Speaker Identification Using Instantaneous Frequencies

Marco Grimaldi and Fred Cummins

***Abstract*—This paper presents an experimental evaluation of different features for use in speaker identification. The features are tested using speech data provided by the CHAINS corpus, in a closed-set speaker identification task. The main objective of the paper is to present a novel parametrization of speech that is based on the AM–FM representation of the speech signal and to assess the utility of these features in the context of speaker identification. In order to explore the extent to which different instantaneous frequencies due to the presence of formants and harmonics in the speech signal may predict a speaker's identity, this work evaluates three different decompositions of the speech signal within the same AM–FM framework: a first setup has been used previously for formant tracking, a second setup is designed to enhance familiar resonances below 4000 Hz, and a third setup is designed to approximate the bandwidth scaling of the filters conventionally used in the extraction of Mel-fequency cepstral coefficients (MFCCs). From each of the proposed setups, parameters are extracted and used in a closed-set text-independent speaker identification task. The performance of the new featural representation is compared with results obtained adopting MFCC and RASTA-PLP features in the context of a generic Gaussian mixture model (GMM) classification system. In evaluating the novel features, we look selectively at information for speaker identification contained in the frequency range 0–4000 Hz and 4000–8000 Hz, as the instantaneous frequencies revealed by the AM–FM approach suggest the presence of structures not well known from conventional spectrographic analyses. Accuracy results obtained using the new parametrization perform as well as conventional MFCC parameters within the same reference system, when tested and trained on modally voiced speech which is mismatched in both channel and style. When the testing material is whispered speech, the new parameters provide better results than any of the other features tested, although they remain far from ideal in this limiting case.**

***Index Terms*—AM-FM representation, instantaneous frequency, speaker identification, speaker recognition.**

I. Introduction

HUMANS are fairly good at identifying speakers based on their voices alone. The large amount of work in the field of speaker recognition over the previous 30 years has been predicated on the belief that automated systems ought to be able to do as well, or even better, than humans. Yet we still lack a solid understanding of those characteristics of speech that index an utterance as originating in one speaker rather than another. The introduction of the idea of a *voice print* by Kersta [21] led to a common perception that the now familiar spectrogram could function in much the same way as a fingerprint. Phoneticians are well aware, however, of the inherent variability in the speech signal, such that no two utterances are ever identical. Moreover, much of what we know about speech comes from an intellectual focus upon features which make linguistic communication possible, which necessarily ignores the idiosyncratic properties of speech that differentiate linguistically similar utterances. The goal of work in speaker recognition, on the other hand, is to find measurable quantities that minimize within-speaker variability and simultaneously maximize between-speaker variability [1], [18], [38].

The general area of speaker recognition encompasses two fundamental tasks: speaker identification and speaker verification [3], [7], [35]. Speaker identification is the task of assigning an unknown voice to one of the speakers known by the system: it is assumed that the voice must come from a fixed set of speakers. Thus, the system must solve a n-class classification problem and the task is often referred to as closed-set identification. On the other hand, speaker verification refers to the case of open-set identification: it is generally assumed that the unknown voice may come from an impostor. Regardless of the specific task at hand, it is common practice to adopt a probabilistic approach that predicts the likelihood that a given speech sample belongs to a given speaker [3], [7], [27], [34]–[36]. The base system for speaker recognition is usually composed of a speech parametrization module and a statistical modeling module [3] which are responsible for the production of a machine readable parametrization of the speech samples and the computation of a statistical model from the parameters. The main difference between speaker identification and speaker verification is that in the first case the system provides one model for each speaker, while, in the second case, the system provides a total of two models: one for the hypothesized speaker and one representing the hypothesis that the speech sample comes from some other speaker—the background model.

The base model for speech parametrization (the front-end of the recognition system) usually adopted is the source-filter model, which leads to the extraction of parameters such as linear predictive coding (LPC), Mel-frequency cepstral coefficients (MFCCs), perceptual linear prediction (PLP) coefficients, etc. (e.g., [8], [17], [24], [27], [34]–[36]). While these parameters have proved highly successful in robust speech recognition, the amplitude spectrum typically employed is highly sensitive to changes in speaking conditions such as changing channels and

Manuscript received February 1, 2008; revised May 2, 2008. Published July 16, 2008 (projected). This work was supported by the Science Foundation Ireland under Grant 04/IN3/I568 to F. Cummins. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mary P. Harper.

The authors are with School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland.

Digital Object Identifier 10.1109/TASL.2008.2001109

speaking style [6], [35], [38]. This is well illustrated by the so-called *great divide* in the KING corpus [15]. Minor changes in the recording setup of the corpus introduced differences in the spectral slope of the data, badly affecting the system performance. To compensate for the shortcomings of standard speech parametrization, researchers often opt to train the decision algorithm using a mix of samples from different sessions/setups [15] and implementing different forms of channel normalization such as cepstral mean normalization or RASTA processing [3], [17], [34], [36]. While these methodologies have potential in some speaker verification scenarios (where it is possible to perform multisession recordings), the same is not true in areas where researchers and investigators do not have full control of the devices used for recording, or of the recording environment, or the amount of recorded material. Moreover, many research studies have stressed the fact that further difficulties may arise due to (possibly volitional) changes in speaking register and speaking style [1], [6], [18], [38].

In the context of text-independent speaker recognition, where there is no prior knowledge of what the speaker will say, one of the most successful methods in modeling human identity from speech has been the Gaussian mixture model (GMM) [3]. The GMM is usually viewed as a hybrid approach between parametric and nonparametric density models: like a parametric model, it has structure and parameters that control the behavior of the density in known ways, but without the constraint that the data must follow a specific distribution [3], [34], [36]. In the literature, other approaches have been proposed to model human identity from speech (e.g., artificial neural networks and support vector machines, [8], [40]–[42]), demonstrating that other techniques may provide comparable performance.

The main objective of this paper is an experimental comparison of different parametrizations of speech for speaker recognition. In order to reduce the degrees of freedom of the problem, we focus on the performance that a generic GMM classification system provides in closed-set identification (speaker identification, hereafter SI). The closed-set scenario is preferred to the open-set scenario (speaker verification), because of the open nature of the latter. As argued by Bimbot *et al.* [3, pp. 435–436], when building a background model for verification, speech should be selected such that it reflects the expected alternative speech to be encountered during the verification procedure. This applies to the type and quality of speech as well as the composition of the set of speakers. Moreover, other than general guidelines and experimentation, there is no objective measure to determine the optimal number of speakers or amount of speech to use in training a background model [3]. On the other hand, by focusing on closed-set identification, differences in the performance of the reference classification system may be directly related to a better separation of the speaker in the feature space, or lack of it. Furthermore, given that both speaker identification and speaker verification rely on the same base techniques, improvements in one technology may also provide improvements in the other.

The speech samples used in this work are extracted from the CHAINS corpus [9]. The CHAINS corpus is a novel speech corpus designed to facilitate research into the characterization of individual speakers. It offers speech samples obtained in two different recording sessions with the unique possibility of studying the effect of nonmodal voice (whispering) in SI. The first recording session used a very high quality recording environment and apparatus, while the second recording session is more typical of data recorded in a quiet office using a near-field microphone. Across the two sessions, each speaker provides recordings in six qualitatively different speaking styles. In this paper, we make use of speech samples selected across the two recording sessions, training and testing GMMs in mismatched channel and speaking style conditions. Speech samples extracted from the first, high-fidelity, recording session are used to train the induction algorithm, while material extracted from the second, lower quality recording session is used in testing the algorithm. The training material consists of utterances recorded from subjects reading a prepared text aloud, while the test material consists of speech samples obtained from subjects who willfully alter their speaking style. A first set of experiments uses test material in which the prepared text is read aloud at a fast rate; a second set of experiments uses test material in which the prepared text is read in a whisper. While our main focus is on the evaluation of a novel parametrization of speech for the purposes of SI, this work also attempts to address outstanding challenges in SI by making use of recordings in which the speakers intentionally modify their voices [37] and to examine the effect of nonmodal voice (whispering) on the performance of the SI system.

This paper introduces a new set of descriptors based on the AM–FM representation of the speech signal. This new signal characterization is obtained by extending the use of the *pyknogram* of the signal [30]. It seeks to exploit the rich set of time-varying frequencies inherent in the speech signal, and to encode them as parameters for speaker identification. Well-known frequency components of voice include both formant resonances and harmonics of the fundamental frequency. Other frequency components of unknown origin (bony structures, idiosyncratic anatomy) may also prove useful in identifying individuals. We therefore explore three alternative ways of parametrizing speech within a single AM–FM framework: the first setup we employ has been used in the literature for formant tracking [30], a second setup was found, during preliminary tests, to enhance familiar resonances below 4000 Hz, and a third setup is designed to mimic the frequency scaling of the filters adopted in the extraction of MFCC coefficients. The recognition rates obtained using the three AM–FM setups are compared with the results obtained adopting MFCC and RASTA-PLP parameterizations within the same generic GMM classification system.

II. AM–FM MODEL

Popular speech processing techniques are conventionally based on spectral analysis, using some variant of linear prediction or cepstral analysis [2], [13], [23]. These parametrizations are commonly used as a front-end for both speech recognition and for speaker identification/verification systems and are based (in one way or another) on the source-filter model of speech production [32], [33]. The source-filter model assumes that the sound source in modally voiced speech is localized in the larynx, while the vocal tract acts as a convolution filter

for the emitted sound. Although this approach has led to great advances in the last 30 years, it is known to neglect some structure present in the speech signal [12]. Examples of phenomena not well-captured by the source–filter model include unstable airflow, turbulence, and nonlinearities arising from oscillators with time-varying masses [12], [25], [39].

In recent years, new ways of modeling and characterizing speech have been proposed in a number of different works (e.g., [10]–[12], [19], [28], [29], [31]). Of particular interest here is AM–FM signal modeling. AM–FM modeling is a technique used especially by electrical engineers in the context of frequency modulated signals, such as FM radio signals. This technique has been applied to speech signal analysis, with varying degrees of success, in areas such as formant tracking [30], speech synthesis [22], speech recognition [10], [11], [28], [29], [31], and speaker identification [19]. The AM–FM model is generally used to decompose a speech signal into decorrelated bandpass channels, each of which is characterized in terms of its envelope (instantaneous amplitude) and phase (instantaneous frequency).

In order to characterize a (single) instantaneous frequency for a real-valued signal, an analytic signal is first constructed: it is a transformation of the real signal into the complex domain and it is adopted because it permits the characterization of the real input in terms of instantaneous amplitude and frequency [14], [33]. More formally: given a real input signal $s(t)$, its analytic signal $s_a(t)$ can be computed as

$$s_a(t) = s(t) + j \cdot \hat{s}(t) \tag{1}$$

where $\hat{s}(t)$ is the Hilbert transform of $s(t)$. The analytic signal $s_a(t)$ can be decomposed as follows:

$$s_a(t) = a(t) \cdot e^{j\phi(t)} \tag{2}$$

where $a(t)$ is the instantaneous amplitude (envelope) of the analytic signal, while $\phi(t)$ is its phase. The instantaneous frequency (IF) $f(t)$ of the analytic signal can be computed directly from the phase

$$f(t) = \frac{1}{2\pi} \cdot \frac{d\phi(t)}{dt}. \tag{3}$$

The importance of the instantaneous frequency (IF) stems from the fact that speech is a *nonstationary* signal with spectral characteristics that vary with time.

When a person speaks, the supra-glottal vocal tract modifies the sound wave originating in the vocal cords in very specific ways. Depending on the message embodied in the speech itself and on the anatomy of the vocal cavities of the speaker, the final result of the act of speaking is a very complex pressure signal containing many different forms of information. Different parts of the skull (soft parts and hard parts of the vocal tract) vibrate under the influence of a common energy generator (the lungs) and the consequent airflow across the glottis. The human ability to change the form of some parts of the vocal tract (e.g., the cross section of the larynx, the size of the mouth cavity) gives rise to the familiar dynamic formant structure found in speech. Changing the stiffness of the vocal chords causes changes in *pitch*. The presence of diseases (e.g., colds) in a speaker has the effect of enhancing the role played in speech production by some structures of the vocal tract (e.g., the nasal cavity). A speaker may voluntarily change some of his speaking habits in order to achieve a predefined goal, e.g., mimicry or disguise [37]. A speaker may involuntarily change some properties of his vocal tract under the influence of a particular state of mind, e.g., anger, stress, sadness. All these intrinsic factors in speech production affect the physical properties of the speech waveform: they collectively ensure that speech is a signal with a rich set of multicomponent time-varying frequencies.

In the context of AM–FM signal modeling, the concept of a single-valued instantaneous frequency for a multicomponent signal becomes meaningless without breaking the signal down into its components. As discussed in [4], the decomposition of a signal is not unique if its frequency components coincide at some points in the time–frequency plane. This is the case for speech, e.g., formants are well known to have points in the time–frequency plane where they appear to join or split. In this case, the decomposition is heuristic in nature and its optimal form will depend on the specific application. [4].

An example of AM–FM signal decomposition for speech analysis was proposed by Potamianos *et al.* [30] in the context of formant tracking. The authors proposed a new way of representing the speech signal through the computation of the *pyknogram*, which is a density plot of the frequencies present in the input signal. The *pyknogram* is computed by the authors using a uniformly spaced Gabor filterbank and a demodulation schema based on the Teager energy-tracking operator [20], [25]. A single instantaneous frequency is computed for each filter output and each frame of speech, yielding a 2-D array of instantaneous frequencies. Potamianos *et al.* demonstrate that the AM–FM approach can overcome some of the limitations of the classic source-filter model and greatly help the difficult task of speech formant tracking.

Jankowski *et al.* [19] adopted the AM–FM model to characterize some fine structures of the human voice for the purpose of speaker identification. The authors proposed a mixed LPC/AM–FM approach to identify, select, and parametrize the first three formants of human speech. Their work demonstrated that this procedure may be beneficial in speaker identification: formant AM–FM parameters substantially improve identification rates on female speakers [19]. A more thorough comparison of AM–FM representations in an SI context has hitherto remained outstanding.

In this paper, we extend the use of the pyknogram and the decomposition of the multifrequency component speech signal to the problem of speaker identification. The approach proposed here seeks to identify *which* instantaneous frequencies are present in the speech signal and to encode them as parameters for speaker identification. Given the nonunique decomposition of a multifrequency component signal and the lack of a theory justifying the adoption of any specific approximation to extract parameters which are invariant in the context of SI, we perform speech parametrization by applying three different decomposition algorithms for the same input signals. Within the same approach (AM–FM), it is possible to vary the way in which the speech pyknogram is computed in order to focus on different

frequencies in the signal: structures due to the presence of resonators in the upper vocal tract (formants) and fine structures related to the presence of quasi-harmonic vibrations within the whole vocal tract.

In the next section we present the main structure of the algorithm used to calculate the pyknogram and the ad hoc modifications applied to selectively attend to different information embodied in the signal.

A. Computing the Pyknogram of the Signal

In order to compute the instantaneous frequencies of the speech signal (and its pyknogram), a *multiband demodulation analysis* (MDA) must be performed. Similar to the process described in [30], the MDA consists of a multiband filtering scheme and a demodulation algorithm. First, the speech signal is bandpassed with the use of a filterbank, and then each bandpass waveform is demodulated and its instantaneous amplitude and frequency computed.

The filterbank adopted consits of a set of Gabor bandpass filters with center frequencies that are uniformly spaced on the frequency axis. The filter bandwidth is either constant or variable depending on which frequency structures we seek to extract (Section II-B). Gabor filters are chosen because they are optimally compact and smooth in both the time and frequency domains. This characteristic guarantees accurate amplitude and frequency estimates in the demodulation stage [30] and reduces the incidence of ringing artifacts in the time domain [19].

The demodulation schema adopted is based on the Hilbert transform demodulation (HTD). Although other demodulation schema are less computationally expensive (e.g., the DESA family of algorithms [25], [30]), the HTD can give smaller error and smoother frequency estimates [25], especially when the first formant is close to the fundamental frequency [30].

The main schema adopted to demodulate the speech signal can be summarized as follows.

- The speech signal $s(t)$ is bandpass filtered and a set of waveforms $w_i(t)$ is obtained (i denotes the output of the ith filterbank).
- For each bandpass waveform $w_i(t)$, its Hilbert transform $\hat{w}_i(t)$ is computed.
- The instantaneous amplitude for each bandpass waveform $a_i(t)$ is computed as

$$a_i(t) = \sqrt{w_i^2(t) + \hat{w}_i^2(t)}. \tag{4}$$

- The instantaneous frequency for each bandpass waveform $f_i(t)$ is computed as the first time derivative of the phase $\phi_i(t)$

$$f_i(t) = \frac{1}{2\pi} \cdot \frac{d\phi_i(t)}{dt} = \frac{1}{2\pi} \cdot \frac{d}{dt} \left[\arctan\left(\hat{w}_i(t)/w_i(t)\right)\right]. \tag{5}$$

Instantaneous amplitude and instantaneous frequency are combined together to obtain a mean-amplitude weighted short-time estimate F_i of the instantaneous frequency for each $w_i(t)$

$$F_i = \frac{\int_{t_0}^{t_0+\tau} \left[f_i(t) \cdot a_i^2(t)\right] dt}{\int_{t_0}^{t_0+\tau} \left[a_i^2(t)\right] dt} \tag{6}$$

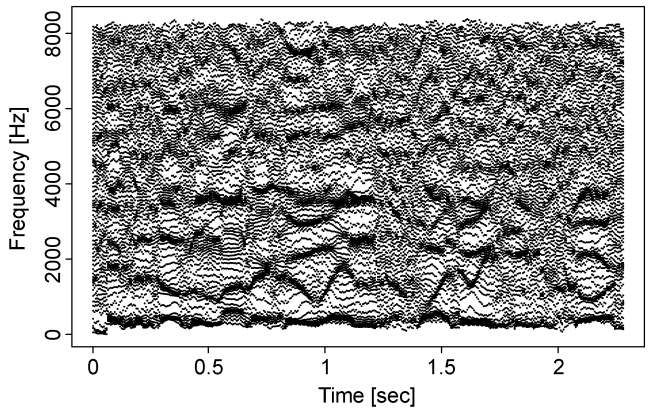

Fig. 1. Pyknogram of "If it doesn't matter who wins, why do we keep score?". 80 filters linearly spaced between 200 and 8200 Hz, constant bandwidth of 400 Hz—Filterbank *setup 1*. See also Fig. 4 for a conventional spectrographic representation.

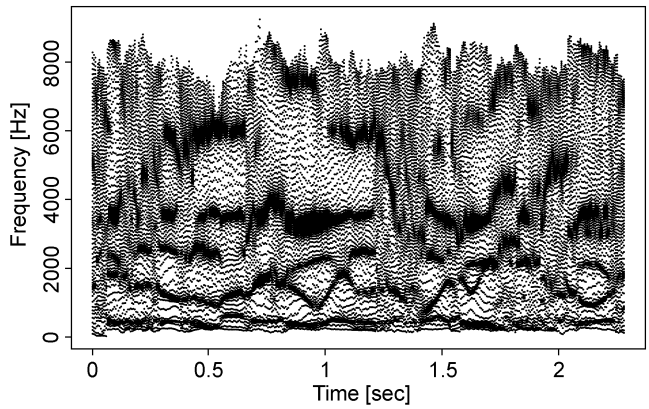

Fig. 2. Pyknogram of "If it doesn't matter who wins, why do we keep score?". Eighty filters linearly spaced between 200 and 8200 Hz, constant bandwidth of 266 Mel—Filterbank *setup 2*.

where τ is the selected length of the time-frame. The short-time frequency estimate is computed for the full length of each $w_i(t)$, with an overlap window of $\tau/2$. The adoption of a *mean amplitude weighted instantaneous frequency* (6) is motivated by the fact that it provides more accurate frequency estimates and is more robust for low energy and noisy frequency bands when compared with an unweighted frequency mean [30].

The computation of the short-time instantaneous frequency for each bandpass waveform i leads to the extraction of the pyknogram of the speech signal. Examples of pyknograms are shown in Figs. 1–3.

The pyknogram, being a density plot of the frequencies present in the input signal, reveals the presence of strong resonances as regions of high density of the estimated short-time frequencies. Harmonic-like structures are identified as regions where the short-time estimates are placed with approximately equal frequency spacing.[1] Note that the pyknogram is *not* a 3-D plot similar to a spectrogram: each individual point represents the location of an (amplitude-weighted) instantaneous frequency, but not its intensity. It is also crucial that in our

[1]A special case is when the signal lacks a marked frequency component within a given band: the estimated instantaneous frequency is then simply the center-frequency of the filter.

approach, the bandwidths of the filters exhibit considerable overlap, allowing the center frequencies from multiple filters to converge on a single strong frequency. The details of the overlap are specific to the setup employed to calculate the pyknogram of the signal, as described in the next section.

B. Different Bandwidth, Different Frequency Structures

An extremely important role in the computation of the pyknogram (and in the encoding of the features derived from it) is played by the bandwidth of each individual Gabor filter within the filterbank. By varying the bandwidth of the Gabor filters, the short-time frequency estimates can be tuned to identify resonances (formants) or structures due to quasi-harmonic vibrations (e.g., during vowel production) in the vocal tract (e.g., the harmonics of the fundamental frequency).

Fine tuning the bandwidth of the filters used for speech analysis and speech characterization is a standard practice found in many approaches. The power spectrum of speech can be fine tuned (varying the bandwidth of the spectral analysis) to enhance structures due to presence of harmonics or formants. Generally, a *broadband* spectrogram is obtained by setting the bandwidth (through the time length of the analysis window) to a value of about 250 Hz; a *narrowband* spectrogram has typical bandwidth of about 50 Hz. The extraction of a conventional MFCC feature vector is based on the definition of a filterbank of triangular filters with uniformly spaced center frequencies and constant bandwidth on the Mel scale: this approach produces filters with variable bandwith. In order to tune the pyknogram to the speech resonances for formant tracking, Potamianos *et al.* [30] used a Gabor filterbank with constant bandwidth of 400 Hz.

In this paper, three different filterbanks are used to compute the pyknogram. In each case. the filterbank is composed of Gabor filters with center frequencies which are uniformly spaced on the Hertz scale while the bandwidths are defined as follows:

- *setup 1*: constant (on the Hertz scale) bandwidth of 400 Hz;
- *setup 2*: constant (on the Mel scale) bandwidth of 266 Mel;
- *setup 3*: constant (on the Mel scale) bandwidth of 106 Mel.

The first setup (*setup 1*) has been used in the literature [30] for formant tracking and is known to be of use in identifying structure due to the first four or five formants in the vocal tract (below about 4000 Hz). The second setup (*setup 2*) has been evaluated in our tests and found to reveal marked structures above 4000 Hz, while enhancing the formant structure below 4000 Hz. The third setup (*setup 3*) has been designed to replicate the bandwidth scaling of the triangular filters used for MFCC extraction. With this configuration it is possible to identify both the formants and the harmonic structures well known in the literature and present in the range of frequencies between 50 and 4000 Hz.

Figs. 1–3 show three different pyknograms obtained using the three different setups. The input sound file used is the same and has been extracted from the CHAINS corpus [9]: *irm01_s01_solo.wav* (*If it doesn't matter who wins, why do we keep score?*). Fig. 4 shows its spectrogram computed using *Praat* [5]. All three pyknograms are computed using 80 Gabor filters with uniform center frequency spacing between 200 and 8200 Hz. The bandwidth of the filters are set as described above. Thus, the three setups differ in the degree of overlap among the filters, and in the relation between overlap and frequency.

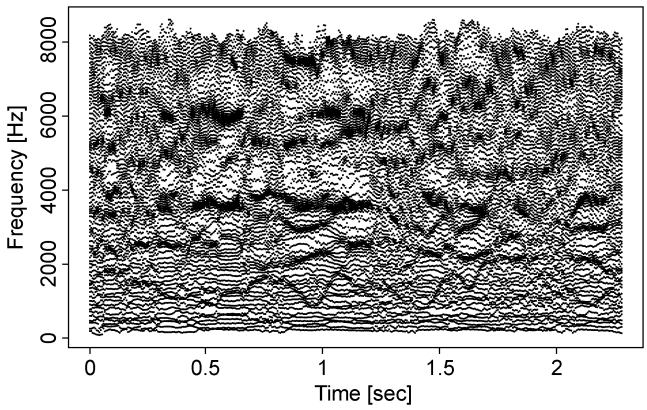

Fig. 3. Pyknogram of "If it doesn't matter who wins, why do we keep score?". 80 filters linearly spaced between 200 and 8200 Hz, constant bandwidth of 106 Mel—Filterbank *setup 3*.

The three different setups adopted in the calculation of the pyknogram of the speech signal reveal different instantaneous frequencies within the signal. Qualitatively, these differences are found not only in the range of frequencies commonly exploited for speaker and speech recognition (below 4000 Hz), but extend to the higher part of the frequency scale, between 4000 and 8000 Hz. In this region, Fig. 3 (*setup 3*) reveals instantaneous frequencies that differ greatly from the ones detected using the other two setups in the same frequency range.

It is interesting to note that standard techniques used in speech analysis are somewhat limited in detecting frequency structures above 4000 Hz: the majority of them are not visible in the spectrogram of the same speech file shown in Fig. 4.[2] On the other hand, the same formants between 0 and 4000 Hz are replicated almost exactly in each of Figs. 1–4.

Given the unknown nature of the high-frequency components of human speech, we will not try to motivate their presence. We will limit ourselves to encoding their differing values (Section III) and verifying their relative utility for speaker identification (Section VI). Moreover, Section VI-C presents a quantitative analysis of the importance of the different instantaneous frequencies in the region 4000–8000 Hz, derived according to the three setups, for the problem of speaker identification.

Table I summarizes the frequency bandwidth employed in *setup 1*, *setup 2*, and *setup 3* for selected Gabor filters.

In the next section, we describe the encoding of the pyknogram into the features used to characterize speech for speaker identification.

III. Speech Parametrization

A. Pyknogram Frequency Estimate Coefficients

The encoding of the pyknogram into a set of feature vectors is straightforward, given its short-time nature and the proba-

[2]No pre-emphasis is applied to the speech file to calculate the pyknograms in Figs. 1–3. Pre-emphasis of 6dB/octave is applied to calculate the spectrogram in Fig. 4.

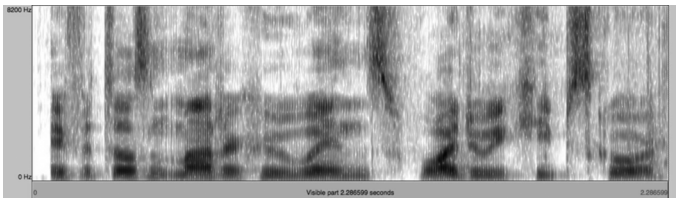

Fig. 4. Spectrogram of a speech file ("If it doesn't matter who wins, why do we keep score?"). The spectrogram shows frequencies between 0 and 8200 Hz.

TABLE I
CENTER FREQUENCY (CF) AND BANDWIDTH (BW) FOR SELECTED GABOR FILTERS USED IN "SETUP 1," "SETUP 2," AND "SETUP 3"

cf	*setup 1*		*setup 2*		*setup 3*	
$[Hz]$	bw $[Hz]$	bw $[Mel]$	bw $[Hz]$	bw $[Mel]$	bw $[Hz]$	bw $[Mel]$
200	400	510	200	266	85	106
400	400	414	260	266	100	106
600	400	350	305	266	120	106
1000	400	266	400	266	160	106
3600	400	105	1020	266	400	106
6200	400	65	1640	266	650	106
8200	400	50	2100	266	840	106

bilistic approach taken in speaker identification and verification. An utterance can be represented by an $N \times M$ matrix, where N corresponds to the number of Gabor filters used in the filterbank and M to the number of times the input signal can be segmented into frames with a given window length—the value of τ in (6). For each frame, N values corresponding to the estimated short-time frequency of each Gabor filter are encoded. This approach is very similar to the standard procedure used for, e.g., MFCC encoding; the main difference being that in the MFCC case, N corresponds to the first N discrete cosine transform (DCT) coefficients used to decorrelate the cepstrum of the signal. In our case, no DCT is applied to the pyknogram, while the frequency values are expressed in kilohertz. During preliminary tests, we found that expressing the short-time frequency estimates in kilohertz allows the feature set to work well together with a Gaussian mixture model classifier (Section IV). Hereafter, we refer to the features based on the estimates of the short-time frequency as *pykfec*: pyknogram frequency estimate coefficients. Our decision not to apply the DCT has the effect of introducing an element of redundancy due to correlation in our encoding, but as we show in Section VI-B, this does not, in fact, seem to adversely affect the accuracy of the reference GMM system.

As a first approximation, we standardize the choice of N (the number of Gabor filters in the filterbank) to 80, with center frequencies uniformly spaced on a linear scale between 200 and 8200 Hz; however, in Section VI-B, we examine the effect on SI of varying the number of filters within the same frequency range. Moreover, in order to test the importance of instantaneous frequencies extracted from different frequency ranges, we evaluate the speaker recognition rate provided by 40 pykfec computed using 40 Gabor filters uniformly spaced between 200 and 4200 Hz (Section VI-C), also varying the setup adopted in the calculation of the pyknogram. Finally, in order to quantitatively test the importance of the different instantaneous frequencies in the region 4000–8000 Hz, we also evaluate the speaker recognition rate provided by 40 pykfec computed using 40 Gabor filters uniformly spaced between 4200 and 8200 Hz (Section VI-C), varying the setup adopted in the calculation of the pyknogram.

The window length chosen to calculate the short-time frequency estimate is 1024 taps (about 23 ms for a signal sampled at 44 100 Hz) with an overlap between successive windows of 512 taps. No channel compensation schema is adopted.

B. MFCC

In order to provide a comparison to the performance obtained using pykfec in the task of speaker identification, we also use standard MFCCs with the same general GMM classifier. Since pykfec are extracted using very different setups and taking into account different frequency ranges, the MFCC are extracted according to three schema roughly matching the ranges analyzed in the pykfec extraction:

- MFCC-0-8000: MFCC extracted using 40 triangular filters spaced between 0 and 8000 Hz;
- MFCC-0-4000: MFCC extracted using 40 triangular filters spaced between 0 and 4000 Hz;
- MFCC-4000-8000: MFCC extracted using 40 triangular filters spaced between 4000 and 8000 Hz.

The number of coefficients used for identification is not selected *a priori*; it is evaluated experimentally (Section VI) in order to maximize the speaker recognition rate. The window length is again set to 1024 taps with an overlap between successive windows of 512 taps, to ensure homogeneous sampling between the two approaches (MFCC and pykfec) for the same input files. The zeroth cepstral coefficient is not used in the Mel-frequency cepstral feature vector, while the value of the coefficients is normalized using cepstral mean removal [34], [36], in order to compensate for the different recording channels used to train and subsequently test the induction algorithm.

C. RASTA-PLP

A second comparison is provided using RASTA-PLP coefficients [16], [17] for speech parametrization. PLP coefficients are a hybrid representation, using aspects of both a filterbank and an all-pole model spectral representation. The spectrum is first passed through a bark-spaced and trapezoidal-shaped filterbank and then fitted with an all-pole model. In this paper, the model order is not fixed, but it varies in order to maximize the recognition rate of the induction algorithm. The spectral representation is transformed to cepstral coefficients and a DCT applied as a final step. The zeroth cepstral coefficient is discarded as a form of energy normalization [34]. In order to compensate for channel effects, the input speech is preprocessed using RASTA filtering [17] before proceding to the extraction of the PLP feature vector. RASTA-PLP features are extracted in the frequency interval 0–8000 Hz and with a time window of about 0.23 ms (with a step equal to half of the time window).

IV. GAUSSIAN MIXTURE SPEAKER MODEL

The focus of this study is the utility of the various parametric representations of speech employed. Accordingly, a rather generic and simple classifier is employed, which is sufficiently powerful to discriminate among alternative features. No attempt is made to optimize the classifier however. Gaussian mixture

models are classifiers commonly used in speaker identification/verification, e.g., [27], [34]–[36]. This classifier is able to approximate the distribution of the acoustic classes representing broad phonetic events occurring in speech production (e.g., during the production of vowels, nasals, fricatives) and often outperforms other algorithms on the problem of speaker identification [36], [40], [42]. During training, the induction algorithm estimates the mixture of Gaussian models that best approximates the distribution of values produced by the SI system front-end for a given speaker.

Formally, a speaker is described by a mixture of M Gaussian models $\Gamma = \{\gamma_1, \ldots, \gamma_M\}$: the mixture density is a weighted sum of the M component densities. Given an input feature vector $\vec{x}$, the conditional probability is computed from the mixture as follows:

$$p(\vec{x}\ |\Gamma) = \sum_{m=1}^{M} c_m \cdot \gamma_m(\vec{x}) \tag{7}$$

where c_m are the mixture weights and $\gamma_m(\vec{x})$ an N-variate Gaussian function. The dimensionality of the Gaussian function (N) coincides with the dimensionality of the feature vector $\vec{x}$, while the M models, and relative weights, are estimated from the training data using a special case of the *expectation-maximization* (EM) algorithm [36].

A set of S speakers $\{s_1, \ldots, s_S\}$ is represented by S Gaussian mixture models $\{\Gamma_1, \ldots, \Gamma_S\}$. A given observation sequence $X = \{x_1, \ldots, x_T\}$ is tested by finding the speaker model which has maximum *a posteriori* probability. By applying Bayes rule and using the logarithm [36], the probability can be computed as

$$\hat{S} = \arg\max_{1 \le s \le S} \sum_{t=1}^{T} \log p(\vec{x}_t\ |\Gamma_s). \tag{8}$$

To evaluate different test utterance lengths, the sequence of feature vectors is divided for each speaker into a sequence of overlapping segments of t_0 feature vectors [36]. Each segment is treated as a separate test utterance. For a test set containing the observation sequence $X = \{x_1, \ldots, x_T\}$, two adjacent segments would be

$$\text{segment}_n = x_n, \ldots, x_{n+t_0}$$

and

$$\text{segment}_{n+1} = x_{n+1}, \ldots, x_{n+t_0+1}$$

where $n + t_0 + 1 \le T$.

The final performance evaluation is computed as the ratio of correctly identified sequences, over the total amount of sequences for the whole set of speakers.

The GMM can assume different forms depending on the choice of covariance matrix used in the estimation of the N-variate Gaussian functions. In this paper, nodal, diagonal covariance matrices are used to model the speakers [36], [40], [42].

A. Algorithm Details

There are some known shortcomings of the GMM induction algorithm, due largely to the initialization procedure used to estimate variance and mean of each Gaussian in the mixture, and to the maximum and minimum values permitted for the variance. As pointed out by Reynolds *et al.* [36], the issue of initializing the GMM algorithm for SI is less challenging than in other areas [26]. In speaker identification, elaborate initialization schemes are not necessary for training Gaussian mixture models [36]. We apply a k-means clustering algorithm to find M clusters in the training data as an initialization procedure. The k-means search stops whenever a stable configuration is found.

On the other hand, the maximum and minimum values allowed for the variance of each Gaussian in the mixture are two parameters that play an important role in the training of the GMM algorithm. In preliminary tests using *pykfec* features, we found that if the nodal variance assumes big values ($\gg$100), the EM algorithm does not always converge and the identification accuracy of the GMM is badly affected. When the minimum nodal variance is not bounded, it may happen that the GMM models singularities due, e.g., to a lack of data to represent a specific speaker or to the presence of outliers in the training set caused by noisy data [36]. To overcome these two problems, we encode the *pykfec*, which express the estimate of short-time instantaneous frequency, in kilohertz. This simple stratagem guarantees that the maximum nodal variance does not exceed a few kilohertz, or the maximum bandwidth of the filters in the filterbanks (Table I). The minimum value for the nodal variance is set to 0.001, which corresponds to a minimum nodal variance of 1 Hz when *pykfec* are used. The same absolute value for minimum nodal variance is used when MFCCs are employed.

V. Database Description

The CHAINS corpus [9] contains the recordings of 36 speakers obtained in two different sessions with a time separation of about two months. The first recording session was carried out in a professional recording studio; speakers were recorded in a soundproof booth using a Neumann U87 Condenser microphone. The second recording session was carried out in a quiet office environment with a Shure SM50 head-mounted microphone connected to a Marantz PMD 670 Compact Flash recorder. Across the two recording sessions, each speaker provided recordings in six different speaking styles. In this paper, we make use of three different speaking styles: NORM,[3] in which speakers read a prepared text alone at a comfortable rate; FAST, in which the same prepared text was read at a fast rate; and WHSP, which is a whispered reading of the same material. Full details of the corpus are provided in [9]. The NORM condition, which belongs to the first recording session of the corpus, is used as training material while speech recorded in the second session in the FAST and WHSP styles is used as test material, in order to examine robustness with respect to both stylistic and channel characteristics.

[3]Which is referred to as SOLO in the CHAINS corpus.

From the speech materials, the CHAINS corpus offers, we select (for each of the speaking styles above) the first nine individual sentences as speech samples for testing (s1 to s9 in the corpus), while the next sentences (s10 to s33 in the corpus) are used to generate the training sets. The training of the GMMs is performed using only NORM recordings from the first recording session.

Four sets of the available speakers are employed (8, 16, 24 and 36 speakers respectively). The *8-speaker*, *16-speaker*, and *24-speaker* sets have equal numbers of males and females and speakers are drawn from the university population in Dublin (native speakers of Hiberno-English). The *36-speaker* set covers the whole set of available speakers in the corpus. Of these, 28 (16 male, 12 female) are from the Eastern part of Ireland (Dublin and adjacent counties). A further eight subjects (four male, four female) are from the U.K. or U.S.: two males from the U.K. and two males from the U.S.; three females from the U.S., and the last remaining female from the U.K. [9].[4]

VI. Experimental Evaluation

A. Methodology

The utility of our three novel encoding methods (*pykfec setup 1*, *setup 2*, and *setup 3*) is evaluated in a closed-set text-independent speaker identification task. We compare performance with encodings using MFCCs and RASTA-PLP coefficients. The evaluation is subdivided in three parts. In the first part (Section VI-B), we estimate the accuracy of the reference GMM system, taking into account two different recording channels and two speaking styles. The training material is extracted from the first recording session of the CHAINS corpus and comprises speech samples recorded in the NORM style; the test material is extracted from the second recording session and employs speech samples recorded in the FAST style. Results are obtained varying the number of parameters used to encode speech, the amount of training material used, the number of models in the Gaussian mixture and the number of speakers at hand. All the parameters (*pykfec*, MFCC, and RASTA-PLP) are extracted within the frequency range of 0–8000 Hz.

In the second part of the evaluation, we present results obtained by restricting the parameterization to two different frequency ranges: we compare the speaker recognition rate obtained extracting parameters from the frequency range 0–4000 Hz and from the range 4000–8000 Hz. This is done to quantitatively evaluate the importance of the instantaneous frequencies revealed by the AM–FM approach within the higher part of the frequency axis (the frequency range 4000–8000 Hz). The training material is extracted from the first recording session of the CHAINS corpus using speech samples recorded in the NORM style; the test material is extracted from the second recording session in which the speech was recorded in the FAST style. Results obtained using both MFCC and *pykfec* are reported.

In the third part of the evaluation, nonmodal (whispered) speech is used as testing material. This is done to provide a very hard test bed for the identification system and to assess the extent to which it is possible for a model trained on modally voiced speech to generalize to whispered speech. The training material is extracted from the first recording session of the CHAINS corpus (NORM); the test material is again extracted from the second recording session, but in the WHSP style. Results obtained using both MFCC and *pykfec* are reported and compared with the accuracy results reported in the first and second part of the evaluation.

In each case, the accuracy of the SI system is expressed as the mean of ten runs, each run having a different random initialization of the GMMs. The error of the identification score is calculated as twice the standard deviation of the mean, corresponding to a confidence interval of about 95%.

B. Varying Recording Channel and Speaking Style

In this section, we compare three novel parameterizations *pykfec* (*setup 1*, *setup 2*, and *setup 3*), along with more standard MFCC and RASTA-PLP encodings. The training material is extracted from the first recording session of the CHAINS corpus (Section V) with speech in the NORM style; the test material is from the second, lower fidelity, recording session, and uses speech spoken in a FAST style.

Fig. 5(a)–(d) summarizes the results obtained encoding speech with *pykfec* (*setup 1*, *setup 2*, and *setup 3*) extracted from the frequency range 0–8000 Hz (the Gabor filters are spaced linearly between 200 and 8200 Hz, with varying bandwith according to the selected setup).

Fig. 5(a) shows the accuracy of the reference GMM classifier varying the number of Gaussian components and varying the AM–FM setup adopted to extract the pyknogram of the signal. The total length of the training material is approximately 25 s per speaker, the number of speakers is 16, while utterances of 10 s are used for testing. Eighty *pykfec* are used to encode the speech signal.

Fig. 5(b) shows the accuracy of the GMM classifier varying the total length of training material per speaker. The number of speakers is 16, while utterances of 10 s are used for testing. Eighty *pykfec* are used to encode the speech signal.

Fig. 5(c) summarizes the accuracy of the GMM classifier while varying the number of speakers. The number of Gaussian components is 64, while utterances of 10 s are used for testing. The amount of training material is approximately 25 s per speaker. Eighty *pykfec* are used to encode the speech signal.

Finally, Fig. 5(d) shows the accuracy of the classifier while varying the number of features (*pykfec setup 1* and *setup 3*). The number of Gaussian components is 32; utterances of 5 s are used for testing. The amount of training material per speaker is approximately 25 s. The number of speakers is 16.

Fig. 5(a)–(d) indicates that a parametrization based on *pykfec setup 3* outperforms parametrizations based on both *setup 1* and *setup 2* [e.g., considering the *16-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, *pykfec setup 1* provides an accuracy of 83% $\pm$ 3%, *pykfec setup 2* provides an accuracy of 72% $\pm$ 5%, and *pykfec setup 3* provides an accuracy of 92% $\pm$ 3%, as plotted in Fig. 5(a)]. *Pykfec setup 3* shows the best identification performance and greater stability while increasing the number

[4]The bulk of the subjects are aged between 19 and 25 years: the mode of the data is equal to 21, the median value equal to 22 and the semi-interquartile range equal to 3.5.

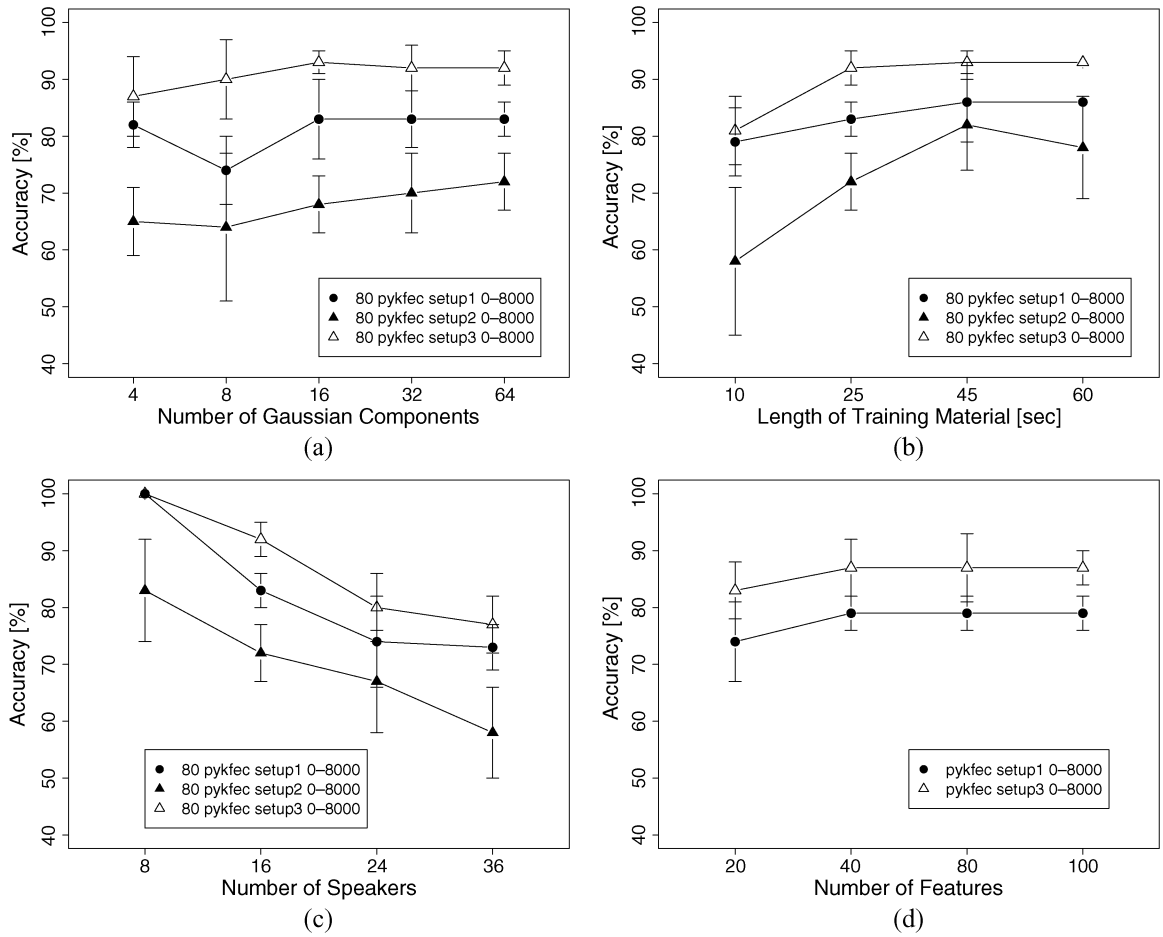

Fig. 5. Accuracy of the reference GMM classifier using *pykfec* for speaker parametrization, varying (a) the number of Gaussian components, (b) varying the length of the training material, (c) varying the number of speaker, and (d) varying the number of features. Parameters are extracted from the frequency interval 0–8000 Hz.

of Gaussian components in the reference classifier, increasing the amount of training material, and varying the number of features. Unsurprisingly, if we increase the number of speakers while keeping the amount of training material per speaker constant, the accuracy of the reference system decreases. However, *pykfec setup 3* still performs better than *pykfec setup 1* and *pykfec setup 2* in all experiments. *Pykfec setup 2* shows the worst recognition rate in all the experiments presented here. It is interesting to note [Fig. 5(d)] that in the case of *pykfec*, the reference system is capable of handling considerable redundancy in the features very well: increasing the number of parameters from 40 to 100 does not seem to produce any overfitting [training 32 GMMs per speaker, using 5-s test utterances and approximately 25 s of training material per speaker, *pykfec setup 2* provides an accuracy of about 79%, *pykfec setup 3* provides an accuracy of about 87%, as plotted in Fig. 5(d)].

Fig. 6(a)–(c) summarizes the results obtained encoding speech with MFCCs extracted from the frequency range 0–8000 Hz.

Fig. 6(a) plots classifier accuracy while varying the number of Gaussian components and the number of MFCCs used to encode the signal. The total length of the training material is approximately 25 s per speakers, the number of speakers is 16, while utterances of 10 s are used for testing.

Fig. 6(b) shows the accuracy curves of the classifier while varying the total length of training material per speaker. The number of speakers is 16, while utterances of 10 s are used for testing.

Fig. 6(c) shows the accuracy of the classifier while varying the number of speakers. The number of Gaussian components is 64 while utterances of 10 s are used for testing. The amount of training material is approximately 25 s per speaker.

Fig. 6(a)–(c) shows that the accuracy of the reference system trained with MFCCs is comparable to the recognition rate obtained using *pykfec setup 3* [e.g., considering the *16-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, *pykfec setup 3* (80 features) provide an accuracy of 92% $\pm$ 3%, 25 MFCCs provide an accuracy of 88% $\pm$ 4%, and 35 MFCCs provide an accuracy of 88% $\pm$ 2%, as shown in Fig. 5(a) and Fig. 6(a)]. However, when MFCCs are adopted for speech encoding, the stability of the GMMs across the various experimental setups is somewhat uneven: increasing the number of cepstral coefficients used to encode the speech may, under some circumstances, harm the recognition rate. Fig. 6(a) shows that both 35 MFCC and 25 MFCC parameterizations yield equivalent accuracy scores while varying the number of Gaussian components in the classifier. On the other hand, Fig. 6(b) suggests that when

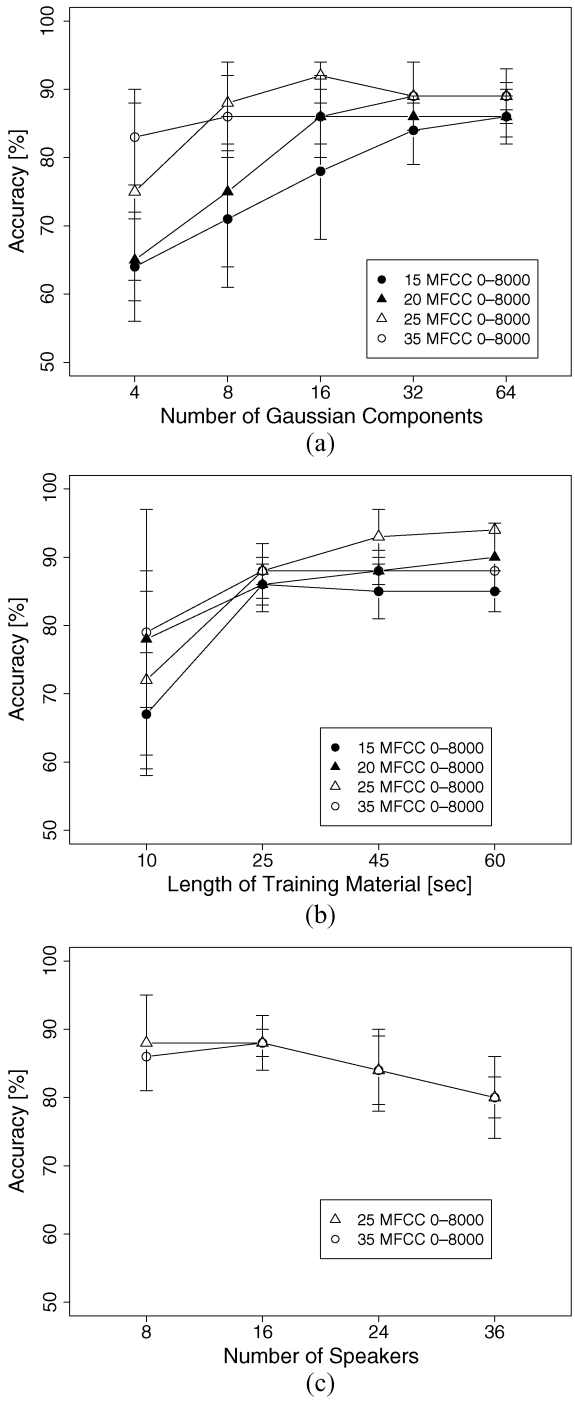

Fig. 6. Accuracy of the reference GMM classifier using MFCC for speaker parametrization, varying (a) the number of Gaussian components, (b) varying the length of the training material, and (c) varying the number of speakers. Parameters are extracted from the frequency interval 0–8000 Hz.

increasing the amount of training material, 25 MFCCs provide better performance than 35 MFCCs: training the GMM classifier with about 60 s of material per speaker and 64 Gaussian components per speaker, 25 MFCCs provide an accuracy of 94% $\pm$ 1%, and 35 MFCCs provide an accuracy of 88% $\pm$ 1%. This is not the case for *pykfec setup 3*: the trained classifier does not show signs of overfitting due to the increased number of features, as shown in Fig. 5(d).

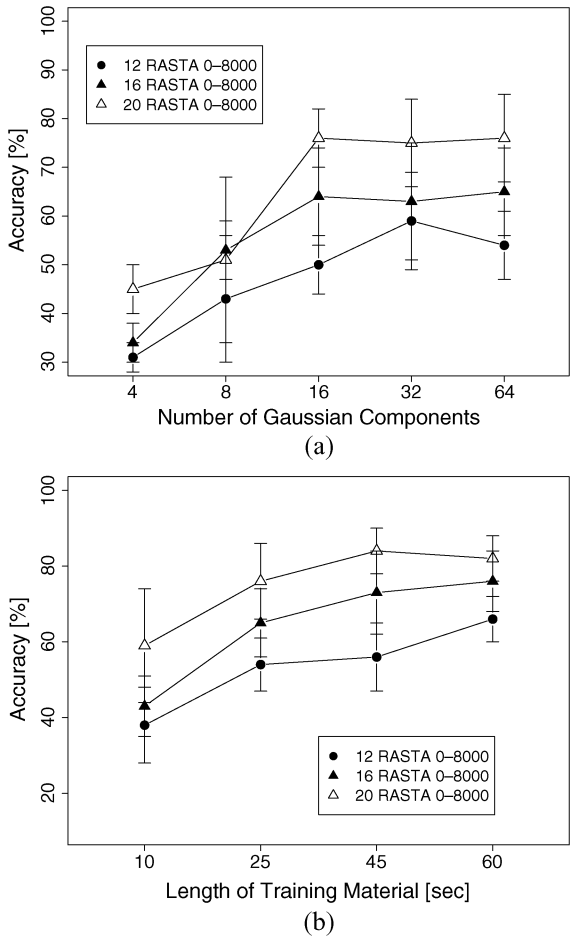

Fig. 7. Accuracy of the reference GMM classifier using RASTA-PLP for speaker parametrization, varying (a) the number of Gaussian components and (b) varying the length of the training material. Parameters are extracted from the frequency interval 0–8000 Hz.

Fig. 7(a) and (b) summarizes the results obtained encoding speech with RASTA-PLP coefficients extracted from the frequency range 0–8000 Hz.

Fig. 7(a) plots accuracy while varying the number of Gaussian components and the number of RASTA-PLP coefficients used to encode the signal. The total length of the training material is approximately 25 s per speakers, the number of speakers is 16, while utterances of 10 s are used for testing.

Fig. 7(b) shows accuracy while varying the total length of training material per speaker and the number of RASTA-PLP coefficients used to encode the signal. The number of speakers is 16, while utterances of 10 s are used for testing.

Fig. 7(a) and (b) indicates that RASTA-PLP coefficients are not as effective in capturing speaker identity as MFCCs and *pykfec setup 3*, within the context of our relatively simple classifier. While the GMM recognition rate using the RASTA-PLP parametrization reaches about 80% when increasing both training material and number of Gaussian components, MFCCs and *pykfec setup 3* yield accuracy scores of about 90% in similar experimental setups—see Fig. 5(a) and (b) and Fig. 6(a) and (b).

In order to facilitate a comparison of the results presented in Fig. 5(a)–(d), Fig. 6(a)–(c), and Figs. 7(a) and (b), we reproduce in Table II and Fig. 8 some of the accuracy results obtained in

TABLE II
ACCURACY OF DIFFERENT ENCODINGS VARYING THE TRAINING AND TEST MATERIAL LENGTH AND NUMBER OF PARAMETERS. THE REFERENCE CLASSIFIER IS TRAINED WITH 64 GAUSSIAN COMPONENTS. RESULTS ARE OBTAINED ON THE "16-SPEAKER" DATASET. THE ACCURACY SCORES ARE COMPUTED AS THE MEAN OF TEN INDEPENDENT RUNS, EACH RUN HAVING A DIFFERENT RANDOM INITIALIZATION OF THE GMMS. THE ERROR OF THE SCORE IS CALCULATED AS TWICE THE STANDARD DEVIATION OF THE MEAN

Training Length [*sec.*]	Test Length [*sec.*]	Number of Features	*pykfec setup 1* ACC. [%]	Error [%]
25	5	20	83	5
25	5	40	87	6
25	5	80	87	5
25	10	80	92	3
60	10	80	93	1
Training Length [*sec.*]	**Test Length [*sec.*]**	**Number of Features**	**MFCC ACC. [%]**	**Error [%]**
25	10	15	86	3
25	10	20	86	4
25	10	25	88	4
60	10	25	94	1
25	10	35	88	3
60	10	35	88	1
Training Length [*sec.*]	**Test Length [*sec.*]**	**Number of Features**	**RASTA-PLPC ACC. [%]**	**Error [%]**
25	10	12	54	6
25	10	16	65	4
25	10	20	76	9
60	10	20	82	6

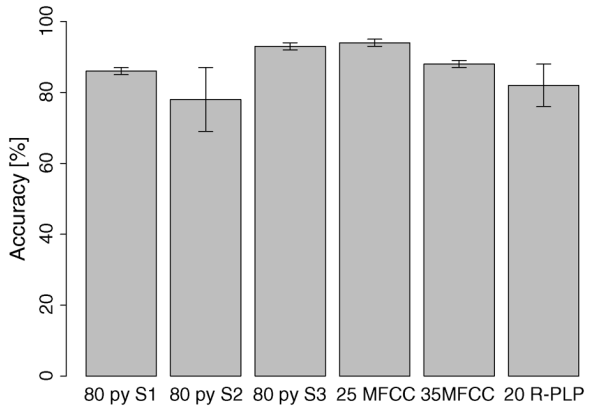

Fig. 8. Accuracy of the reference GMM classifier using *pykfec setup 1*, *setup 2*, and *setup 3* compared with the accuracy of the same classifier using 25 MFCC, 35 MFCC, and 20 RASTA-PLP for speaker parametrization. The reference classifier is trained with 64 Gaussian components; results are obtained on the "16-speaker" dataset and about 60 s of training material per speaker is used; 10-s utterances are used for testing. The different parameters are extracted from the same frequency interval 0–8000 Hz.

the different experimental setups. "Training Length" indicates the amount of training material per speaker used. In Table II, "Test Length" gives the length of the utterance used for testing; "Number of Features" indicates the number of parameters used to encode the speech signal.

C. Role of Different Frequency Intervals

In this section, we present results obtained by restricting our parameter extraction to two limited frequency ranges, 0–4000 Hz and 4000–8000 Hz. The motivation for this is to investigate the potential role in speaker identification for the structures above 4 kHz that are apparent in the pyknogram, but less obvious from a spectrographic representation. In general,

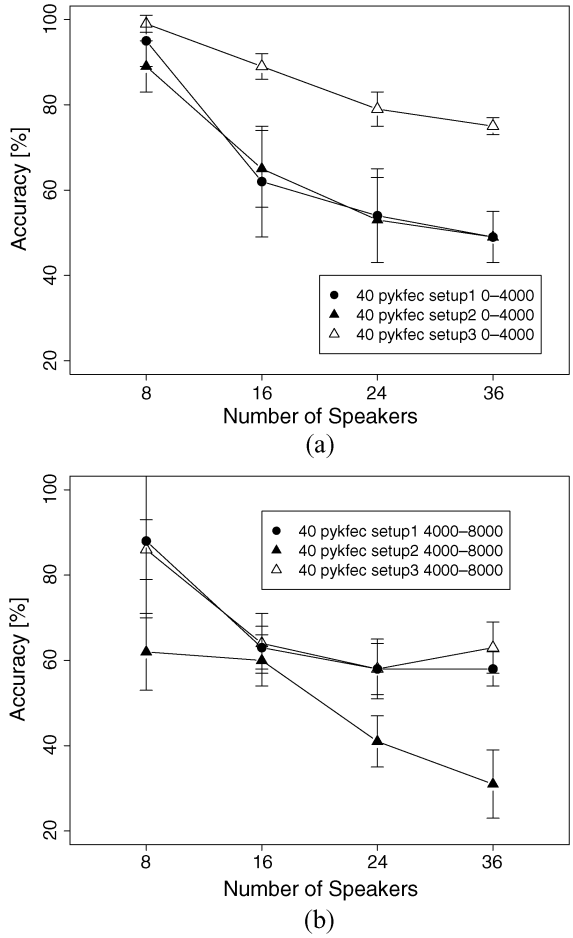

Fig. 9. Accuracy of the reference GMM classifier using *pykfec* for speaker parametrization, varying the number of speakers considered and the frequency interval from which the parameters are extracted (0–4000 Hz and 4000–8000 Hz).

little is known about the relevance or origin of such structures, or how they might vary from individual to individual. The training material is extracted from the first recording session of the CHAINS corpus and uses NORM speech; the test material is extracted from the second recording session and spoken in the FAST style. The reference classifier is trained with 64 Gaussian components and utterances of 10 s are used for testing. The results obtained are also compared with MFCCs extracted from the same two frequency ranges (Fig. 10).

Fig. 9(a) shows that encoding speech with 40 *pykfec setup 3* features extracted from the interval 0–4000 Hz provides better identification performance than the use of either *pykfec setup 1* or *pykfec setup 2*, with accuracy curves that are comparable to those obtained using 80 *pykfec setup 3* features extracted from the larger frequency range of 0–8000 Hz [e.g., considering the *24-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, *pykfec setup 3* 0–4000 Hz (40 features) provide an accuracy of 79 ± 4 while *pykfec setup 3* 0–8000 Hz (80 features) provide an accuracy of 80 ± 6, as shown in Fig. 9(a) and Fig. 5(c)]. On the other hand, Fig. 9(b) clearly shows that parameters extracted from the higher frequency range 4000–8000 Hz are less selective for speaker identity than

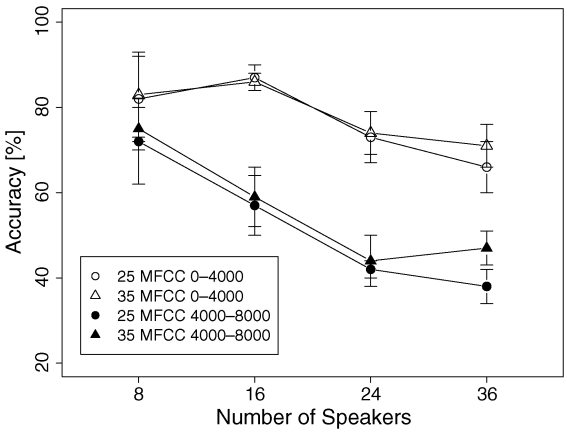

Fig. 10. Accuracy of the reference GMM classifier using MFCC for speaker parametrization, varying the number of speakers considered, the number of coefficients and the frequency interval from which the parameters are extracted (0–4000 Hz and 4000–8000 Hz).

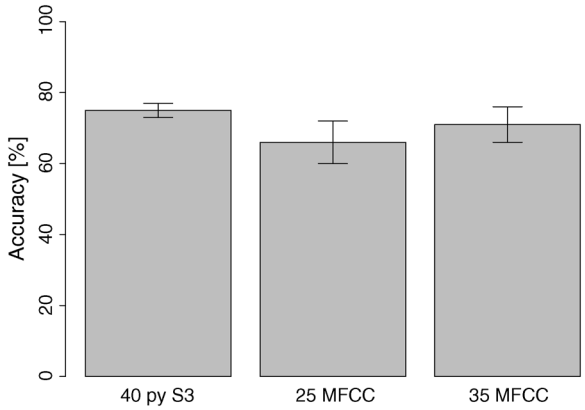

Fig. 11. Accuracy of the reference GMM classifier using 40 *pykfec setup 3* compared with the accuracy of the same classifier using 25 and 35 MFCCs for speaker parametrization. The reference classifier is trained with 64 Gaussian components, results are obtained on the "36-speaker" dataset and about 25 s of training material per speaker is used; 10-s utterances are used for testing. The different parameters are extracted from the same frequency interval 0–4000 Hz.

those from 0 to 4000 Hz, e.g., considering the *24-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, *pykfec setup 3* 4000–8000 Hz (40 features) provide an accuracy of 58 $\pm$ 7; *pykfec setup 3* 0–4000 Hz (40 features) provide an accuracy of 79 $\pm$ 4 [Fig. 9(a)].

It is interesting to note that *pykfec setup 1* and *pykfec setup 2* provide equivalent accuracy in the frequency range 0–4000 Hz, while in the frequency range of 4000–8000 Hz equivalent accuracy is obtained by adopting *pykfec setup 1* and *pykfec setup 3*. This corresponds well with the impression of similarity seen in comparing the three setups in Fig. 1, Fig. 2, and Fig. 3. The pyknogram obtained adopting *setup 1* and *setup 2* appear to represent formant-like information below 4000 Hz similarly, while *setup 3* draws out individual harmonics more. Conversely, *setup 1* and *setup 3* represent broad resonances above 4 kHz in similar fashion, while these get washed out with the broad bandwidths of *setup 2*.

Fig. 10 shows the results obtained using 25 and 35 MFCCs to encode the same speech signals as above, varying the number of speaker and the frequency intervals from which MFCCs are extracted. The reference classifier is trained with 64 Gaussian components and utterances of 10 s are used for testing.

Fig. 10 shows that 25 and 35 MFCCs provide somewhat equivalent identification rates when extracted from the same frequency intervals (e.g., considering the *36-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, 25 MFCC 0–4000 Hz provide an accuracy of 66 $\pm$ 6 while 35 MFCC 0–4000 Hz provide an accuracy of 71 $\pm$ 5). Moreover, it confirms that parameters extracted from the frequency range 4000–8000 Hz are not good indicators of speaker identity, e.g., considering the *36-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, 35 MFCCs 0–4000 Hz provide an accuracy of 71 $\pm$ 5, while 35 MFCCs 4000–8000 Hz provide an accuracy of 47 $\pm$ 4. It is interesting to note that the recognition rate obtained using 40 *pykfec setup 3* features is roughly equivalent to the performance obtained with both 25 and 35 MFCCs, irrespective of the frequency range from which the parameters are extracted, e.g., considering the *36-speaker* set, training 64 GMMs per speaker, using 10-s test utterances and approximately 25 s of training material per speaker, *pykfec setup 3* 0–4000 Hz (40 features) provides an accuracy of 79 $\pm$ 4 [Fig. 9(a)] while 35 MFCCs 0–4000 Hz provide an accuracy of 71 $\pm$ 5 (Fig. 10).

Fig. 11 summarizes some of the results presented in this section.

D. Different Phonation: Whispering

In this section, nonmodal, whispered, speech is used to test the GMM reference classifier. This is done to provide a relatively hard test case for the identification system and to verify the extent to which it is possible to generalize an identification model (the mixture of Gaussian models trained for each speaker) based on modally voiced speech to whispered speech. The training material is NORM speech from the first recording session of the CHAINS corpus; the test material is from the second recording session in which speech was whispered throughout. The reference classifier is trained with 64 Gaussian components and utterances of 10 s are used for testing. The results reported in this section are obtained using both *pykfec* and MFCCs for speech parametrization (Fig. 12).

Fig. 12 shows that neither MFCCs nor *pykfec* provides a speech parametrization that is invariant with respect to drastic changes in phonation: the identity model extracted from modal speech (the NORM style) is not useful when test samples are whispered. On the other hand, when the same models are tested using FAST speech, the reference system is well capable of recognizing speakers' identities—Fig. 5(a)–(d) and Fig. 6(a)–(c).

Finally, Fig. 13 shows the effect of extracting MFCCs and *pykfec* from two different frequency ranges: 0–4000 Hz and 4000–8000 Hz using the same experimental conditions as above. The training and test materials are as before. The reference classifier is trained with 64 Gaussian components and utterances of 10 s are used for testing.

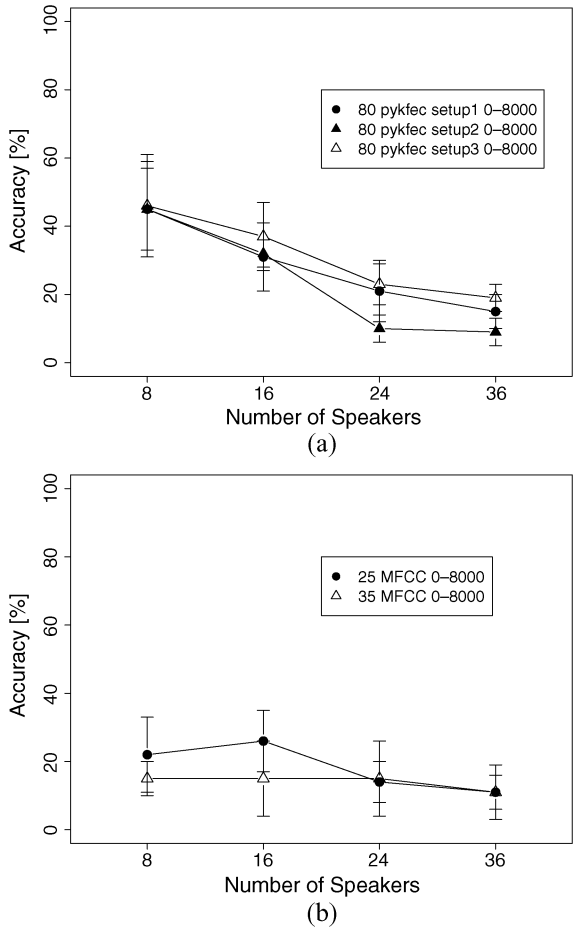

Fig. 12. Accuracy of the reference GMM classifier trained on SOLO speech and tested using WHSP speech varying the number of speakers considered and the parametrization of the speech samples. Parameters are extracted from the frequency interval 0–8000 Hz.

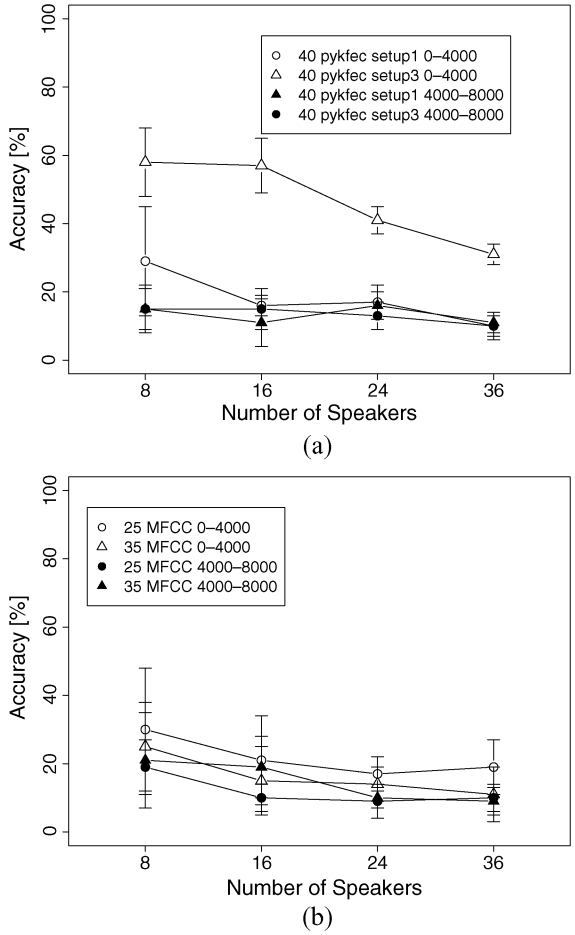

Fig. 13. Accuracy of the reference GMM classifier trained on SOLO speech and tested using WHSP speech varying the number of speakers considered and the parametrization of the speech samples. Parameters are extracted from the frequency intervals 0–4000 Hz and 4000–8000 Hz.

Fig. 13(a) and (b) shows that 40 *pykfec setup 3* extracted from the frequency range 0–4000 Hz provides a better recognition rate than any other encoding tested. Nevertheless, the accuracy of the reference system is far from ideal, confirming that the parametrizations explored are not invariant with respect to drastic changes in phonation. For the purpose of comparison with Fig. 1, Fig. 14 shows a sample pyknogram of "If it doesn't matter who wins, why do we keep score?", derived from whispered speech. Finally, Fig. 15 summarizes some of the results presented in this section.

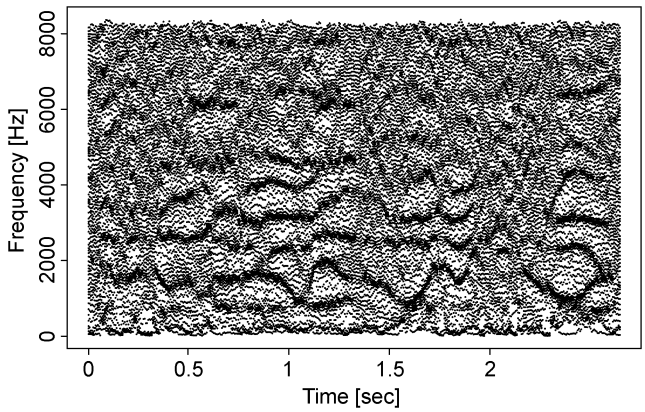

Fig. 14. Pyknogram of "If it doesn't matter who wins, why do we keep score?" (whisper). Eighty filters linearly spaced between 200 and 8200 Hz, constant bandwidth of 400 Hz—Filterbank *setup 1*.

VII. Conclusion

This work has introduced a new set of descriptors that capture the identity of speakers well and that demonstrate robustness with respect to changes in recording channel and speaking style, without requiring any advanced channel compensation schema, as it is usually the case for standard encoding approaches such as MFCCs. Our experimental evaluation indicates that the characterization of the different instantaneous frequencies within the speech signal play a significant role in capturing the identity of a speaker. The identity of a human speaker can be exploited robustly by looking at *which* instantaneous frequencies are produced by the speech production system.

The great plasticity and complexity of the human speech production system ensure that the problem of speech parametrization is not uniquely solvable. For this reason, this work explores

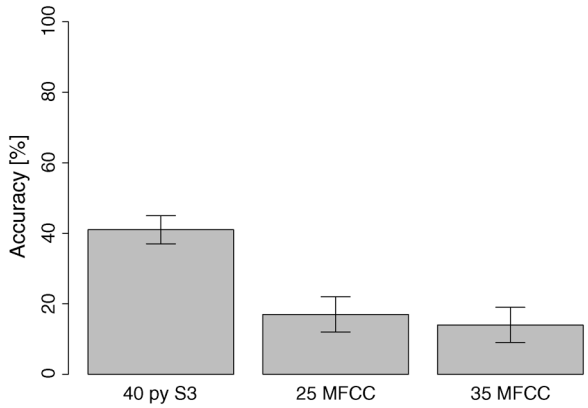

Fig. 15. Accuracy of the reference GMM classifier using 40 *pykfec setup 3* compared with the accuracy of the same classifier using 25 and 35 MFCCs for speaker parametrization. The reference classifier is trained with 64 Gaussian components, results are obtained on the "24-speaker" dataset and about 25 s of modal speech (NORM) per speaker is used; 10-s ("whisper") utterances are used for testing. The different parameters are extracted from the same frequency interval 0–4000 Hz.

three different way to decompose the speech signal within the same framework. Three different setups were tested for computing the pyknogram of the signal and the derived *pykfec* parameters; three setups that differ in their specificity for the various harmonics and resonances typical of the speech signal.

When compared with *pykfec setup 1* and *pykfec setup 2*, *pykfec setup 3* show the best performance in terms of speaker identification and stability, as we increase the number of Gaussian components in the reference classifier, the amount of training material and the number of features. With respect to the latter, despite the introduction of clear redundancy among the descriptors, the reference GMM classifier shows no loss in recognition rate as we increase the number of parameters from 40 up to 100. Moreover, with the proposed AM–FM approach, channel normalization is not required, as the (instantaneous) amplitude is used only for identifying the short time frequency estimate within a single band.

In the experiments in which we selectively extracted parameters from restricted frequency ranges, we found that although the *pykfec* features appear to reveal interesting structures above 4 kHz, these are not necessarily of use in speaker identification, and our novel features did not fare significantly better than the MFCC representation.

In the first set of experiments, we demonstrated that the novel representations are capable of generalizing well from high-quality speech recorded at a comfortable rate to lower-quality recordings in which the speaker tried to speak quickly. This represents a promising degree of robustness with respect to both channel characteristics and style. The robustness has its limits, however, as was demonstrated when we used whispered speech as test material. Even here though, it was clear that the new *pykfec* may be of use in the future, as results obtained with setup 3 when features were extracted below 4 kHz were demonstrably better than either of the other two setups, or either of the MFCC parameterizations employed.

Our examination of the robustness of these novel representations are necessarily limited in scope. However, the robust performance across the board exhibited by the novel AM–FM derived features are promising and merit the attention of the speaker identification and verification community for consideration in further work.

ACKNOWLEDGMENT

The authors would like to thank the SFI/HEA Irish Center for High-End Computing (ICHEC) for the provision of computational facilities and support. They would also like to thank the anonymous reviewers for their very constructive feedback.

REFERENCES

[1] T. B. Alderman, *Forensic Speaker Identification, A Likelihood Ratio-Based Approach Using Vowel Formants*, ser. Lincom Studies in Phonetics. Munich, Germany: LINCOM, 2005.

[2] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, 1971.

[3] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 430–451, 2004.

[4] B. Boashash, "Estimating and interpreting the instanteneous frequency of a signal—Part 1: Fundamentals," *Proc. IEEE*, vol. 80, no. 4, pp. 519–538, Apr. 1992.

[5] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, no. 9/10, pp. 341–345, 2001.

[6] J. F. Bonastre, F. Bimbot, L. J. Boe, J. P. Campbell, D. A. Reynolds, and I. Magrin-Chagnolleau, "Person authentication by voice: A need for caution," in *Proc. Eurospeech 2003*, Genoa, Italy, Sep. 2003, pp. 33–36.

[7] J. P. Campbell, Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[8] K. Chen, L. Wang, and H. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification," *Int. J. Pattern Recognition Artif. Intell.*, vol. 11, no. 3, pp. 417–445, 1997.

[9] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in *Proc. SPECOM'06*, St. Petersburg, Russia, 2006, pp. 431–435.

[10] D. Dimitriadis and P. Maragos, "Robust energy demodulation based on continuous models with application to speech recognition," in *Proc. Eurospeech'03*, 2003, pp. 2853–2856.

[11] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Process. Lett.*, vol. 12, no. 9, pp. 621–624, Sep. 2005.

[12] M. Foundez-Zanuy, S. McLaughlin, A. Esposito, A. Hussain, J. Schoentgen, G. Kubin, W. B. Kleijn, and P. Maragos, "Nonlinear speech processing: Overview and applications," *Control Intell. Syst.*, vol. 30, pp. 1–10, 2002.

[13] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[14] D. Gabor, "Theory of communication," *JIEE*, vol. 93, no. 3, pp. 429–457, Nov. 1946.

[15] J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification," in *Proc. ESCA Workshop Automatic Speaker Recognition, Identification, Verification*, Martigny, Switzerland, Apr. 1994, pp. 39–42.

[16] H. Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *J. Acoust. Soc. Amer.*, pp. 1738–1752, 1990.

[17] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoust. , Speech, Signal Process. (ICASSP'92)*, Mar. 23–26, 1992, vol. 1, pp. 121–124.

[18] H. Hollien, *Forensic Voice Indentification*. New York: Academic, 2002.

[19] C. R. Jankowski, Jr, T. F. Quatieri, and D. A. Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP'95)*, 1995, pp. 325–328.

[20] J. K. Keiser, "On Teager's energy algorithm and its generalization to continuos signals," in *Proc. IEEE DSP Workshop*, New York, 1990, CD-ROM.
[21] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, pp. 1253–1257, 1962.
[22] G. Li, L. Qiu, and L. K. Ng, "Signal representation based on instantaneous amplitude models with application to speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 353–357, May 2000.
[23] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
[24] Y. Mami and D. Charlet, "Speaker recognition by location in the space of reference speakers," *Speech Commun.*, vol. 48, no. 2, pp. 127–141, 2006.
[25] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct. 1993.
[26] G. McLachlan and K. E. Basford, *Mixture Models*. New York: Marcel Dekker, 1987.
[27] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 554–568, Sep. 1999.
[28] K. K. Paliwal, "Usefulness of phase in speech processing," in *Proc. IPSJ Spoken Lang. Process. Workshop*, Gifu, Japan, 2003, pp. 1–6.
[29] K. K. Paliwal and B. S. Atal, "Frequency-related representation of speech," in *Proc. Eurospeech'03*, Sep. 2003, pp. 65–68.
[30] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoust. Soc. Amer.*, vol. 99, pp. 3795–3806, 1996.
[31] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 196–200, Mar. 2001.
[32] L. R. Rabiner and R. W. Shafer, *Digital Signal Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1989.
[33] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.
[34] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 639–643, Oct. 1994.
[35] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, pp. IV-4072–IV-4075.
[36] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.
[37] R. D. Rodman, "Computer recognition of speakers who disguise their voice," in *Proc ICSPAT'00*, 2000, CD-ROM.
[38] P. Rose, *Forensic Speaker Indentification*, ser. Forensic Science. New York: Taylor and Francis, 2002.
[39] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*, ser. NATO Advanced Study Institute Series D, W. J. Hardcastle and A. Marchal, Eds. Bonas, France: Kluwer, Jul. 1989, vol. 55.
[40] V. Wan and S. Renals, "Evaluation of kernel methods for speaker verification and identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'02)*, 2002, vol. 1, pp. 669–672.
[41] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 203–210, Mar. 2005.
[42] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 447–456, Sep. 2003.

Marco Grimaldi received the Laurea (M.S.) degree in physics from the University of Trento, Trento, Italy, in 1999 and the Ph.D. degree in computer science from Trinity College Dublin, Dublin, U.K., in 2004.

From 2004 to 2005, he was a Research Assistant in the Department of Computer Science, Trinity College Dublin. In the Fall of 2005, he joined the UCD School of Computer Science and Informatics, University College Dublin, as a Research Fellow, conducting full-time research on speaker identification. His research interests include robust speaker identification and verification, speaker recognition, and general problems in signal analysis and classification.

Fred Cummins received the B.A. degree in computer science, linguistics, and German from Trinity College Dublin, Dublin, U.K., in 1991, the M.A. degree in linguistics from Indiana University, Bloomington, in 1996, and the Ph.D. degree in linguistics and cognitive science from Indiana University in 1997.

He completed postdoctoral research positions at Northwestern University, Evanston, IL, and in the Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland. He has been a College Lecturer in University College Dublin (UCD) School of Computer Science and Informatics since 1999. He has published over 60 refereed publications, mainly in the area of phonetics and cognitive science. His current interests encompass enactive and dynamical models of cognition, visual perception, speaker identification, and coordination.