Proceedings of Meetings on Acoustics

Volume 20, 2013 http://acousticalsociety.org/

166th Meeting of the Acoustical Society of America
San Francisco, California
2 - 6 December 2013

Session 2pSCa: Speech Communication

2pSCa4. Detecting voice disguise from speech variability: Analysis of three glottal and vocal tract measures

Talal B. Amin*, James S. German and Pina Marziliano

***Corresponding author's address: School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Media Technology Lab, S2.2-B4-02, Singapore, 649491, Singapore, Singapore, talal1@e.ntu.edu.sg**

The deliberate attempt by speakers to conceal their identity (voice disguise) presents a challenge for forensics and for automated speaker identification systems. Using a database of natural and disguised voices of three professional voice artists, we build on earlier findings in [Amin et al., 2014] by exploring how certain glottal and vocal tract measures, such as glottal timing (Open Quotient) and vowel formants are manipulated by the artists to create novel voice identities. We also investigate whether there are any features from these measures that can be useful for discriminating natural and disguised voices. As expected, variation in Open Quotient was speaker-dependent, and corresponded closely to social attributes (i.e., age) of the voice identities involved. By modelling the overall variability of speakers in the vowel space, we propose a new method for automatically classifying natural and disguised voices. The proposed method is found to outperform several state-of-the-art methods.

Published by the Acoustical Society of America through the American Institute of Physics

©2014 Acoustical Society of America [DOI: 10.1121/1.4879257]
Received 7 Apr 2014; published 12 May 2014

1 Introduction

The automatic detection of disguised voices is an interesting and challenging problem. Current speaker recognition systems are becoming increasingly more accurate. However, their robustness to attacks of voice disguise is still a relatively unexplored area. Several studies, including [Künzel et al., 2004, Zhang and Tan, 2008], have reported that the presence of disguised voices can significantly reduce the performance of speaker recognition systems. For example, when making a phone call, criminals will typically try to conceal their identity by disguising their voices. In such cases, having information about the nature of the voice (disguised/ natural) can prove useful for uncovering the true identity of the speaker. More generally, a disguise detection system can also be utilized for reducing the misclassification rate of speaker recognition system. By using a disguise detection as front end to a speaker recognition system, disguised voices can be filtered out before any attempt is made to relate them to the voices of authorized individuals in the system.

The simplest approach for disguise detection, as was used in [Perrot and Chollet, 2008], is the same as that used for speaker verification [Reynolds et al., 2000]. In the case of speaker verification, the goal is to verify whether a given set of speech samples belong to the claimed speaker or not. The baseline method [Reynolds et al., 2000] utilizes Mel Frequency Cepstral Coefficients (MFCCs) as features which are parameterized representations of the spectral envelope. These are extracted from short-time frames (25 ms) over the voiced segments of speech samples. During the training phase, Gaussian Mixture Models (GMM) are used to build two statistical speaker models. The first model, called the *speaker model*, corresponds to the claimed speaker, while the *background model* theoretically represents all possible speakers except the claimed speaker. When an unknown speech sample is presented to the system during testing, the likelihood ratio test is used for making a binary decision, that is, whether the given speech samples belong to the claimed speaker model or to the background model. This MFCC-GMM method works best if the spectral features across the speaker and background models are well-separated; in other words, the speakers in the two models have very different spectral characteristics [Reynolds and Rose, 1995]. For the case of disguise detection, the two models that are built are the disguised and natural voice models [Perrot and Chollet, 2008]. Although the formulation of the problem of disguise detection is similar to speaker verification, there are reasons to believe that using spectral features for disguise detection might not be the best possible approach. For speaker verification, the assumption is that the spectral features of the claimed speaker and the rest of the speakers are highly distinct. However, this assumption might not hold for disguise detection where the spectral features related to the natural and disguised voices of the "same" speaker can have significant overlap. In particular, some speech segments across the natural and disguised voices of a speaker may be very similar (poorly-disguised) while some might be highly dissimilar (well-disguised). In this study, we found that MFCC-GMM approach failed to give reasonable performance when presented with samples of natural and disguised voices from a database of three professional voice artists.

In an earlier study [Amin et al., 2014], we found that human subjects could identify disguised voices correctly only slightly better than chance. A key finding from that study was that the natural voices of speakers exhibited much less variability than their disguised voices in the vowel space. Based on this idea, an objective metric relating to the variability of speakers vowel productions in the F1-F2 space was found which consistently ranked the natural voices of speakers higher than their disguised voices. Previously, vowel space variability has also been found to be a relevant factor for speech recognition and speech intelligibility [Wade et al., 2007]. Some studies, including [Sturim et al., 2002], have also shown that considering only certain speech segments for analysis rather than all voiced segments of speech helps to bring out fine-grained phonetic differences of a speaker's productions. Thus in this study, we build upon these ideas and propose a new method which addresses the shortcomings of the traditional approaches a) by modelling the variability of speakers' productions rather than comparing spectral features directly and b) by focusing only on certain speech segments (vowels in this case) rather than using all the voiced segments of speech. In other words, we take higher level linguistic information, the vowel category, into consideration as opposed to using only a simple voiced/ unvoiced discrimination. Another key advantage of this approach is that less data is needed for modelling the disguised and natural voices. This is particularly useful in forensic scenarios where there the amount of data maybe limited. In this study, we also used Electroglottographic (EGG) signals to understand how the artists modulated their vocal fold patterns when producing natural and disguised voices. The EGG signal is independent of vowel category [Epstein, 2002] and primarily depends on the anatomical characteristics of a speakers vocal folds [Childers and Krishnamurthy, 1985]. Therefore, it is useful to investigate whether there are any vocal fold characteristics which are distinct to disguised or natural voices and thus useful for their automatic discrimination.

This paper is organized as follows: Section 2 describes the data collection method and the materials; in Section 3 an analysis of the EGG signals is presented together with the vowel space analysis. In Section 4, we first describe a baseline method for disguise detection and then present a new method, the results are discussed in Section 5 and finally Section 6 concludes the paper.

2 Data Collection

Synchronous speech and electroglottgraphic signals were collected from three (one female, two male) professional voice artists. We refer to the artists as 1F, 2M and 3M, respectively. English was the first language of all three artists with some differences in dialectal features (South Asian, Southeast Asian and North American for 1F, 2M and 3M, respectively). The voice artists impersonated various fictitious characters. The artists had the freedom to choose the characters, though they provided labels that indicated certain social variables such as approximate age and gender. Each artist in total produced nine distinct voice identities including their natural voice. The recordings took place inside a sound-attenuated room. Speech signals were recorded using an AKG C520L head-mounted condenser microphone. To obtain the EGG signals, two electrodes were placed externally over the larynx of the subject. The electrodes were connected to the EG2-PCX2 Electroglottogram by Glottal Enterprises [Rothenberg, 1992] which measures the electrical conductance of the vocal folds during voiced phonation. The analog speech and EGG signals were digitized using a Zoom H4n recorder and stored in WAV format at a sampling rate of 44.1 kHz with 16-bit resolution.

The speech material consisted of nine short sentences each one having two monosyllabic target words. The target words included one of the following vowels: /æ/, /ʌ/, /ɪ/, /i/, /u/, and /ɛ/. The target vowels were chosen because a) they are representative of the overall vowel space of English, and b) they are relatively robust to dialectal differences. The target words were placed in positions of maximum prosodic prominence. For each voice, the artists produced the nine sentences in a sequence, and then repeated the sentences in the same voice. The sentences that were used in this study can be found in the Appendix.

3 Analysis

In this section, we present an analysis of the Electroglottographic (EGG) signals and the vowel space. The EGG signals are related to the periodic movements of the vocal folds and are useful for understanding how the artists modulated their vocal fold patterns when producing different voice identities. The vowel space on the other hand, depends on the variations in the vocal tract shape and is useful for understanding the strategies and variability exhibited by the artists when producing different vowels. To prepare the data for analysis, a phonetic parse of the samples was first obtained automatically using the Penn Phonetics Lab Forced Aligner [Yuan and Liberman, 2008]. The results of the automatic segmentation were then manually checked and corrected by a trained phonetician.

3.1 Electroglottographic signals

Electroglottography is a non-invasive technique for estimating the vocal fold contact area, and it does so by measuring the electrical conductance between the vocal folds. It was first introduced by Fabre [Fabre, 1957]. Since the EGG signal is free from the effects of the vocal tract, it is useful for analyzing the complex periodic movements of the vocal folds during voiced phonation [Childers and Krishnamurthy, 1985]. Previous studies, including [Campbell et al., 2003], have indicated the usefulness of the EGG signals for speaker identification. The motivation for using EGG signals in this study was therefore to investigate whether the artists were utilizing variations in their vocal fold patterns for producing distinct voice identities.

One useful parameter which can be studied from the EGG signals is the Open Quotient (OQ). Over a vocal fold vibratory period, the OQ represents the percentage of time for which the glottis remains opened. The OQ has been reported to be related to the age and gender of speakers. The OQ was found to decrease with increasing age for females in [Higgins and Saxman, 1991, Winkler and Sendlmeier, 2005], while for males the OQ was found to

Table 1: Mean (μ) and standard deviation (σ) of open quotient

Voice Artist	**Label**	μ (%)	σ(%)
3M	OM	52	17
2M	OM	59	17
1F	OM	60	11
1F	OF	60	20
2M	OM	61	13
1F	YM	61	13
1F	YM	70	13
1F	YF	72	12
1F	YF	73	09
2M	YM	75	08
2M	YF	79	07

increase with age [Higgins and Saxman, 1991]. Since the OQ has been reported to be mostly independent of the vowel category [Epstein, 2002], all the target vowels associated with a voice were used to obtain the mean (μ) and standard deviation (σ) of the OQ. Table-1 shows the mean OQ and its standard deviation for the voices of all three artists. The method in [Amin and Marziliano, 2013] was used for estimating the OQ. The voices listed in Table-1 were assigned age and gender labels by the artists based on the characters they were projecting. These voice labels represent "O" (old), "Y" (young), "M" (male), "F" (male) voices. Thus "OM" in Table-1 specifies an "old male" voice for an artist. The voices have been sorted according to their mean OQ values.

Table-1 shows that there is a correspondence between the perceived age and the mean OQ value. Consistent with previous studies, we find that the "old" voices of the artists had lower mean OQ values than their "young" voices. This was also confirmed by a one-way ANOVA which indicated that there was a significant effect of voice on the mean OQ values. Although, no particular relationship between the the OQ and disguised/ natural voices was found, it is clear from the results that variation in OQ was one strategy adopted by the artists for projecting an age-related identity.

3.2 Vowel space analysis

In this section, we demonstrate how the variability of voices in the vowel space can be a useful parameter for separating natural and disguised voices. Acoustic variability such as that related to formant variances in vowels has been shown to be an important factor for speech recognition and intelligibility [Wade et al., 2007]. Based on some of our earlier findings in [Amin et al., 2014], we predict that the artists will be more practised (less variable) in their natural voice vowel productions than their disguised voices.

The vowel space is a two dimensional plot of the first two formant frequencies (F1 and F2). F1 and F2 are the primary acoustic correlates of perceptual differences among vowel categories. A vowel space is therefore useful for visualizing the inter- and intra-speaker variability exhibited by speakers in their vowel productions. If we consider the disguised and natural voices of the artists as two separate groups in the vowel space, then a greater inter-class separation and smaller intra-class variability is desirable for accurate discrimination between the disguised and natural voices.

The vowel space analysis was performed by first obtaining the first two formant frequencies (F1 and F2) for the target vowels. These were estimated using the Burg method in Praat [Boersma and Weenink, Retrieved October 21, 2011]. With the number of poles set to 12, the F1 and F2 values were obtained from 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. The F1 and F2 values obtained were then manually checked, with less than 7% values identified as having potentially erroneous formant estimates based on what is typical for each vowel. These values were then corrected by a visual inspection of the Short Time Fourier Transform (STFT) time-frequency distribution (spectrogram).

The F2 vs. F1 plot of vowel /u/ for the three natural voices and a single disguised voice of each artist is shown in Figure 1. The ellipses in Figure 1 represent the 95% confidence region with natural and disguised voices represented by red and blue colored ellipses respectively. From Figure 1, we find that the natural and disguised voices of the artists occupy similar regions in the F1-F2 space. The artist 1F in particular shows significant overlap between her natural and disguised voice. Considerable overlap between the

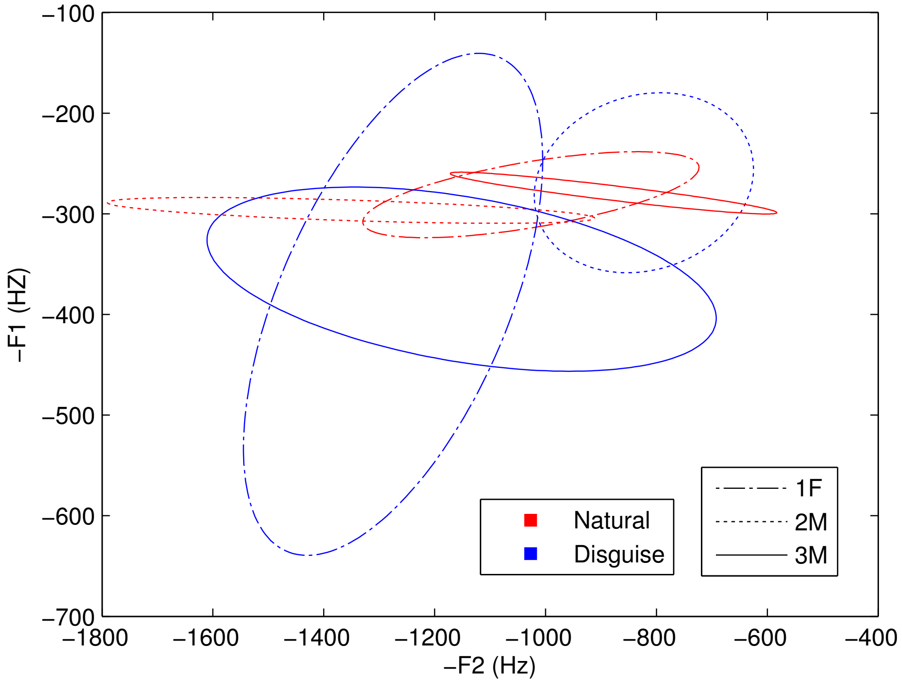

Figure 1: The natural and disguised voices for vowel /u/. The ellipses represent the 95% confidence region

disguised and natural voices was also observed for the rest of the target vowels as well. From these results, it is clear that any approach which relies on the formant values or in general the spectral features will have a very high chance of confusing disguised and natural voices. Another key observation from Figure 1 is that the ellipses associated with natural voices occupy much smaller area in the vowel space compared to the disguised voice ellipses. This suggests that the artists were more consistent in their natural voices when producing the target vowels. A multivariate analysis of variance (MANOVA) also confirmed that the variances in F1 and F2 associated with natural voices were among the lowest when averaged across all vowels. Moreover, there was no evidence found suggesting that the artists were simply changing their vocal tract length for producing different voices. The artists made adjustments to the formats on a vowel-by-vowel basis, so that the changes cannot be explained by a simple contraction or expansion of the vowel space. More details on this can be found in [Amin et al., 2014]. In Section 4.2, we build upon these findings and show how the vowel space variability can be modeled as a feature vector for discriminating disguised and natural voices.

4 Methods

In this section, we first describe the state-of-the-art baseline method which has already been used in [Perrot and Chollet, 2008] for the detection of disguised and natural voices. Then we propose a new method which relies on the vowel space variability for discriminating natural and disguised voices.

4.1 MFCC-GMM method

The most popular method used for speaker identification [Reynolds and Rose, 1995] and speaker verification [Reynolds et al., 2000] involves use of Mel Frequency Cepstral Coefficients (MFCCs) [Davis and Mermelstein, 1980] as features together with a Gaussian Mixture Model (GMM) as a supervised classifier. The MFCC-GMM approach was also used in [Perrot and Chollet, 2008] for the purpose of automatic disguise detection.

For an L dimensional feature vector $\boldsymbol{x}$ the Gaussian mixture density is a weighted sum of M unimodal Gaussian densities and is defined as

$$p\left(\boldsymbol{x}|\lambda\right) = \sum_{k=1}^{M} w_k p_k\left(\boldsymbol{x}\right). \tag{1}$$

Here w_k are the weights and p_k represents the individual Gaussian densities. The sum of all the weights must be equal to one, i.e. $\sum_{k=1}^{M} w_k = 1$. The Gaussian density function is defined as

$$p_k\left(\boldsymbol{x}\right) = \frac{1}{\left(2\pi\right)^{L/2}\left|\boldsymbol{\Sigma_k}\right|^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})'\left(\boldsymbol{\Sigma_k}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu_k}\right)\}, \tag{2}$$

where $\boldsymbol{\mu_k}$ is a $L \times 1$ vector of means and $\boldsymbol{\Sigma_k}$ is a $L \times L$ covariance matrix. Thus, a Gaussian mixture model can be completely described by a set λ as

$$\lambda = \left\{p_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}\right\}, \qquad \text{where } k = 1, \ldots, M. \tag{3}$$

Given a set of feature vectors, the expectation-maximization algorithm [Dempster et al., 1977] can then be used to obtain the maximum likelihood estimates of the model parameters given by λ.

For the purpose of disguise detection, two Gaussian mixture models are constructed: one for the natural voices denoted as λ_N and one for the disguised voices denoted as λ_D. To perform the binary classification of a sequence of feature vectors $X = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_T}\}$, the following likelihood ratio test is used

$$\Lambda\left(X\right) = \log p\left(X|\lambda_N\right) - \log p\left(X|\lambda_D\right). \tag{4}$$

Thus if $\Lambda\left(X\right) \geq 0$ the feature vectors X belong to a natural voice and if $\Lambda\left(X\right) < 0$ the feature vectors X belong to a disguised voice. For a sequence of feature vectors X, the log likelihood function in Equation (4) can be calculated as the summation of the logarithm of the Gaussian density function in Equation (1). It is thus defined as

$$\log p\left(X|\lambda\right) = \sum_{t=1}^{T} \log p\left(\boldsymbol{x}|\lambda\right). \tag{5}$$

The GMM classifier is typically used together with MFCC feature vectors which are computed over short-time windows over the voiced segments of the speech signal. The MFCCs provide an efficient representation of the spectral envelope and are considered free from the influences of the vocal folds. The dimensionality L of the MFCCs is usually 36 including the delta and double-delta coefficients [Reynolds and Rose, 1995].

4.2 Ellipse-QDA method

Now we show how the vowel variances in the F1-F2 space can be modeled for the purpose of disguise detection. Let $\mathbf{G}^k$ be the matrix which contains the first two formants for a vowel k. It is defined as

$$\mathbf{G}^k = \left[\mathbf{f}_1^k \ \mathbf{f}_2^k\right] \tag{6}$$

where $\mathbf{f}_1^k$ and $\mathbf{f}_2^k$ are $n \times 1$ vectors representing the F1 and F2 values of vowel

k in Hertz respectively. They are defined as

$$\mathbf{f}_1^k = \left[f_{11}^k f_{12}^k \dots f_{1n}^k\right]^\intercal \tag{7}$$

and

$$\mathbf{f}_2^k = \left[f_{21}^k f_{22}^k \dots f_{2n}^k\right]^\intercal \tag{8}$$

where n are the number of tokens (samples) of vowel k.

As a first step, we make the vectors $\mathbf{f_1}^k$ and $\mathbf{f_2}^k$ zero mean

$$\mathbf{f}_1^k = \mathbf{f}_1^k - \overline{\mathbf{f}}_1^k \tag{9}$$

$$\mathbf{f}_2^k = \mathbf{f}_2^k - \overline{\mathbf{f}}_2^k. \tag{10}$$

The matrix $\mathbf{G}^k$ in Equation (6) represents a set of points (tokens) of a vowel k in the F1-F2 space. The arrangement, position and variability of the vowel k can be captured by drawing an ellipse around these points. The ellipse can be constructed such that it covers, for example, 95% of the data points as shown in Figure 2. Now we show how the relevant parameters of such an ellipse can be extracted given the matrix $\mathbf{G}^k$. Let $\mathbf{C}^k$ be the correlation matrix for vowel k defined as

$$\mathbf{C}^k = \begin{bmatrix} \langle \mathbf{f}_1^k, \mathbf{f}_1^k \rangle & \langle \mathbf{f}_1^k, \mathbf{f}_2^k \rangle \\ \\ \langle \mathbf{f}_2^k, \mathbf{f}_1^k \rangle & \langle \mathbf{f}_2^k, \mathbf{f}_2^k \rangle \end{bmatrix} \tag{11}$$

Let $\mathbf{v}_1^k$ and $\mathbf{v}_2^k$ represent the eigen vectors and α_1^k and α_2^k represent the eigen values of the correlation matrix in Equation (11), where $\alpha_1^k \geq \alpha_2^k$. The eigen vectors here represent the axes of the ellipse. The coordinates of the points projected on these axes follow a Gaussian distribution with zero mean and variance which is given as

$$\sigma_1^k = \sqrt{\frac{\alpha_1^k}{n-1}} \tag{12}$$

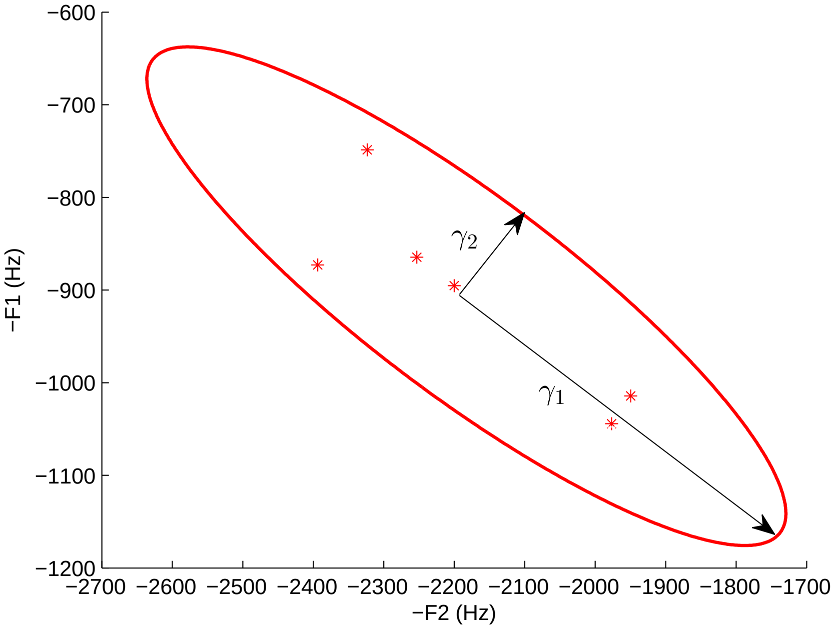

Figure 2: The 95% confidence ellipse for vowel /ae/. The major and minor axis of the ellipse are represented by γ_1 and γ_2.

$$\sigma_2^k = \sqrt{\frac{\alpha_2^k}{n-1}} \tag{13}$$

Let U^k and V^k represent the coordinates of the ellipse which lie along the direction of $\mathbf{v}_1^k$ and $\mathbf{v}_2^k$ respectively. Then the ellipse is of the form

$$Z^k = \left(\frac{U^k}{\sigma_1^k}\right)^2 + \left(\frac{V^k}{\sigma_2^k}\right)^2. \tag{14}$$

The ellipse in Equation (14) follows a chi-squared distribution. For a 95% confidence ellipse, using a lookup table we find that $P(Z^k < 5.991) = 0.95$. Using this in Equation (14), we get

$$\left(\frac{U^k}{\sigma_1^k}\right)^2 + \left(\frac{V^k}{\sigma_2^k}\right)^2 = 5.991. \tag{15}$$

From Equation (15), we find that the major and minor axis of the ellipse which covers 95% of the data points are given by

$$\gamma_1^k = \sqrt{5.991}\sigma_1^k \tag{16}$$

$$\gamma_2^k = \sqrt{5.991}\sigma_2^k. \tag{17}$$

Using these major and minor axis of the 95% ellipse, the feature vector $\boldsymbol{\delta}^k$ for the k^{th} vowel is defined as

$$\boldsymbol{\delta}^k = \left[\gamma_1^k \ \gamma_2^k\right]. \tag{18}$$

Given a set of N vowels, the feature vector for a voice is then defined as

$$\boldsymbol{\beta} = \sum_{k=1}^{N} \boldsymbol{\delta}_k. \tag{19}$$

Here $\boldsymbol{\beta}$ represents the overall variation exhibited by a voice in the vowel space. Ellipses having smaller area in the vowel space will have smaller values of δ_k in both dimensions. Smaller values of $\boldsymbol{\beta}$ on the other hand imply an "overall" smaller area occupied by a voice in the vowel space. Thus natural voices are predicted to have smaller values $\boldsymbol{\beta}$ (higher consistency) compared to disguised voices. Given the lower dimensionality of $\boldsymbol{\beta}$ compared to the MFCC features (2 vs. 36), the Quadratic Discriminant Analysis (QDA) was found suitable for classification purposes. The QDA classifier uses a quadratic boundary to separate different classes [Hastie et al., 2009]. It also has the advantage of having a lower computational efficiency than the GMM classifier.

5 Results

In this section, we compare the performance of the two methods which were presented in Section 4, that is, the MFCC-GMM method and our newly proposed ellipse based disguise detection method. For the MFCC-GMM method, the MFCC feature vectors were computed over all the voiced segments of the speech signals. The number of Gaussians, M, was set to 32. Increasing the number of Gaussians did not result in an increase in either the disguise or

Table 2: The confusions matrices for the the different matrices indicating their performance for disguised and natural voices. A high performing system has higher values along the diagonal entries and lower values of the off diagonal entries.

	Natural	Disguise
Natural	**100 %**	0.00 %
Disguise	4.17 %	**95.83 %**

(a) Ellipse-QDA

	Natural	Disguise
Natural	58.28 %	41.72 %
Disguise	55.17 %	44.83 %

(b) MFCC-GMM

	Natural	Disguise
Natural	7.59 %	92.41 %
Disguise	5.43 %	94.56 %

(c) MFCC-QDA

	Natural	Disguise
Natural	15.40 %	84.59 %
Disguise	14.76 %	85.23 %

(d) MFCC-SVM

natural voice detection accuracy. For the proposed Ellipse-QDA method, the number of vowels, N, was set to 6. In order to have a more fair and thorough comparison, we also compared the proposed method with two other variants of the MFCC based method, which use a Quadrature Discriminant Analysis classifier (MFCC-QDA) and a Support Vector Machine classifier (MFCC-SVM). This is to ensure that the gain in performance observed for the proposed method is not due to a difference in the type of method used for classification.

Table-2 shows the performance of all the different methods in terms of confusion matrices. Larger values of the diagonal entries and smaller values of the off diagonal elements are desirable for a high performing system. From Table-2, we find that the Ellipse-QDA method correctly classifies a high percentage (100 and 95.83) of natural and disguised voices while only confusing a small number of disguised voices (4.17 %) as being natural. The MFCC-GMM system on the other hand is found to be at best guessing about the voice type while both the MFCC-QDA and MFCC-SVM systems seem to be overfitting the disguised voices. Thus their accuracy for correctly detecting natural voices is very low (7.59 % and 15.40 % respectively). Together, the results indicate a) the superiority of using the vowel variability features over the spectral features and b) the usefulness of using only key segments of speech (target vowels) rather than utilizing all voiced portions.

6 Conclusions and future work

We proposed a new approach based on utilizing the variability of the vowel distributions of speakers in the formant space. Compared to traditional approaches which model the spectral envelope for the voiced segments of speech, we instead model the overall variability of a voice for specific vowels. The results indicate that for the task of discriminating natural and disguised voices, the proposed vowel consistency features clearly outperform the traditional spectral features. In conclusion, vowel variability is found to be highly important feature for the discrimination of disguised and natural voices. On the other hand, the EGG analysis did not indicate any particular differences between disguised and natural voices. However, it did reveal the reliance of the artists on their vocal fold vibratory patterns for impersonating different age groups.

Further improvements in the performance of the segmentation and formant tracking algorithms will eliminate the reliance on hand-correction, thereby allowing the proposed method to be readily implemented in a fully-automated way. In future work, it will be important to test the method using a larger number of speakers. A method that applies weighting to different vowels depending on their contribution to disguise discrimination may also lead to further gains in reliability. Finally, in this study, we only considered a subset of the vowel inventory. A method that uses variability in other segments, including both vowels and consonants, may provide better resolution and therefore better overall performance.

7 Acknowledgements

The authors would like to thank Charmaine Hon and Zhixin Tan for the manual refinement of the phonetic segmentation and the voice artists: Noella Menon, Marc X Grigoroff and Rishi Budhrani.

References

T.B. Amin and P. Marziliano. Glottal activity detection from differentiated electroglottographic signals using finite rate of innovation methods. In *IEEE International Conference on Information, Communications and Signal Processing (ICICS)*, 2013.

T.B. Amin, P. Marziliano, and J.S. German. Glottal and vocal tract characteristics of voice impersonators. *IEEE Transactions on Multimedia*, 6(3): 668–678, 2014.

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

W.M. Campbell, T.F. Quatieri, J.P. Campbell, and C.J. Weinstein. Multimodal speaker authentication using nonacoustic sensors. In *Workshop Multimodal User Authentication*, pages 215–222, 2003.

D.G. Childers and A.K. Krishnamurthy. A critical review of electroglottography. *Critical reviews in biomedical engineering*, 12(2):131, 1985.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing,*, 28(4):357–366, 1980.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

M.A. Epstein. *Voice Quality and Prosody in English.* Phd thesis, University of California, Los Angeles, 2002.

P. Fabre. Un procédé électrique percuntané d'inscription de l'accolement glottique au cours de la phonation: Glottographie de haute fréquence; premiers résultats (a non-invasive electric method for measuring glottal closure during phonation: High frequency glottography; first results). *Bull. Académie Nationale de Médecine.*, 141:66–69, 1957.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning : data mining, inference, and prediction.* Springer series in statistics. Springer, New York, second edition, 2009.

M.B. Higgins and J.H. Saxman. A comparison of selected phonatory behaviors of healthy aged and young adults. *Journal of Speech, Language and Hearing Research*, 34(5):1000–1010, 1991.

H.J. Künzel, J. Gonzalez-Rodriguez, and J. Ortega-García. Effect of voice disguise on the performance of a forensic automatic speaker recognition system. In *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

P. Perrot and G. Chollet. The question of disguised voice. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, 1995.

D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1):19–41, 2000.

M. Rothenberg. A multichannel electroglottograph. *Journal of Voice*, 6(1): 36–43, 1992.

D.E. Sturim, D..A. Reynolds, R.B. Dunn, and T.F. Quatieri. Speaker verification using text-constrained gaussian mixture models. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–677. IEEE, 2002.

T. Wade, A. Jongman, and J. Sereno. Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, 64 (2-3):122–144, 2007.

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*, 40:213–225, 2005.

J. Yuan and M. Liberman. Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics*, pages 5687–5690, 2008.

C. Zhang and T. Tan. Voice disguise and automatic speaker recognition. *Forensic science international*, 175(2):118–122, 2008.

APPENDIX

List of sentences

1. The little boys **dad** had a large collection of **beads**.
2. Mrs. Schumann asked the gardener if he recognized the **buds**, and he swore that he **did**.
3. Whenever Josh is feeling **bad**, he tries to do at least one good **deed**.
4. Though the agent was a bit of a **dud**, John was enthusiastic about the **bid**.
5. There once was a lady named **Sid** who was often mistaken for a **dude**.
6. In spite of the newly planted **seeds**, the garden appeared **dead**.
7. Sasha saw the treasure lying on the **bed**, and it made her feel **sad**.
8. When the jester slipped on the **suds**, the restless crowd **booed**.
9. If Georgia had lost the **deed**, the bank might have **sued**.