ABSTRACT

Recorded speech signals convey information not only for the speakers’ identity and the spoken language, but also for the acquisition devices used for their recording. Therefore, it is reasonable to perform acquisition device identification by analyzing the recorded speech signal. To this end, recording-level spectral, cepstral, and fusion of spectral and cepstral features are employed as suitable representations for device identification. The feature vectors extracted from the training speech recordings are used to form overcomplete dictionaries for the devices. Each test feature vector is represented as a linear combination of all the dictionary columns (i.e., atoms). Since the dimensionality of the feature vectors is much smaller than the number of training speech recordings, there are infinitely many representations of each test feature vector with respect to the dictionary. These representations are referred to as collaborative representations in the sense that all the dictionary atoms collaboratively represent any test feature vector. By imposing the representation to be either sparse (i.e., to admit the minimum $\ell_1$ norm) or to have the minimum $\ell_2$ norm, unique collaborative representations are obtained. The classification is performed by assigning each test feature vector the device identity of the dictionary atoms yielding the minimum reconstruction error. This classification method is referred to as the sparse representation-based classifier (SRC) if the sparse collaborative representation is employed and as the least squares collaborative representation-based classifier (LSCRC) in the case of the minimum $\ell_2$ norm regularized collaborative representation is used for reconstructing the test sample. By employing the LSCRC, state of the art identification accuracy of 97.67% is obtained on a set of 8 telephone handsets, from Lincoln-Labs Handset Database.

Keywords: Audio Forensics, Collaborative Representation, Digital Speech Forensics, Sparse Representation, Telephone Handset Identification

1. INTRODUCTION

Speech is the most natural way to communicate between humans. Nowadays, speech communication systems acquire, transmit, store, and process the information in digital form. However, the digital speech content can be imperceptibly altered by malicious, even amateur, users by using a variety of low-cost audio editing software. This creates a serious threat to the knowledge life cycle. Indeed, when hearing is no longer believing, the process of
going from data to information, knowledge, understanding and, decision making is severely compromised (Farid, 2008). The consequences of this threat permeate a wide variety of fields, such as intellectual property, intelligence gathering, forensics, and news reporting to name a few. Currently, the methods to combat this threat in the field of digital speech forensics are still in their infancy. Therefore, there is an urgent need to advance the state-of-the-art in this field (Garcia-Romero & Espy-Wilson, 2010).

A first step to remedy the aforementioned threat is to extract forensic evidence about the mechanism involved in the generation of the speech recording by analyzing only the speech signal (Garcia-Romero & Espy-Wilson, 2010). That is, to identify the acquisition device by assuming that the devices along with their associated signal processing chain leave behind intrinsic traces in the speech signal. Indeed, the electronic devices, especially when including a microphone, cannot have exactly the same frequency response due to tolerances in the production of their electronic components and the different designs employed by the various manufacturers (Hanilci, 2012). This implies that the recorded speech can be considered as a signal whose spectrum is the product of the genuine speech spectrum, driving the acquisition device, and the frequency response of the latter. Consequently, the recorded speech signal can be exploited in device identification, following a blind-passive approach, as opposed to active embedding of watermarks or having access to input-output pairs (Garcia-Romero & Espy-Wilson, 2010).

Although there are significant advances in image forensics (Farid, 2008), audio forensics are less developed (Maher, 2009). Few exceptions include the authentication of MP3 (Yang, 2008) and the authentication of speakers’ environment (Oermann, 2005; Kraetzer, 2007; Malik & Farid, 2010). Similarly, a few automatic acquisition device identification systems have been developed. For instance, a method for the classification of 4 microphones has been proposed in (Kraetzer, 2007). The speech signal is parameterized by employing time domain features and the mel-frequency cepstral coefficients (MFCCs). The identification of the microphones is performed by a Naive Bayes classifier at a short-time frame level. Accuracies on the order of 60-75% have been reported. In (Garcia-Romero & Espy-Wilson, 2010), the identification of 8 landline telephone handsets and 8 microphones is addressed. In particular, the intrinsic characteristics of the device are captured by a template constructed by appending together the means of a Gaussian mixture which have been trained using linear and mel-scaled cepstral coefficients extracted by speech recordings of each device. Classification accuracies higher than 90% have been achieved, when a support vector machine (SVM) classifier was used. Recently, a robust system for the identification of cell-phones has been proposed in (Hanilci, 2012). In particular, when the MFCCs, extracted from speech recordings, are classified by an SVM, 14 different cell-phones are identified with an accuracy of 96.42%.

In this paper, a novel blind-passive method for landline telephone handset identification is proposed. The method resorts on spectral and cepstral based features extracted from speech recordings and their collaborative representation (Zhang, 2011), revealing the identity of the recording device. Figure 1 outlines the proposed method. In particular, 5 recording-level features are used for device characterization. The random spectral features (RSFs) (Panagakis & Kotropoulos, 2012a) and the labeled spectral features (LSFs) (Panagakis & Kotropoulos, 2012b) are obtained by applying unsupervised and supervised feature selection on the mean spectrogram of each speech recording, respectively. That is, for unsupervised feature selection, the dimensionality of the mean spectrogram is reduced by random projections (Bingham & Mannila, 2001) yielding the RSFs (Panagakis & Kotropoulos, 2012a). In the supervised setting, the label information (i.e., the class where each device belongs to) of the training speech recordings is taken into account in order to derive a mapping between the feature space where the mean spectrograms lie onto and the label space. The mapping between