This is a bizarre claim, but plausible since the human voice has a great degree of natural variation. It poses a difficult forensic problem: instead of detecting impersonation one must now prove the *absence* of it, and instead of matching the evidence with the natural voice of the person one must show that they cannot *not* have a common originator. In this paper we address the problem of disproving the denial of voice ownership from an articulatory-phonetic perspective, and propose a hypothesis testing framework that may be used to solve it. We demonstrate our approach on data comprising voices of prominent political figures in USA, and their expert impersonators.

1. Introduction

Recent political stories [1, 2] have highlighted a largely ignored challenge in voice forensics – that of *denial* of voice ownership. A voice recording appears to be the voice of a speaker "Q". Q, however, denies authorship of the recording, claiming it to be the artwork of an expert impersonator. How do we prove or disprove the denial? This is the problem we address in this paper.

Voice impersonations are possibly among the oldest forms of mimicry – and forgery. In the Odyssey (Homer, 850 BC) Helen of Troy, suspecting treachery, allegedly circled the wooden horse, calling out to various Greek soldiers in the voices of their wives and sweethearts. This story is remarkable for two reasons. It is one of the oldest recorded references to voice impersonation. It is also possibly the oldest reference to voice impersonation for the purpose of *deception*. Happily for the invading Greeks, the hero Odysseus, recognizing the deception, prevented his soldiers in the horse from responding, avoiding an immediate grisly end to the legend. References to voice impersonations abound in various mythologies, works of literature, and historical record. They have been performed for entertainment, deception, or even depictions of deception for entertainment (as when a character mimics another in a play). Their popularity has only increased with time – a YouTube search reveals literally hundreds of people, both expert and amateur, attempting to

impersonate dozens of celebrities, political figures, lay persons, or even unfortunate victims of harassment or bullying. Impersonation has in fact been deeply studied in form and implication in the literature within multiple disciplines ranging from science and philosophy to politics and drama. Of special relevance to the subject of this paper is the conundrum posed by self-impersonation [3]: *is this me masquerading as myself, or is this someone masquerading as me?*

1.1. The components of a successful impersonation

At the outset it is important to understand what comprises successful impersonation. Regardless of whether it is performed for entertainment or deception, the objective of the impersonation in all of the contexts mentioned above is to *sound* like (a possibly caricatured version of) the target of the impersonation. Since the primary objective is perceptual, the mimicry itself tends to focus on, or to even overemphasize, perceptually distinctive aspects of the target's speech rather than the fine detail. For example, Zetterholm [4] reports that impersonators frequently mimic or caricature pitch register, voice quality, dialect, prosody and speech style, albeit to different degrees of success. Indeed, it may be physically impossible to replicate all the fine detail of the target speaker's speech. Eriksson and Wretling [5] report that mimics successfully imitate fundamental frequency patterns and the global speech rate of the target, but often fail to replicate segmental timing patterns. Mimics are also reportedly able to imitate key formant patterns [6, 7] for several phonemes; however we find no reports of successful imitation of formant *bandwidths*, which are an intrinsic characteristic of the speaker's vocal tract.

The distinction between the success in imitating high-level patterns and the relative intransigence of fine-level detail in being replicated may be attributable to several reasons. Primary, of course, is that of intent. From the qualitative standpoint of a human listener, a successful impersonation is that which creates an impression of the target that is sufficient for the listener to believe that the voice is indeed that of the target speaker. Once this objective is achieved any additional effort in imitating fine detail is unnecessary. Thus, most successful impressions are auditory illusions generated by the impersonator, and are created by taking advantage of the limitations and shortcomings of human hearing and cognition [4, 8]. The secondary reason, *and what makes aural illusions necessary for mimicry*, is biological. The human voice is the output of a very complex vocal apparatus, many aspects of which are very specific to the speaker. It is physically impossible for the mimic, with an entirely different vocal apparatus, to replicate every detail of the speech of the target speaker. The constraints are not merely physical – they are also cognitive. McGettigan [9] reports from studies of brain images that mimicry is a deliberate activity invoking regions of the brain that are not activated similarly during natural speech. Thus, even cognitively, mimicry is not a natural act and, regardless of training and skill, will not duplicate natural speech.

It must be clarified that the above discussion deals mainly with mimicry that has the implicit or explicit intent of deception. There are other forms of mimicry that occur naturally in human speech, such as children mimicking adults, or in conversational phenomena such as convergence and accommodation [10]. These are not the subject of this paper.

1.2. Speaker-verification systems and impersonated speech

Human speech, including mimicked speech, is an active process, invoking cognitive processes that adapt continuously to feedback [11]. We may hence hypothesize that even for speech characteristics that the mimic *does* successfully imitate from the target, the mimic will revert, at least partially, to his or her "natural" mode of speaking when doing so does not appear to affect the

perceived quality of the mimicry. Thus, the *variation* in the patterns of even these well-mimicked features will be greater than when either the target or the mimic speak in their natural voice. This has a forensic consequence. Voice impersonation has also been suggested (and attempted) as a way to break into voice authentication systems. However, although expert mimics may deceive human listeners, they cannot similarly deceive state-of-art speaker-verification systems which consider the aggregate distribution of the features in the speaker's speech signal. Mariéthoz and Bengio [12] demonstrate that even a simple GMM-based verification system cannot be deceived by expert voice mimics. Hautamäki et. al. [13] show that the detection-error tradeoff curve of a state-of-art speaker-verification system stays essentially unchanged when the impostor is an expert voice mimic. As a matter of fact, even *human* listeners who may be deceived when they listen to impersonations by the mimic in isolation will not be deceived when they can also hear the actual speech by the target for comparison [14]. Hence, the risks posed to voice-based authentication systems by voice mimicry are not great. In fact, the most successful attempts at breaking into voice-authentication systems have been with *synthetic* voices, obtained either by direct synthesis or by modifying an actual voice [15], although, paradoxically, these voices would not fool a human ear.

1.3. The problem and our approach to its solution

With this understanding we are now in a position to examine the problem of denial of voice ownership that is addressed in this paper. This problem is diametrically opposite to the forensic challenge of detecting (and rejecting) an impersonator who *claims to be the target*. Here we have been presented with a recording has been accepted by both human experts and automated systems as belonging to the subject, but the subject is denying authorship, claiming it to be the work of an expert impersonator. We must accept or reject this claim.

In spite of this reversal of perspective, the problem may still be approached as one of verification. The obvious solution is to "tighten" the threshold of acceptance in a state-of-art statistical speaker-verification system until the probability of false acceptance is sufficiently small. If the recording is nevertheless accepted by this system, the denial of authorship can be rejected. While this solution is perfectly acceptable, it has its drawbacks. The system can reject true recordings (and accept their denial) by the speaker for many reasons including channel noise, variations in the subject matter of the speech, etc. On the other hand, even if the denial is rejected by the test (i.e. the verification system accepts the recording as authentic), the target may counter that the recording is by a very skilled impersonator who has fooled the system which only considers statistical aggregate information. As additional evidence against the counter-claim, we must hence demonstrate that the recording does not exhibit any of the defects that are inherent to mimicry.

In this paper we build a solution with these constraints in perspective. Our solution is predicated on the goal of explicitly identifying and characterizing *defects* of impersonated speech that a) are not attributable to normal variations in the speaker's natural speech alone and b) can potentially be employed to reject the claim of expert impersonation of the speaker's voice.

To this end, we show that attempts at impersonation even by *expert* impersonators working under clean conditions have *fine-detail* features that show statistically significant differences from the target speaker – differences which cannot, from the experimental setup, be attributed to poor impersonation or noisy recordings. We have previously shown that under similar conditions even one of the world's best impressionists cannot modify his formant loci for specific phonetic contexts, regardless of who he tries to imitate [16]. Here we show the reverse – that there are some patterns by the target speaker that are simply not replicated by even the expert impersonators.

2. Articulometric basis and features for testing

To study the differences and similarities between impersonated and natural voices we propose an approach that is based on articulatory-phonetic guidelines.

Speech signals carry the resonances of the speaker's vocal tract. Every speaker's speech production apparatus is unique in the structure and shape of its vocal tract, as also in several associated factors such as muscle agility, tissue elasticity, moisture levels, length and thickness of the vocal folds, lung capacity, tracheal diameter etc. As a result, we may expect the resonance patterns produced by the apparatus to also exhibit speaker-specific characteristics that cannot be completely mimicked to their smallest detail. Articulatory-phonetic units of speech are associated with specific patterns of vocal-tract resonances. Consequently, we may expect articulatory-phonetic units to have fine-level speaker-specific characteristics that cannot be mimicked in their entirety.

For our problem it is sufficient to focus on features that definitely contribute to the intelligible content of speech and clearly do relate to the physical characteristics of the speaker. The spectral peak structure of speech, or *formants*, satisfy these criteria. Note that these are not the *only* features that satisfy these criteria; rather, it is our claim that they provide *sufficient* evidence of the nature of impersonated speech that could be exploited to verify denial of voice.

Formant-related measurements may be expected to be useful in this context. We give the specific example of formant transitions between different phonemes in continuous speech. The timing and duration of transitions is related to the degree of agility and control that the speaker's vocal tract is able to exert while changing the vocal tract configurations during continuous speech. The formant energy patterns during these transitions capture many of these speaker characteristics, some of which are relatively invariant to impersonation. Fig. 1 shows an example. Fig. 1(a) is the spectrogram of multiple instances of the word "great" spoken by a target and his impersonators. The transitions between the phonemes R and EY in this word are visible as the high-energy sloping contours in the figure. We see that the slope of these formant transitions is different across different speakers. However, the slopes are distinctive for each speaker. Note that remarkably, the slopes are similar for the same speaker regardless of the actual duration of the phoneme. In addition, note that the formant transition patterns persist in spite of the differences in fundamental frequency seen via the harmonics visible on the spectrogram. The patterns thus also appear to be invariant to pitch changes in a speaker's voice. Fig. 1(b) shows the spectrogram of the same word spoken by eight different people in their natural voices. We see again that the transition region slopes are distinctive and differ for each speaker. This is of course only one of many transitions that can be evaluated. We have found such transition patterns to be largely invariant to impersonation for many phoneme combinations.

Qualitative evidence of this kind suggests that the *distribution* of the formant features produced by the mimic in producing any articulatory-phonetic unit of sound may be expected to differ from that of the natural speech by the speaker, and the difference will vary with the unit and the impersonator. Based on this hypothesis, in Section 4 we develop a test to evaluate the denial of voice ownership. To do so effectively, however, we first need to isolate the effects of impersonation from the effects of normal factors that affect (and cause variations in) a person's speech. To some extent, this can be resolved at the feature extraction level itself, as we explain below.

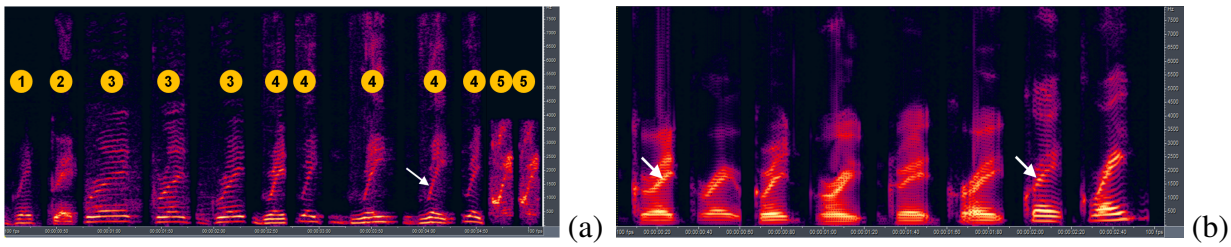


Fig. 1. ***(a)*** *Spectrogram of the word "great" spoken by multiple people. 1,2 and 3 are impersonators of the target 4, and 5 is the target's voice over a noisy, nonlinear, lossy channel. Formant transitions between phonetic units R and EY are seen as the high-energy slopes within each instance of the word, as indicated by the white arrows.* ***(b)*** *Spectrogram of the same word spoken by eight different people in their natural voices. Transitions between R-EY are marked with white arrows.*

3. Extracting stable spectral peak features from articulatory-phonetic units

Much of the natural variation in a person's speech is manifested in the contextual variations in spectral peak trajectories of the phonemes. In this section we explain how we eliminate these to retain only those parts of phonemes that directly relate to the speaker's invariant vocal tract characteristics.

The locus theory of phonemes states that every phoneme has a "locus" corresponding to a canonical arrangement of the vocal tract for that phoneme, and that the articulators move towards it in the production of the phoneme [17]. In continuous speech the loci of phonemes may not be fully reached as the articulators move continuously from one set of configurations to another. Fig. 2 shows an example. The interior regions of the phoneme that are representative of the locus are *specific to the speaker's vocal tract and articulators*. They are also much more invariant to adjacent phonetic context [18] and longer-term prosodic and expressive trends in the speech signal than the boundaries. In our approach we therefore first identify these interior stable "locus" regions of the phonemes, and extract formant features from them.

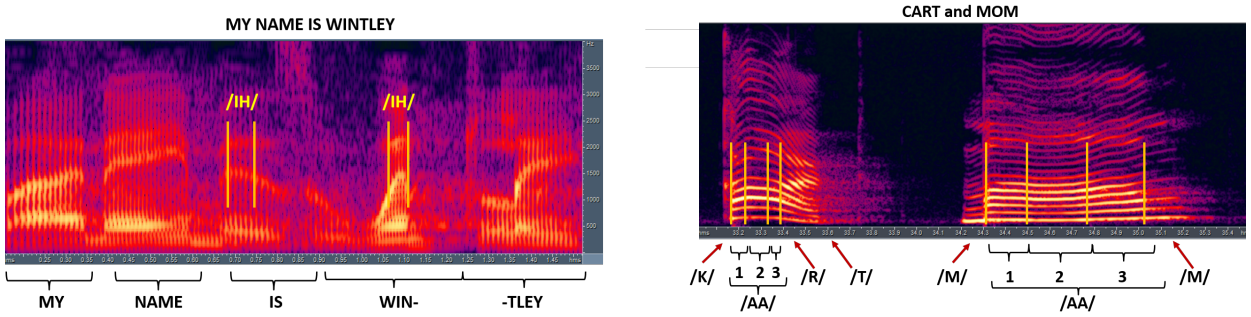


Fig. 2. ***(a)*** *The sentence "My name is Wintley" spoken by an adult male. Note the long-term spectral patterns that flow continuously across the entire utterance while moving towards different phoneme loci in the process. The phoneme IH is marked . It occurs in two different phonemic contexts, and the influences can be seen at the beginning and end of the phoneme in each instance.* ***(b)*** *State level segmentations for the phoneme AA in the American English pronunciation of the word CART (K AA R T) and MOM (M AA M).*

The exact interior region that represents the locus may, however, vary with the instance of the

phoneme. It need not be at the center of the phoneme, and its position must be carefully estimated. In order to identify this section accurately, we build on the fact that we may expect the inter-instance variation between different instances of the phonemes to be lower in these "locus regions" than in other regions of the instances. We embody this principle into an algorithm designed to identify the locus region of phonemes, which we outline below.

We model each phoneme by a Hidden Markov Model (HMM), with Gaussian-mixture state output densities. Each phoneme is modeled by a three-state HMM with a strict left-to-right topology for transitions. The HMM topology ensures that any sequence of states drawn from the HMM must include a segment from the central state. Conventional maximum-likelihood HMM training utilizes the Baum-Welch algorithm to derive the estimate $\hat{\Lambda}$ of the parameters of the HMM for a stochastic process (in this case a phoneme) to maximize the log likelihood of the training instances $\mathbf{X}$:

$$\hat{\Lambda} = \arg\max_{\Lambda} \log P(\mathbf{X}; \Lambda)$$

We modify the above to include an *entropic prior* on the distribution of the central state:

$$\hat{\Lambda} = \arg\max_{\Lambda} \log P(\mathbf{X}; \Lambda) - \alpha H(\Lambda_c) \tag{1}$$

where Λ_c represents the parameters of the Gaussian-mixture density of the *central* (i.e. the second) state, and $H(\Lambda_c)$ represents the entropy of a Gaussian-mixture density with parameters Λ_c. α is a regularization parameter.

The above optimization simultaneously attempts to both maximize the likelihood of the observations, and minimize the entropy of the distribution of the central state. This has the effect that the observations from the central state, as modeled by the HMM, have least variation and are hence maximally consistent across different observation sequences. Combined with the fact that every state sequence through the HMM must visit the central state, for phonemes modeled by such entropically trained HMMs we may expect the central state to represent the consistent "locus" region of the phoneme.

The parameter estimation of Eq. 1 can be performed using a modified version of the Baum-Welch algorithm described in [19]. Given trained HMMs for any phoneme, the locus region of a novel instance of the phoneme can be segmented out by aligning the instance against the states of the HMM. The segment of the instance that aligns against the central state represents the locus region of the phoneme. An example is shown in 2(b).

We extract two kinds of formant features from the locus regions of each phoneme: the formant peaks and formant bandwidths. Formant transition slopes can be calculated dynamically from these and are not part of the initial extraction process. It must be noted that the entropic HMM conservatively focuses on matching the steady-state regions of the phonemes; however this comes at the cost of accurate identification of the boundaries of these regions, and hence the durations of the steady state.

The formant measurements are obtained through standard procedures [20] which we repeat here for reference: we fit a 12^{th} order LPC spectrum to 50ms slices of the signal, with a 40ms overlap between adjacent frames. The segments are pre-emphasized, and windowed with a Hamming window. The 12 LPC pole frequencies represent candidate formant positions. To derive the actual formants we run a simple dynamic programming algorithm to track formant values, subject to constraints on the actual values of the formants. The first formant F1 is assumed to lie in the range 200-900 Hz, F2 is assumed to lie within 900-2800 Hz, F3 is assumed to lie between 1500-3500 Hz, and F4 and F5 are assumed to lie above 3500 Hz and 4500 Hz respectively. A subsequent

refinement is performed by scanning for the frequency location of an actual spectral peak in the vicinity of the pole frequency at which a formant has been identified by the dynamic program. The formants are tracked only within speech regions.

Note that the dynamic program may not always find all formants in every frame, although F1 and F2 are generally found for most (but not all) phonemes. We found the estimation of F4 and F5 to be most unreliable. Once formant positions are determined, the bandwidth of each formant is determined by finding the frequencies on either side of the spectral peak at the formant where the spectral energy falls by 3dB with respect to the peak frequency.

4. A hypothesis testing method for category comparison

The crux of our problem is to determine if (the set of formant features from) a phoneme uttered by a target speaker differs statistically from imitations by an expert mimic. It is to be expected that impersonators will generally perform a better job of mimicking some phonemes than others [16]. So, a secondary task is that of determining *which* phonemes they mimic more successfully (and the complement of this set).

We are given collections of formant feature vectors from the target speaker and from the impersonator. To determine if the impersonator succeeds in mimicking the speaker we must determine if both sets of vectors are drawn from the same, or very similar distributions. This is a classical hypothesis testing problem, which we tackle using conventional two-sided *t*-tests to compare distributions of scalar values (although non-parametric tests such as the Kolmogorov-Smirnoff test [21] and the Wilcoxon test [22] are also applicable). We use the Hotelling T-squared test [23] to compare multivariate measurements.

The above hypothesis tests, meant for direct comparison of the distributions of random variables, are suitable to compare renditions of phonemes by a target speaker with those by a mimic. An alternate comparison is needed to distinguish variations between different renditions of the same phoneme by a target speaker from those due to impersonation. In the abstract, this can be stated as the comparison of the statistical distance between the distributions underlying two sets of vectors A_1 and A_2 (representing two renditions by the target speaker) to that between two other sets B_1 and B_2 (which may represent renditions by the target speaker and an impersonator).

The usual approach to such verification is through the comparison of a divergence, such as the Kullback-Leibler (KL) divergence, between the distributions underlying A_1 and A_2 to that between the distributions for B_1 and B_2. However, this requires estimation of these distributions as well as the KL divergences between the estimated distributions, both of which are challenging tasks, particularly when the data do not follow any simple distributions and are too few for accurate estimation of more complex distributions, as is the case with formant-value vectors from phoneme segments.

Once again, the setting is more appropriate to a statistical hypothesis test. The natural test here is the multivariate Fisher test [24]. However, this implicitly assumes a Gaussian distribution for the variables, an assumption that is not valid for our setting. Instead, we use an alternative hypothesis-testing framework proposed in [25], which evaluates the comparison of sets using a synthetic proposal distribution over random vectors.

Mathematically, we state the problem as follows: given two sets of samples A_1 and A_2 (representing two sets of samples from the same speaker) and as a contrast two sets B_1 and B_2 (representing samples from the speaker and the impostor), our goal is to determine if the two groups of samples A_1, $A_2 \subset \mathbb{R}^N$ are closer than B_1, $B_2 \subset \mathbb{R}^N$.

To perform the comparison, we compute distances of the sets with respect to "query" vectors drawn from some distribution Q. Given any "query" vector $q \in \mathbb{R}^N$ we compute the distance to the sample set A_1 as

$$d(q, A_1) = \frac{\sum_{x \in A_1} \|q - x\|}{|A_1|}$$

We define the distance between q and A_2 in a similar manner. We can then define the dissimilarity between A_1 and A_2 given a query q as

$$D(q, A_1, A_2) = d(q, A_1) - d(q, A_2)$$

For any distribution Q, and $q \sim Q$, we should have

$$\mathbb{E}\left[d(q, A_1) - d(q, A_2)\right] = 0$$

In practice, we set Q to be the global distribution of the data – all speech in this case. Given a set of k independent queries drawn from Q, we can now compute

$$\hat{d}(q_1, ..., q_k; A_1, A_2) = \{D(q_i, A_1, A_2),\ i\ =\ 1 \cdots k\}$$

According to the central limit theorem $\hat{d}(q_1, ..., q_n; A_1, A_2)$ is normally distributed. We now have a univariate normal statistic which can be evaluated using the Fisher test. We note that the degree of similarity between A_1 and A_2 is related to the variance of $\hat{d}(q_1, ..., q_n; A_1, A_2)$, which should be small if A_1 is very similar to A_2. To determine if A_1 and A_2 are closer than B_1 and B_2, we now compare the variance of $\hat{d}(q_1, ..., q_k; A_1, A_2)$ with the variance of $\hat{d}(q_1, ..., q_k; B_1, B_2)$. If the queries used to compute the former statistics are independent of the queries used to compute the latter, we can do so using the standard Fisher test in a procedure that is analogous to ANOVA.

In the Fisher test we have two samples, $x_1, ..., x_n$ and $y_1, ..., y_m$, each one i.i.d. from two populations which each have a normal distribution. The hypothesis tested is that the variances of the two are equal. In particular, if $\overline{X}$ and $\overline{Y}$ are the corresponding sample means, we can define

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \text{ and } S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left(Y_i - \overline{Y}\right)^2$$

where S_X^2 and S_Y^2 are the sample variances. Then the test statistic

$$F = \frac{S_X^2}{S_Y^2}$$

has an F-distribution with $n - 1$ and $m - 1$ degrees of freedom if the null hypothesis of equality of variances is true. Otherwise it has a non-central F-distribution. The null hypothesis is rejected if F is either too large or too small, *i.e.* if the CDF of the F-distribution at the computed F value is either too small (e.g. 0.01) or too high (e.g. 0.99) showing the F value to be highly unlikely to have been generated by a central F distribution.

Applying this procedure to our problem, we define F as the ratio between the variance of $\hat{d}(q_1, ..., q_k; A_1, A_2)$ and the variance of $\hat{d}(q_1, ..., q_k; B_1, B_2)$. If F is smaller than 1 and the Fisher test shows statistical significance, we can say that the groups A_1 and A_2 are closer than B_1 and B_2.

5. Experimental evaluation

For our experiments we used a corpus of publicly available data of the natural voices of seven different US politicians who were either presidents of the USA or presidential candidates. We also collected data from public performances (on national TV) of different publicly-acclaimed impersonators of each of these personalities. The corpus included several hours of data from these target personalities and their impersonators. The speaker composition of the data used is given in Table 1, which also assigns IDs to each speaker that are used for brevity of presentation in the rest of this paper. The data include both male and female voices. All impersonators were of the same gender as their target. 30 minutes or more of target (natural voice) and impersonator speech (both natural voice and impersonation of target) were used for the experiment.

Target	Impersonated by (ID)
Bill Clinton	Darrell Hammond (BC-1), Steve Bridges (BC-2)
Sarah Palin	Tina Fey (SP-1)
George Bush	Steve Bridges (GB-1)
Barack Obama	Steve Bridges (BO-1)
Donald Trump	Anthony Atamanuik (DT-1), Bob DiBuono (DT-2), Darrell Hammond (DT-3), Jimmy Fallon (DT-4), Taran Killam (DT-5)
Hilary Clinton	Amy Poehler (HC-1), Tina Fey (HC-2), Kate McKinnon (HC-3), Rosemary Watson (HC-4)
Bernie Sanders	James Adomian (BS-1)

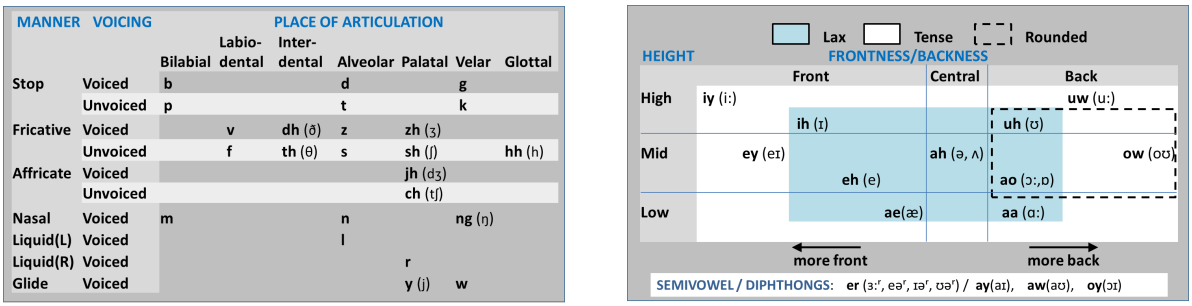
Table 1 Speaker composition of the data used for experimentation.

Presidential candidates in USA must be American citizens by birth. The target speakers in our database were all born and brought up in America. All were native speakers of American English, although some had traces of accents from places in America that they grew up in. The most pronounced accent was that of Bernie Sanders. We transcribed the data carefully to mark all speech and non-speech sounds, including breaths, partial words, and other filled pauses such as UHs and UMs. The portions of speech that could not be deciphered were marked as unintelligible and modeled by a universal background model for segmentation purposes.

For our analysis we used a phonetic pronunciation dictionary that comprised 50 phonemes, including those that represented the filled pauses, silences and other non-speech sounds. This is the specific set of phonemes used by the CMU Sphinx automatic speech recognition (ASR) system [26], which we used for segmentation and for feature extraction purposes.

Fig. 3 shows the key membership of this specific phoneme set along with their IPA symbols. In the rest of this paper we will use the non-IPA symbols as shown in Fig. 3, since we believe they are more readable than IPA symbols.

To derive the locus regions of phonemes, we segmented the speech using 3-state HMMs trained on 5000 hours of clean speech from a collection of standard speech databases in English available from the Linguistic Data Consortium. The CMU-Sphinx ASR system [28] was used for this. For this purpose we used high-temporal-resolution MFCC vectors [29], computed over 20ms analysis frames 200 times a second, to achieve a temporal resolution of 5ms. Figure 2 shows typical segmentations that are achieved by the algorithm. Note the similarity in the structure of the data in the central (locus) state across the different instances of the phoneme. Subsequent to segmentation, formant features were derived from the central-state regions of the phonemes.



(a) (b)

Fig. 3. ***(a)*** *Classification of consonants based on the involvement of different parts of the vocal tract (place of articulation) and the manner of their movements (presence of absence of voicing, and the manner or articulation).* ***(b)*** *Vowels in American English, referenced from [27] to represent the specific vowels modeled by the CMU Sphinx ASR system used in our experiments. The IPA symbols for these phonemes are also shown in these charts. We however choose to use plain alphabetical symbols to represent them, for easier reading.*

5.1. Results

We have reasoned that in order to invalidate the denial of a voice recording we must demonstrate that it is statistically infeasible for anyone else to have produced the specific set of sounds in the recording. Our experiments are founded on the hypothesis that impersonators will fail to exactly replicate spectral resonance characteristics, namely formants, formant bandwidths, and the manner in which they change. We also expect that different impersonators vary in their ability to imitate different features, and that this variation, which depends on the degree of control over articulators, will also be phoneme dependent. In general, we expect some features of some phonemes to be difficult for *anyone* to mimic (unless it is the same person!), while others may vary in their susceptibility to imitation. We run two sets of tests to verify these hypotheses.

5.1.1. Pair-wise testing: We evaluated the ability of each impersonator to mimic each of several formant features, for different phonemes. Specifically, the features tested were the first five formants F1-F5 and their respective bandwidths B1-B5. If an impersonator is able to mimic a particular feature well within a given phoneme, we expect the distribution of that feature within instances of the phoneme in impersonated speech to be close or identical to the distribution for the target speaker. This is evaluated within a hypothesis testing framework.

For each impersonator, for each phoneme, we compare the distribution of each feature with the corresponding distribution from the target speaker. Since our objective is to validate denial of voice ownership, where the claim is that the voice recording is rendered by a skilled impersonator, our default (null) hypothesis is that the impersonator *is* able to mimic the features of the target speaker very well, *i.e.* that values of samples of the feature from impersonated speech are drawn from the same distribution as feature values from the target speech. The alternate hypothesis is that they differ. For each of the features F1-F5 and B1-B5, we test this using a regular two-sample *t*-test. In addition, we also test the *joint* distributions of F1-F5, B1-B5 and (F1-F5, B1-B5). We use Hotelling's T-squared test for this purpose [23].

Both the *t*-test and Hotelling's T-squared test assume that the underlying variables have Gaussian distributions, although they are known to be robust to changes in the form of the distribution.

Formant and formant-bandwidth features do not have Gaussian distributions. To compensate, and to minimize the possibility of type-1 errors, we use a low P value threshold of 0.001 to reject the null hypothesis. Thus, when we reject the null hypothesis, the two distributions are almost certainly different, and the impersonator has definitely failed to mimic the target speaker accurately in his or her rendition of the phoneme. However, failing to reject the null hypothesis does not imply that the impersonator has been successful in his or her imitation; thus our tests are highly conservative.

Fig. 4(a) illustrates our findings. Each bar in the figure represents a specific feature. Each cell in each bar represents a phoneme. A phoneme is shown against a feature if at least one of the impersonators in our setup failed to mimic the feature within that phoneme.

We refer to any phoneme that differs significantly from the target during impersonation as an *invariant* since the impersonator is not able to vary his rendition of the phoneme accurately during impersonation. These phonemes are good candidates to test to evaluate voice denial. We note that formants F1-F4 are good at identifying invariants. F5 is not as reliable a cue, possibly because it is hard to estimate accurately in the first place.

We also note that formant bandwidths B3-B5 are not useful at all for identifying invariants. B1 and B2 are not particularly useful either. This could be because they are poorly estimated – the LPC-spectrum-based estimator that we and other researchers use may not be sufficiently accurate to discern differences in bandwidth between target speakers and impostors. In subsequent experiments we only use the joint feature comprising the formants F1-F5, since they are most sensitive to variations in the maximum number of phonemes. Fig. 4(b) shows this specific combination for five imitators of one of the targets in our database – Donald Trump. Each bar represents a phoneme, and the number of cells in each bar shows the number of impersonators who failed to mimic that phoneme accurately. As before, we see that different impersonators are able to manipulate different phonemes as expected, but some phonemes are invariant across all impersonators. Specifically, BREATH, AE, D, EY, IY, L and N are universally invariant. Note that BREATH is often treated as a phoneme in standard ASR systems, and is known to have formant structure that can be used to differentiate between speakers.

Another feature that is often a cue to identity is formant *trajectory* – the manner in which a formant moves from one position to another, as illustrated in Fig. 1. While such trajectories are highly evident across phoneme *transitions*, they are less so within a phoneme, particularly within the steady state of a phoneme. Nevertheless we compute formant trajectory slopes by fitting a linear regression to formant trajectories, for each of the formants. Fig. 4(c) shows results with the combined slopes of the first three formants, which we found to be the most informative combination of formant slopes through a similar analysis as in Fig. 4(a). We find that within-phoneme formant trajectories are not as informative as the formants themselves. However, we have not explicitly evaluated phoneme-transition trajectories – formant trajectories in the transition between phonemes – primarily due to paucity of data.

5.1.2. Multi-way testing: The criterion given above for rejecting the hypothesis of impersonation is based only on *expected* characteristics of the acoustic-phonetic units. An additional hypothesis stated earlier was that even for phonemes that an impersonator *is* able to replicate nearly perfectly, he or she will not do so in *every* instance, reverting to more natural (for the impersonator) modes when this will not affect the auditory illusion. This leads us to our second test: we now evaluate the inter-instance variation – is the variation between different instances of the phoneme by the target speaker comparable to the variation between renditions by the target and that by the

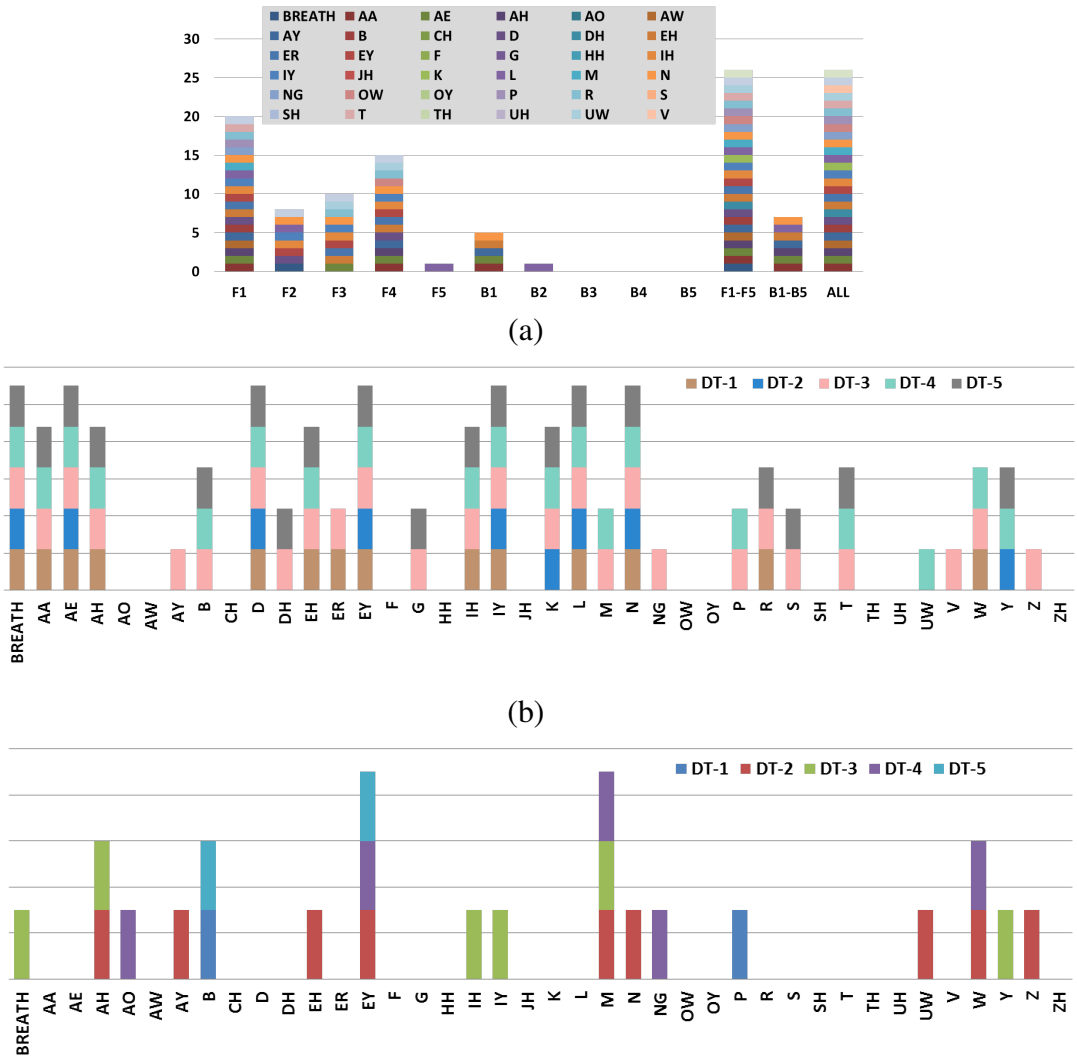


Fig. 4. *Identifying invariants in impersonations.* ***(a)*** *Evaluation of all features. Each cell represents a phoneme that at least one impersonator has failed to mimic.* ***(b)*** *Evaluation of impersonators of Donald Trump using F1-F5.* ***(c)*** *Evaluation of the same impersonators using the slopes of F1-F3.*

potential impostor? If a systematic statistically significant difference can be found, then the denial of voice ownership may be disproved. To test this hypothesis we must now draw upon the multi-way hypothesis test described in Section 4.

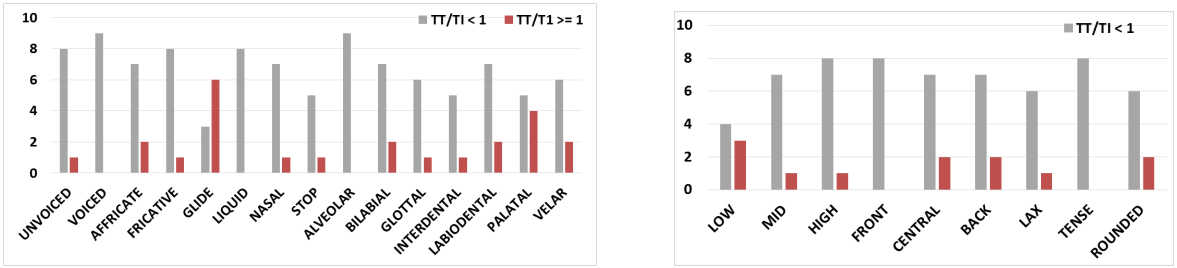
For the test(s) measurements were taken from the central regions of the phonemes which were derived as explained in Section 3. Feature vectors were computed at 0.01 second intervals. Our test comprises comparing the closeness of two random *instances* of the same phoneme from the target speaker, and the closeness of a random instance of the phoneme from the target to one from the impersonator. We used $k = 100$ to compute the F values in all cases. The test was conducted over all pairs of (relevant) targets and their impersonators. We finally considered only those pairs for which the computed F value was statistically significant to the 0.01 level.

When we consider the formants F1-F3 and their bandwidths B1-B3 as a joint 6-dimensional

measurement, tests run on individual phonemes showed that vowels and consonants fell into patterns that aligned with their articulatory-phonetic categorization. Following these observations, we grouped the vowels and consonants according to different criteria following the guidelines shown in Fig. 3. Fig. 5 shows these results at the level of individual categories. The bars in gray are those that had target-target (TT) closeness greater than target-impersonator (TI), so that the ratio TT/TI was less than 1. The cases where TT/TI is greater than (or equal to) 1 are shown as red bars.

When each consonant was tested separately the results were observed to be best for categories that followed the *place of articulation* categorization, and there was no significant distinction between the results based on the manner of articulation. This can be seen in Fig. 5(a). Similarly, when each vowel is tested separately the results were roughly observed to group into categories that followed the front-mid-back categorization of vowels. There was no significant distinction between the results based on the height of the vowels. This is seen in Fig. 5(b). There was also no significant difference in these results between males and females and their impersonators.

From Fig. 5 we see that voiced, liquid and alveolar consonants provide the greatest evidence against denial of voice ownership, as do front tense vowels. Note that these are not necessarily the most difficult phonemes to impersonate, but are the ones where the impersonator is unlikely to hold the mimicry across instances.



(a) (b)

Fig. 5. *Multi-way hypothesis test on* ***(a)*** *Consonants* ***(b)*** *Vowels. The vertical axis shows the number of speaker-imposter pairs for which the F value was found to be statistically significant.*

5.2. Testing voice denial cases

Since we do not have real cases of voice denial for this paper, we create the following experimental setup using the impersonation dataset of Table 1. We randomly selected 15 utterances of about 10 seconds in duration from each target and impersonator mentioned in Table 1, except for the case of Donald Trump which we have used extensively in the analysis above. For each target, we also selected an additional 15 utterances in the target's natural voice from a heldout set of recordings. Since these are data collected from the wild, the recording conditions and channel characteristics were not matched. We did not attempt to correct for channel mismatches where present, since in real cases such mismatches may be present and may not be correctable. For each target speaker we then assumed that all his or her impersonators' samples were those that were denied ownership by the target, and also the additional 15 natural-voice samples of the target from the heldout recordings were also denied by the target. To test our procedure, we did the following: a) Applying the guidelines provided by Fig. 5, we computed TT/TI ratios using only voiced-alveolar consonants (D,Z,N,L) and the front-high-tense vowel IY. These are categories of phonemes for which TT/TI < 1 with high confidence for most speaker-imposter pairs, as shown in Fig. 5. We note that these phonemes (other than Z) are also among the most discriminative

according to Fig. 4. We call this subset of phonemes "selected" in the table. If the ratio TT/TI was less than 0.9 (for impersonations), we labeled the voice denial case as "true", implying that the voice was indeed not of the target, and "false" otherwise (TT/TI $\geq$ 0.9) , implying that the voice belonged to the target. The results of this experiment are shown in the first column of in Table 5.2 in terms of accuracy of correct resolution of the case. b) In a contrastive experiment, we used all phonemes (consonants and vowels) to make a similar prediction keeping the threshold of 0.9 constant. The resolution accuracies are shown in the second column of Table 5.2 for this case. We note immediately that considering *all* phonemes results in significantly poorer ability to reject fake denial, than when only the phonemes identified as "good" candidates from the analyses in the previous sections. This also matches our expectation that impersonators try to create auditory illusions, and considering all phonemes includes those that support these illusions, while focusing on the discriminative ones lets us focus on those that do not.

Target	True speaker	Resolution acc. (%)		Target	True speaker	Resolution acc. (%)	
		Selected phonemes	All phonemes			Selected phonemes	All phonemes
B. Clinton	BC-1	100	67	H. Clinton	HC-1	100	53
	BC-2	100	60		HC-2	100	66
	Self	100	73		HC-3	100	73
S. Palin	SP-1	100	60		HC-4	100	73
	Self	100	66		Self	100	73
G. Bush	GB-1	100	66	B. Sanders	BS-1	93	47
	Self	100	73		Self	93	53
B. Obama	BO-1	100	87				
	Self	100	87				

Table 2 Resolution accuracies for contrived voice denial cases

6. Conclusions

This work addresses an important problem in forensic scenarios: voice *denial* – the allegation by the speaker identified in a voice recording that it is in fact by an impersonator. This reversal of perspective from the usual one of matching or authentication converts the problem from one of confident classification or detection to one of hypothesis testing, where the need is to establish a probability that the recording was that of the target. Since such a denial is generally a refutation of identity matches obtained through evaluation of large-grained statistical characteristics, we must focus on fine-detail features to reject the denial.

The strategy we use in this paper is simple in itself: demonstrate that some patterns in the voice of the target speaker are simply not replicable by an expert impersonator, and that a real impersonator *will* make mistakes. The forensic consequence of this is that in any case of denial, if the supposed impersonator in the denied recording has *not* been found to make such mistakes, he is either an improbable new talent, better even than the (best) ones evaluated, or the claim that the voice is that of an impersonator is not true. Our experiments validate this approach.

They also validate a number of other assumptions. Evidence and literature suggest that large-scale characteristics like prosody and style are well-imitated in impersonations. However, as we note from our current results and have noted in earlier work, even the best impersonators do not

succeed in mimicry of every sound. Our hypothesis that by considering the characteristics of individual sounds, we may be able to better identify the absence of impersonation is validated by the results. In essence, we expected the distribution of appropriately chosen features of two instances of a given sound to be closer when both are from the target speaker, than when one is from the target speaker and one is from the impostor. Moreover, we also hypothesized that inconsistency of delivery will result in larger variation between instances by an impostor, than would be observed between instances by a target speaker. These hypotheses have been also shown to be valid.

We conclude with a caveat: our results have been derived from study of impersonations of a small number of subjects. While we expect the broader conclusions to hold in general, the details may vary for other subjects. Our future plans include organizing a much larger corpus to repeat this study on, so that more detailed observations may be made with confidence.

7. References

[1] CNN News Channel, USA, "Donald Trump on recording: Not me," http://edition.cnn.com/2016/05/13/politics/donald-trump-recording-john-miller-barron-fake-press/, May 14, 2016.

[2] High Court of Justiciary, Edinburgh, Scotland, "Her Majesty's Advocate v Thomas Sheridan and Gail Sheridan," Decided 23 December 2010.

[3] W. Doniger, *The Woman who Pretended to be who She was: Myths of Self-imitation*. Oxford University Press, 2004.

[4] E. Zetterholm, "Voice imitation: a phonetic study of perceptual illusions and acoustic success," Ph.D. dissertation, Lund University, 2003.

[5] A. Eriksson and P. Wretling, "How flexible is the human voice? – a case study of mimicry," in *In Proc. EUROSPEECH 97*, vol. 2, 1997, pp. 1043–1046.

[6] A. Eriksson, "The disguised voice: imitating accents or speech styles and impersonating individuals," in *Language and identities*, C. Llamas and D. Watt, Eds. Edinburgh University Press, 2010, ch. 8, pp. 86–98.

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

[8] D. Deutsch, "Auditory illusions, handedness, and the spatial environment," *Journal of the Audio Engineering Society*, vol. 31, no. 9, pp. 606–620, 1983.

[9] C. McGettigan, F. Eisner, Z. K. Agnew, T. Manly, D. Wisbey, and S. K. Scott, "T'ain't what you say, it's the way that you say it – left insula and inferior frontal cortex work in interaction with superior temporal regions to control the performance of vocal impersonations," *Journal of Cognitive Neuroscience*, vol. 25, no. 11, pp. 1875–1886, 2013.

[10] C. Gallois and H. Giles, "Communication accommodation theory," *The International Encyclopedia of Language and Social Interaction*, 2015.

[11] C. McGettigan, "The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication," *Frontiers in Human Neuroscience*, vol. 9, no. 129, 2015.

[12] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a GMM-based speaker verification system?" IDIAP, Tech. Rep., 2005.

[13] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A.-M. Laukkanen, "I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry." in *INTERSPEECH*. Citeseer, 2013, pp. 930–934.

[14] F. Schlichting and K. P. Sullivan, "The imitated voice – a problem for voice line-ups?" *Forensic Linguistics*, vol. 4, pp. 148–165, 1997.

[15] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4401–4404.

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

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

[18] C. T. Ferrand, *Speech Science: An Integrated Approach to Theory and Clinical Practice (with CD-ROM)*. Allyn & Bacon, 2006.

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

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

[21] D. A. Darling, "The Kolmogorov-Smirnov, Cramer-Von Mises tests," *The Annals of Mathematical Statistics*, vol. 28, no. 4, pp. 823–838, 1957.

[22] W. H. Kruskal, "Historical notes on the Wilcoxon unpaired two-sample test," *Journal of the American Statistical Association*, vol. 52, no. 279, pp. 356–360, 1957.

[23] H. Hotelling, "A generalized T test and measure of multivariate dispersion," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 23–41, 1951.

[24] T. W. Anderson, "An introduction to multivariate statistical analysis," Wiley New York, Tech. Rep., 1962.

[25] A. Jimenez and B. Raj, "A three-way hypothesis test to compare multivariate sets," *Arxiv*, 2016.

[26] P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf, "The CMU SPHINX-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, vol. 1. Citeseer, 2003, pp. 2–5.

[27] W. Labov, S. Ash, and C. Boberg, *The atlas of North American English: phonetics, phonology and sound change*. Walter de Gruyter, 2005.

[28] "The CMU Sphinx suite of speech recognition systems," http://cmusphinx.sourceforge.net/, 2013.

[29] R. Singh, B. Raj, and J. Baker, "Short-term analysis for estimating physical parameters of speakers," in *4th IEEE International Workshop on Biometrics and Forensics (IWBF)*, Cyprus, March 2016.