Speaker identification using phonetic segmentation and normalized relative delays of source harmonics*

Diana Mendes[1], Aníbal Ferreira[1]

[1] *Universidade do Porto - Faculdade de Engenharia, Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal*

Correspondence should be addressed to Aníbal Ferreira (ajf@fe.up.pt)

ABSTRACT

Current state-of-the-art speaker identification systems achieve high performances in reasonably well controlled conditions. However, some scenarios still elicit significant challenges, particularly in audio forensics when voice records are typically just a few seconds long and are severely affected by distortion, interferences, and abnormal speaking attitudes. In this paper we are inspired by the concept of minutiae in the context of fingerprinting, and try to extract localized, phase-related singularities from the speech signal denoting glottal source idiosyncratic information. First, we perform MFCC+GMM experiments in order to find the most effective phonetic segmentation of the speech signal for speaker modelling and discrimination. Secondly, we rely on effective phonetic segmentation and, in addition to MFCC features, we extract Normalized Relative Delays (NRDs) obtained from the phase of spectral harmonics. We use a Nearest Neighbour generalized classifier for speaker modelling and identification. Our results indicate that combining a careful phonetic segmentation and the inclusion of phase-related information, performance in speaker identification may increase significantly.

1. INTRODUCTION

During the last few decades the practical importance of biometric systems [1] has increased significantly, and today we can find mature technology namely in the area of image analysis involving fingerprinting [2] and iris pattern recognition. However, biometric systems based on voice analysis are not widely deployed [3]. Although a significant amount of research work has been carried out improving the performance of current voice identification or verification systems [4], the reality shows that these systems are highly dependent on signal acquisition conditions, namely the microphone, the acoustics of the environment, signal alterations due to the communication channel, and spurious interferences due to the overlap of multiple acoustic events, including multiple voice signals. For these reasons, practical applications of voice-based biometry can be found mainly in contexts were those factors are reasonably well controlled, such as home-banking.

Typically, in the area of audio forensics, voice data is highly corrupted with noise and interferences, is produced under altered emotional or behavioural conditions and, especially, voice records are of very short duration such as just two or three seconds long. In these cases, statistical-based voice modelling such as Gaussian Mixture Models (GMMs) can not simply be applied because the amount of voice data is clearly insufficient to produce representative GMM models. As an alternative, we are inspired by the concept of minutiae in the context of fingerprinting recognition [2] and look for opportunities to identify single occurrences or singularities in the signal that might denote the unique sound signature of a specific voice, either audible or not by a human. The importance and even competitiveness of phase-related features in speaker identification has been shown recently by at least two independent research studies [5, 6]. These studies highlight that carefully chosen phase-related features exhibit a discrimination capability which is compa-

*This work was supported by the Portuguese Foundation for Science and Technology, an agency of the Portuguese Ministry for Education and Science, under research project PTDC/SAU-BEB/14995/2008. URL: http://gnomo.fe.up.pt/∼voicestudies/artts/, last accessed on March 31st 2012.

rable to that of conventional spectral magnitude-related information such as Mel-frequency Cepstral Coefficients (MFCCs). Taking as a reference the source-filter model of voice production as originally proposed by Fant [7], those results also suggest that by segregating features related to the source (i.e., glottal information) and filter (i.e., vocal tract filter) of the phonation system, we may develop perceptually-motivated signal analysis and recognition approaches as it is known that Humans use several levels of perceptual cues for speaker identification [8].

The remaining of this paper is structured as follows. In section 2 we discuss the main concepts underlying the voice production system, we review the associated source-filter model, and we discuss the underlying opportunities for robust voice analysis and feature extraction, in the perspective of speaker modelling and identification.

In section 3 we use traditional MFCC+GMM speaker modelling and classification in order to assess which specific regions in the speech signal are more rich containing individual idiosyncratic information. The purpose of this preliminary stage is to help devise a suitable phonetic segmentation paving the way for the development of specific strategies exploiting different levels of the idiosyncratic information pertaining to a speaker, namely source information (i.e., related to the glottal pulses produced by the vocal folds) and filter information (i.e. related to the vocal tract and nasal tract filters).

In section 4 the concept of Normalized Relative Delay is reviewed given its importance in representing source-related phase information, which is used in section 5 in speaker identification experiments.

In section 5 speaker identification experiments are reported which use two databases of pre-segmented and very short vowel regions only. The Weka data mining environment (that is publicly available) is used. Given that GMM modelling is not viable since the available amount of data is insufficient to build, populate and test Gaussian-based models, we use a rather simple but effective classifier which consists in the Nearest Neighbour generalized classifier. Results reveal that NRDs help to the improve the automatic speaker identification, relative to MFCCs-based signal classification only, which confirms that phase-related features possess important information and properties helping speaker discrimination and identification tasks.

Finally, section 6 summarizes the mains results and conclusions of this paper.

2. THE SOURCE-FILTER MODEL AND IDIOSYNCRATIC ASPECTS OF THE VOICE SIGNATURE

The source-filter model of voice production [7] is schematically represented in Fig. 1. It consists of three main processing stages: source excitation, vocal tract filtering, and lips and nostrils radiation [9]. The source ex-

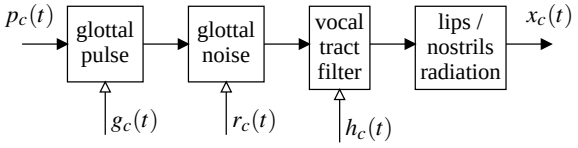

Fig. 1: Simplified source-filter model of voice production (adapted from [9]).

citation results from air expelled by the lungs and which flows through the glottis, an opening (i.e., gap) between the vocal folds located at the larynx. If the air flow is interrupted periodically as a result of a quasi periodic cycle (driven by the pulse train $p_c(t) = \sum_{\ell=-\infty}^{\infty} \delta(t - \ell T_0)$) of opening and closing of the vocal folds, the source excitation has the form of a train of glottal pulses (whose prototype is represented by $g_c(t)$) which are filtered by the vocal tract filter (represented by $h_c(t)$), giving rise to a voiced sound. If a constriction exists either at the glottis or downstream in the vocal tract (e.g., teeth or lips), the air flow becomes turbulent giving rise to an unvoiced sound. In general, speech sounds involve a combination of both types of sound.

In Fig. 1, $r_c(t)$ represents the noise produced at the glottis, and the lips/nostrils radiation is modelled as a high-pass filter (approximated by taking the derivative of the air flow volume velocity) and reflects the acoustic coupling between the vocal/nasal tract cavities and the outside surrounding space. The vocal tract filter is usually modelled as an all-pole filter that shapes the spectrum of the source according to the resonant frequencies (also known as formants) of the vocal tract.

All signal components and processes in Fig. 1 contribute to the sound signature of a specific speaker. For example, the period of oscillation of the vocal folds (T_0 or its reciprocal $F_0 = 1/T_0$, also known as the fundamental

frequency, or pitch) is traditionally an important physical feature of a given speaker sound signature, as well as its micro-variations in terms of *jitter* (or local perturbations of the fundamental frequency), and in terms of *shimmer* (or local perturbations in the magnitude of the glottal pulses).

However, a larger perspective must necessarily involve a discussion of the relative relevance of different 'layers' of the voice, namely the segmental versus suprasegmental characteristics of the voice, the linguistic and meta-linguistic aspects of speech production which are learned or acquired due to cultural influence or exposure. A reasonably simplified discussion of these aspects may also be expressed as identifying a speaker using two types of voice features which may be broadly characterized into high-level and low-level features.

High-level features, which are not in focus in this paper, concern specific speaking styles, peculiar articulatory gestures due to specific cultural influences and thus are characterized at the linguistic level. On the other hand, low-level features concern the physical properties of the voice signal due to the idiosyncrasies of the speaker, such as fundamental frequency and formant frequencies, and thus are characterized at the segmental level [4].

Since low-level features are more persistent and technically easier to analyse using signal processing algorithms, they are usually preferred for feature extraction, speaker modelling, speaker verification and speaker identification. Low-level features are typically extracted using short-term analysis (in the order of 20 ms) of the voice signal, which also facilitates data reduction for modelling and training of speaker models. Thus, in this paper, we focus on low-level acoustic cues that are more suitable for automatic voice analysis and processing.

An interesting discussion concerns the desired properties of the voice features [10]: they should be reasonably frequent in habitual speech, they should not depend on the subject health or emotional state, they should be robust to natural voice alterations due to ageing, and to channel influences and interferences. In this sense, in this paper we focus on vocal tract filtering and associated resonances, and we focus especially on the very fundamental features of the voice source: those resulting from the influence of the vocal folds located at the larynx, namely phase information.

As a curiosity, during several visits we made in 2010 to different universities and research centres including CMU (Pittsburgh, USA), Georgia Tech (Atlanta, USA), EPFL (Lausanne, Switzerland) and NJIT (New Jersey, USA), we asked several researchers active in the field as to what they feel is likely to contribute more to the uniqueness of a voice signature: either source or filter idiosyncratic features of the speaker. In general, researchers agree that both factors play a role but they are not confident stating that one predominates over the other. Possibly, this is an issue that is also subject/speaker dependent.

3. IMPACT OF PHONETIC SEGMENTATION IN MFCC+GMM SPEAKER IDENTIFICATION

Typically, in the context of speaker verification or identification, speech features result from a spectral analysis of the voice signal using uniform segmentation, for example, using a 20 ms window and 50% overlap. Most often, data is collected irrespective of the fact its corresponds to voiced or unvoiced regions of the speech signal. However, voiced and unvoiced speech have very dissimilar features. Simply stated, while voiced speech corresponds to a reasonably periodic signal including magnitude and phase glottal information, unvoiced speech has a stochastic nature.

Since voiced and unvoiced speech possess quite distinct characteristics, and thus are amenable to the extraction of distinct and specific voice features, as will be exploited in section 5, we tested the hypothesis that the segmentation of the voiced parts of speech can contribute to higher identification percentages in a speaker identification task. This possibility was studied using the most popular voice features and modelling technique: MFCCs and GMMs. The feature vectors extracted were 13-dimensional excluding the 0^{th} order coefficient. The frame size used was 256 samples, with 50% overlap.

On the other hand, it is known that the amount of data available during the training of a statistical speaker model (or speaker enrolment) such as a GMM, is of paramount importance to the performance of speaker classification using that statistical model. The amount of data available for testing obviously also influences the performance; it is for example fairly common to consider for training as much as three times more data than for testing [11]. It has also been shown that this ratio may lead to a significant reduction of the Equal-Error-Rate [12] (the EER is a performance criterion corresponding

to the point of the performance curve where the probability of false reject equals that of false accept). Other factors affecting the identification performance include channel and microphone variations, echoes and reverberation due to poor acoustic conditions, the emotional state of the speaker, as well as voice alterations due to ageing and health conditions [11].

Thus, we also studied the impact on the speaker identification performance resulting from varying the amount of test data since our purpose is to recognize a speaker using the shortest possible voice record. Training and test data are mutually exclusive.

The training and test segments for this set of experiments were obtained from the widely used database TIMIT. 40 speakers from this database were selected to participate in the tests. Three types of phonetic segmentation were performed: segmentation of the voiced part of the speech, of the unvoiced part of the speech, and also segmentation of the vowels. The segmentation was based on the TIMIT supporting documentation, which provides a phonetic transcription of each speech segment. Preliminary tests have been conducted to set the number of Gaussians in GMM speaker modelling. The number of Gaussians was adjusted from 8 till 16. In all tests, classification is performed using the log-likelihood obtained for the test data considering the GMM model. It was concluded that the best results for the test conditions assumed in this section, as well as in section 5, were obtained when the number of Gaussians was set to 12.

The speaker identification results using the different types of speech segmentation are presented in Figs. 2 and 3. Due to the short duration of the segments and in order to be more informative, the duration values are expressed in number of MFCC vectors. In these experiments, the size of the test segment was varied from 50 to 300 MFCC vectors, while the size of the training data was fixed at 301 MFCC vectors in one experiment (Fig. 2), and 578 vectors in another experiment (Fig. 3). The latter case of training duration corresponds to the maximum amount of training data available from the segmented speech.

Figure 2 shows the percentage of correct identifications obtained for all three types of segmentation, as well as for the complete voice signal with just the silence regions removed. Given that the unvoiced part of the speech represents just about 30% of all speech segments, only 200 (test) vectors were available for all considered speakers. For this reason, the tests using 250 and 300 MFCC vectors as the test duration could not be performed for the unvoiced data segments.

From Fig. 2 it can be concluded that the GMM-based modelling and classification achieves considerably higher performance using just the vowels. The complete voice signal provides the lowest percentage of correct identifications, while the voiced and unvoiced segments have comparable performance in the present test scenario.

This comparison was repeated for a larger training data set, although not for the pre-segmented unvoiced part of speech (exclusion was dictated by data insufficiency as indicated above), and the results are shown in Fig. 3. Since the vowels constitute about 45% of all available speech segments, and the unvoiced parts are excluded, the maximum training duration available is now 578 MFCC vectors. This test confirmed the previously obtained results: the vowels obtain consistently the highest performance, as the percentage of correct identifications was up to 10% higher than the percentage obtained with the complete voiced parts, and between 20% and 30% higher than the percentage obtained with the complete voice (with just silence regions removed). This is a somewhat surprising result suggesting that because the complete speech signal (after silence removal) is more undifferentiated that the other signals consisting of phonetically segmented speech regions, the building of corresponding Gaussian models is more disperse making speaker automatic identification more difficult. In brief, this just suggests that phonetic segmentation of speech is not only pertinent, but also useful.

From these experiments (involving data from 40 speakers), it can thus be concluded that the vowels, as they are constituted by the more stationary parts of speech including magnitude and phase glottal information, are more likely to capture richer speaker specific information than the other considered regions of the voice.

4. CAPTURING SOURCE-INFORMATION USING NORMALIZED RELATIVE DELAYS

Given that in section 5 we use a voice feature (NRD) which denotes glottal source phase-related information, we present here a brief perspective on NRDs. The meaning of the NRD concept has been detailed in [13] and may be explained with the help of Fig. 4. In this fig-

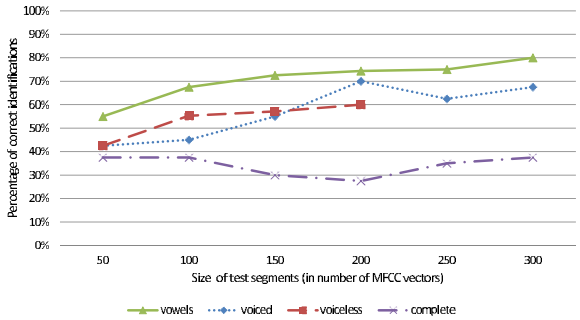

Fig. 2: System performance using different types of phonetic segmentation and considering 301 MFCC vectors as training data.

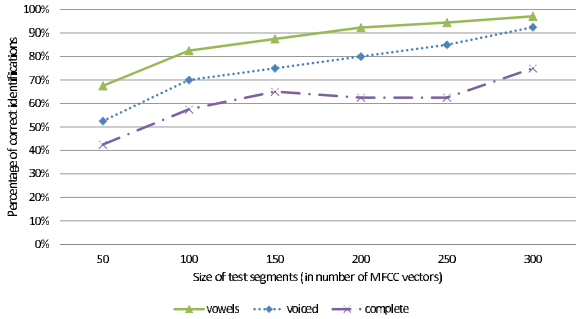

Fig. 3: System performance using different types of phonetic segmentation and considering 578 MFCC vectors as training data.

ure, n_0 represents the delay of the fundamental frequency sinusoid relative to a time reference that in our context corresponds to the center of the time analysis window defining the time support of the time-frequency transformation. The delay of a harmonic sinusoid relative to the same time reference is represented by n_1. As illustrated in Fig. 4, n_0 is equal to Δ plus an integer number of periods of the harmonic sinusoid that fit within n_0. If the period of the harmonic sinusoid is represented by P_1, then $n_0 = \Delta + \lfloor n_0/P_1 \rfloor P_1$. The relative delay is therefore given by $d = n_1 - \Delta = n_1 - n_0 + \lfloor n_0/P_1 \rfloor P_1$. If $d < 0.0$, then $d = d + P_1$ so that d is always a positive number less than P_1. When d is divided by P_1 it becomes normalized between 0.0 and 1.0 and corresponds to the NRD.

Instead of describing a quasi-harmonic signal by using the triplet A_ℓ, ω_ℓ, and ϕ_ℓ, i.e., the magnitude, frequency and phase of all harmonic partials, the NRD concept allows to describe that signal by using the magnitude, the frequency and the delay of each partial of the harmonic

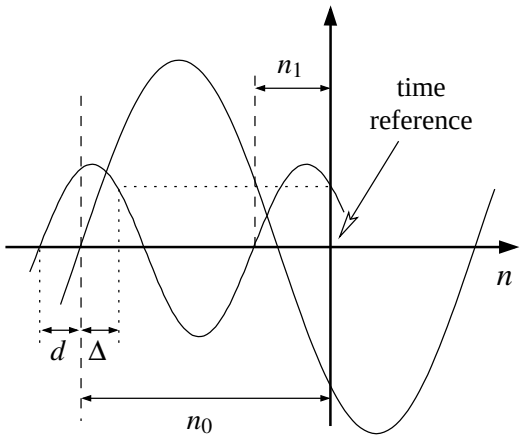

Fig. 4: Illustration of the relative delay d between a harmonic sinusoid whose delay to a time reference is n_1, and the fundamental frequency sinusoid whose delay to a time reference is n_0. (adapted from [13])

structure relative to the fundamental frequency sinusoid. The advantage is that the only variable time information is the delay of the fundamental frequency sinusoid since the delays of all partials are relative to it. Thus, the NRD consists in an important signal feature characterizing the shape of the waveform of that signal, independently of its overall time shift, and independently of its fundamental frequency. This is extremely important not only for signal analysis and identification, but also for signal transformation since the NRD of one harmonic signal may be imprinted on any other harmonic signal of a different pitch frequency. Provided that the magnitude and frequency relations among all partials are preserved, shape invariance will also be preserved. The algorithm computing the NRDs is further detailed in [13] and requires accurate frequency, phase and magnitude estimation [14, 15, 16] of sinusoids.

5. SPEAKER IDENTIFICATION TESTS USING MFCC AND NRDS

Taking advantage of the conclusions obtained in section 3, the objective of the tests described in this section is to determine the speaker identification accuracy for short vowel records only, obtained with Normalized Relative Delay features, and to determine whether these features, in combination with MFCC features, can improve the performance achieved with MFCCs alone.

Given that the focus is on speaker identification using very short voice segments (in the order of 200 ms), we use two databases of pre-segmented singing voice vowels and normal voice (i.e., spoken) vowels. The first database is described in [5] and consists of 40 singing voice recordings. The recordings were extracted from *arpeggio* vocalizations performed by 8 singers. For each singer, a stable segment of 200 ms pertaining to each one of the sung vowels [a], [e], [i], [o] and [u], was manually selected. The second database is described in [17] and consists of records of 5 spoken vowels by 44 speakers, 27 of whom are children and 17 are adults (11 females and 6 males). For each vowel, only a stable segment of 100 ms was manually selected and entered into the database. Thus, in total, this database includes 220 short vowel records. In all cases the sampling frequency is 22050 Hz.

The database consisting of sung vowels was first used. Because of the small amount of data available in this database, an approach based on GMMs, such as the one presented in section 3, is clearly not appropriate. Thus, we decided to use a classification method consisting of the Nearest Neighbor generalized classifier ('NNge'), which is available in the publicly available Data Mining tool, Weka. This approach was selected because our preliminary tests indicated that among the simplest classifiers available in Weka, the NNge consistently provided better results.

The MFFCC feature vectors were obtained using the publicly available Voicebox Matlab toolbox, and the NRD feature vectors (of 5 NRD coefficients since for some vowels only a rather small number of harmonics is effectively detected) were obtained as detailed in [13].

The performance results obtained for the first database are summarized in Table 1. It can be concluded that the NRDs, when used in combination with MFCCs, are able to slightly improve the percentage of correct identifications. In fact, the percentage obtained with MFCCs alone was 97%, and with the addition of NRDs that percentage increased to 98%. Alone, NRDs achieve 94% correct identifications which is a very significant result indicating that phase information reflects important speaker individuality characteristics. These results are quite in line with those obtained in [5], despite the fact that the classification approach is not exactly the same. It should be emphasized that in this experiment the number of different speakers (or more specifically: singers) is 8.

Table 1: Percentage of correct identifications using MFCC and NRD features, extracted from segments included in the database and consisting of sung vowels.

Features	% of correct identifications
MFCCs	97%
NRDs	94%
MFCCs+NRDs	98%

Concerning the second database (which includes 44 different speakers), the performance results in speaker identification are presented in Table 2. This table shows that the NRDs achieve an improvement in performance of 7%, when compared to the performance obtained with the MFCCs features only. Despite the fact that all classification results are below 90%, which is certainly due to the fact that this database includes a large number of speakers (44), and also to the fact that a significant fraction of the speakers are 'difficult' speakers (i.e., children whose voice spectra is not only very sparse because of the typically high pitch, but is also somewhat unstable since the vocal tract apparatus is not mature), it should be emphasized that NRDs alone achieve a remarkable identification performance. This confirms that speaker phase-related information denotes important idiosyncratic characteristics.

Table 2: Percentage of correct identifications using MFCC and NRD features, extracted from segments included the database and consisting of spoken vowels.

Features	% of correct identifications
MFCCs	82%
NRDs	76%
MFCCs+NRDs	89%

Overall, these results show that the NRDs, despite reflecting phase information only, are highly speaker discriminative features and that possess important complementary information to the information provided by the MFCCs (which reflect spectral magnitude information only).

6. CONCLUSION

We have described in this paper a number of experiments involving speaker identification tasks in which we

have included phase-related features in addition to spectral magnitude-related features for speaker modelling. The pertinence of this approach has been first highlighted by reviewing the source-filter model of speech production and by emphasizing, in particular, that the glottal source signal contains important speaker idiosyncratic information which is imprinted in the voiced regions of the speech. We have confirmed this perspective in MFCC+GMM speaker modelling and classification tests. These tests have revealed that if a careful phonetic segmentation of the speech signal is performed by just retaining the vowel regions, the automatic speaker identification performance is consistently better than if speaker modelling and classification is based on speech segmentation involving only voiced regions, only unvoiced regions, or the complete speech signal with silence regions removed. We have briefly reviewed the concept of Normalized Relative Delays which we use as phase-related features denoting speaker idiosyncratic information, namely concerning the glottal source. We also described specific speaker identification experiments we have conducted, by just using for speaker modelling and classification, pre-segmented and very short vowel regions of the singing or speech by different speakers. In these experiments, we have used as signal features both MFCCs (which denote spectral magnitude information) and NRDs (which denote spectral phase-related information). Given the small amount of data resulting from the phonetic segmentation of the speech, we have used a Nearest Neighbour generalized classifier (within the Weka environment) for signal modelling and classification. Results have revealed that NRDs alone achieve very significant and competitive speaker identification scores (in the order of 70% or better) and that, when combined with MFCCs, help to improve the speaker identification performance (by as much as 7%), especially in the case of spoken vowels by a significant number of different speakers (44, to be precise), and despite the fact that the pre-segmented vowel regions are just about 100 ms long. Overall, these results confirm previous conclusions by other authors that the phase information of voiced regions of the speech should be taken into consideration in order to improve the speaker identification performance, especially when the amount of speech data is small, as it frequently happens in audio forensics.

7. REFERENCES

[1] N. Ratha, A. Senior, and R. Bolle, "Automated biometrics", in *Proceedings of ICAPR*, 2001, pp. 445–454, Rio-Brazil.

[2] Davide Maltoni, "A tutorial on fingerprint recognition", *Lecture Notes in Computer Science (Advanced Studies in Biometrics)*, vol. 3161, pp. 121–138, 2005.

[3] Joseph P. Campbell, "Speaker recognition: A tutorial", *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[4] Haizhou Li and Bin Ma, "Techware: Speaker and spoken language recognition resources", *IEEE Signal Processing Magazine*, pp. 139–142, November 2010.

[5] Ricardo Sousa and Aníbal Ferreira, "Singing voice analysis using relative harmonic delays", in *12th Annual Conference of the International Speech Communication Association (Interspeech-2011)*, 2011, pp. 1997–2000.

[6] Inma Hernáez, Ibon Saratxaga, Jon Sanchez, Eva Navas, and Iker Luengo, "Use of the harmonic phase in speaker recognition", in *12th Annual Conference of the International Speech Communication Association (Interspeech-2011)*, p. 2757.

[7] G. Fant, *Acoustic Theory of Speech Production*, The Hague, 1970.

[8] Harvey Richard Schiffman, *Sensation and Perception*, John Wiley and Sons, Inc., 1989.

[9] Sandra Dias, Ricardo Sousa, and Aníbal Ferreira, "Glottal inverse filtering: a new road-map and first results", in *Speech Processing Conference*, June 2011, Tel-Aviv, Israel.

[10] Jared J. Wolf, "Efficient acoustic parameters for speaker recognition", *Journal of the Acoustical Society of America*, vol. 51, no. 6B, pp. 2044–2056, 1972.

[11] D. E. Sturim, W. M. Campbell, and D. A. Reynolds, "Classification methods for speaker recognition", *Lecture Notes in Computer Science (Speaker Classification I)*, vol. 4343, pp. 278–297, 2007.

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

[13] Ricardo Sousa and Aníbal Ferreira, "Importance of the relative delay of glottal source harmonics", in *39th AES International Conference on Audio Forensics - practices and challenges*, 2010, pp. 59–69.

[14] Aníbal J. S. Ferreira, "Accurate estimation in the ODFT domain of the frequency, phase and magnitude of stationary sinusoids", in *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 21-24 2001, pp. 47–50.

[15] Aníbal Ferreira and Deepen Sinha, "Accurate and robust frequency estimation in the ODFT domain", in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2005, pp. 203–206.

[16] Ricardo Sousa and Aníbal J. S. Ferreira, "Non-iterative frequency estimation in the DFT magnitude domain", in *4th International Symposium on Communications, Control and Signal Processing*, March 2010.

[17] Aníbal J. S. Ferreira, "Static features in real-time recognition of isolated vowels at high pitch", *Journal of the Acoustical Society of America*, vol. 112, no. 4, pp. 2389–2404, October 2007.