Digital Signal Processing 35 (2014) 75–85

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Source cell-phone recognition from recorded speech using non-speech segments

Cemal Hanilçi [a],*, Tomi Kinnunen [b]

[a] Department of Electrical-Electronic Engineering, Bursa Technical University, 16190, Bursa, Turkey
[b] University of Eastern Finland, FI-80101, Joensuu, Finland

ARTICLE INFO

Article history:
Available online 27 August 2014

Keywords:
Source cell-phone recognition
Mel-frequency cepstrum coefficients
Mutual information
Source microphone identification
Gaussian mixture model

ABSTRACT

In a recent study, we have introduced the problem of *identifying cell-phones using recorded speech* and shown that speech signals convey information about the source device, making it possible to identify the source with some accuracy. In this paper, we consider recognizing source cell-phone microphones using non-speech segments of recorded speech. Taking an information-theoretic approach, we use Gaussian Mixture Model (GMM) trained with maximum mutual information (MMI) to represent device-specific features. Experimental results using Mel-frequency and linear frequency cepstral coefficients (MFCC and LFCC) show that features extracted from the non-speech segments of speech contain higher mutual information and yield higher recognition rates than those from speech portions or the whole utterance. Identification rate improves from 96.42% to 98.39% and equal error rate (EER) reduces from 1.20% to 0.47% when non-speech parts are used to extract features. Recognition results are provided with classical GMM trained both with maximum likelihood (ML) and maximum mutual information (MMI) criteria, as well as support vector machines (SVMs). Identification under additive noise case is also considered and it is shown that identification rates reduces dramatically in case of additive noise.

1. Introduction

Development of digital technology has led to development of low-cost portable tools such as pocket cameras, dictation machines, cellular phones and smart phones that form an integral part of our daily life. Such tools are used for recording and transmitting multimedia data that play an increasing role as an evidence in forensic investigation [1]. Thus, there is an increasing need for accurate analysis and classification of forensic multimedia data.

Forensic multimedia analysis has largely focused on digital images. Determining the integrity and authenticity of an image, identifying the source camera the image was taken with [2], digital image and video watermarking [3] and image *steganalysis* [4] – the problem of detecting the presence of hidden messages in images – are common applications. From these, *source camera recognition* [2] is one of the most challenging problems. Given an image, the task is to determine the source device that the image was taken with [5]. This is possible because imperfections in acquisition devices leave their device-specific footprints to the images. For example, in [6] and [7], identification of digital cameras based on sensor pattern noise and dust was proposed. In [8], a similar approach was used for identifying source scanners from scanned images based on dust, dirt and scratch traces on the scanner platen.

When a recorded speech sample appears as forensic evidence, it is often necessary to trace the recording device or the environment. To this end, identification of the microphone and recording environment have been addressed in several studies [9–15]. For example, [10] studied classification of 4 different microphones and 10 different environments (rooms) using different time-domain features while [11] used Fourier coefficients to classify 7 different microphones. The authors of [13] studied identification of 10 telephone handsets using Mel- and linear-frequency cepstral coefficients (MFCCs and LFCCs), reporting higher than 90% classification accuracy. Similarly, [14] used MFCCs to identify microphones in NIST 2008 speaker recognition evaluation (SRE) corpus. More recently, [15] studied identification of 8 different landline telephone handsets (four carbon-button and four electret) and 8 different microphones using cepstral features while [12] used higher-order statistic features for microphone identification.

In our recent study [16], we addressed a similar problem to handset identification, *source cell-phone identification* [16]. As typical in speech-related classification tasks, we used features extracted from the whole signal. Even though high recognition

* Corresponding author.
E-mail addresses: cemal.hanilci@btu.edu.tr (C. Hanilçi), tomi.kinnunen@uef.fi (T. Kinnunen).

http://dx.doi.org/10.1016/j.dsp.2014.08.008
1051-2004/© 2014 Elsevier Inc. All rights reserved.

accuracy was achieved, for the present work we hypothesize that information about the source device might be more pronounced in the *non-speech* parts of the signal. Therefore, our study gives the first detailed account into whether speech or nonspeech parts convey more device-specific information. We attack the problem both with classification experiments utilizing Gaussian mixture model (GMM) and support vector machine (SVM) classifiers, as well as studying device-specific information in different acoustic features with the aid of a mutual information criterion. Inspired by the use of sensor noise to detect source cameras [6] and establishment of such methodology as state-of-the-art in image forensics, we are curious to study the relative importance of non-speech segments in cell-phone identification. We approach the problem from an information-theoretic perspective, specifically, using maximum mutual information (MMI) criterion to analyze the amount of device-specific information in speech and non-speech parts. We show that features extracted from non-speech parts of the signal contain higher mutual information compared to those extracted from the speech segments. This naturally somewhat avoids the irrelevant information (disturbance) and thus yields higher recognition rates. This result can be justified from signals and systems point of view as well. As the non-speech parts contain only noise-like signals, which have a flatter spectral density compared to those of speech signals and provide relatively uncorrelated excitation to the recorder, they capture the transfer function of the recording circuitry of source devices (the device footprint) much closer to its original. Thus, noise-like signals (non-speech segments) help us to discriminate source devices easier than the speech segments because they transfer the electro-acoustical properties of the recording device to the recorded signal. For comparison, we examine the performance of classical GMM trained with maximum likelihood (ML) criterion as well as the state-of-the-art pattern classification method, support vector machines (SVM), and provide experimental results.

While the classifiers selected for this study are well-explored in speaker recognition [17], it is unknown of how they apply to source cell-phone recognition. Besides comparison of classifiers, in this paper, we investigate the source device recognition performance under additive noise conditions which has not been considered in the previous studies. Mel-frequency and linear frequency cepstral coefficient (MFCC and LFCC) feature representations are also compared for both clean and noisy conditions. Our purpose in this paper is to compare the performance of established methods on the source device identification using the proposed feature extraction technique.

2. Source cell-phone recognition system

Source cell-phone recognition can refer to two different tasks: *identification* and *detection*. Both tasks consist of two steps: *training* and *recognition*. In the training step, features are extracted from the training speech samples of each cell-phone in the database and a cell-phone model is created. In the recognition step of an identification system (Fig. 1(a)), features are extracted from a test signal and a similarity score is computed for each of the cell-phone models in the database. The cell phone that gives the largest similarity score is designated as the detected cell phone. In detection (Fig. 1(b)), a similarity score between the features extracted from test speech and hypothesized cell-phone model is computed; if the score is above the threshold the hypothesis is accepted and rejected otherwise. Note that our goal is to recognize a specific physical cell-phone which exist in our database, rather than recognizing the brand or the model. Thus, we use the term *source cell-phone recognition*.

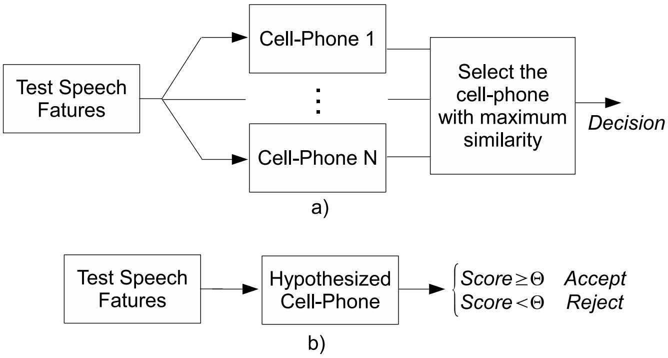

Fig. 1. Decision logic for cell-phone (a) identification and (b) detection systems.

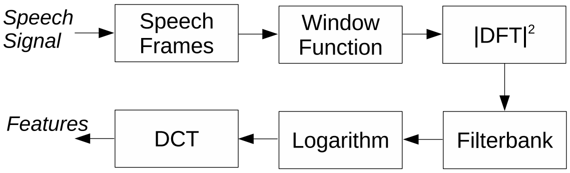

Fig. 2. MFCC extraction process.

2.1. Feature extraction

Most of the feature sets in speech processing are extracted from the spectrum of the signal. The most popular features for the recognition systems are the mel-frequency cepstral coefficients [17]. In [16], we considered the recording device (cell-phone) leaving its foot-prints in the recorded speech as device-specific information in the form of a convolutional distortion. This information can then be captured and represented by MFCC feature vectors in additive form with the contributions of speech signal and source device. Since the information from the speech signal is itself irrelevant to the aim of the task, in this paper, we use the non-speech parts of the recorded signal to remove less relevant information.

Considered to be stationary in short-term, the speech signal is first divided into overlapping frames and windowed using an appropriate window function. The power spectrum is computed using the discrete Fourier transform (DFT), which is then smoothed with a bank of triangular filters whose center frequencies are uniformly spaced on the Mel-scale. Finally, logarithmic filterbank outputs are converted into MFCCs by taking the discrete cosine transform (DCT). The MFCC extraction procedure is shown in Fig. 2. In the paper, we use 30 millisecond frames with 15 millisecond overlap and a Hamming window.

Different from the MFCC features, *linear* frequency cepstral coefficients (LFCCs) are extracted the same way, except that the triangular filters are spaced in linear, rather than Mel-scale. We compare the source identification performances of these two different acoustic feature sets. Detailed comparison of different feature extraction and normalization methods on source cell-phone identification can be found in [18].

2.2. Speech/nonspeech detection

We use adaptive energy-based speech activity detector (SAD) to locate the speech and the non-speech parts. Our energy SAD measures the energy of each frame and compares it with a threshold. The energy of a speech frame is measured as,

$$E_i = \left(\frac{1}{S}\sum_{k=1}^{S}\left(s_i(k) - \bar{s_i}\right)^2\right)^{1/2}, \tag{1}$$

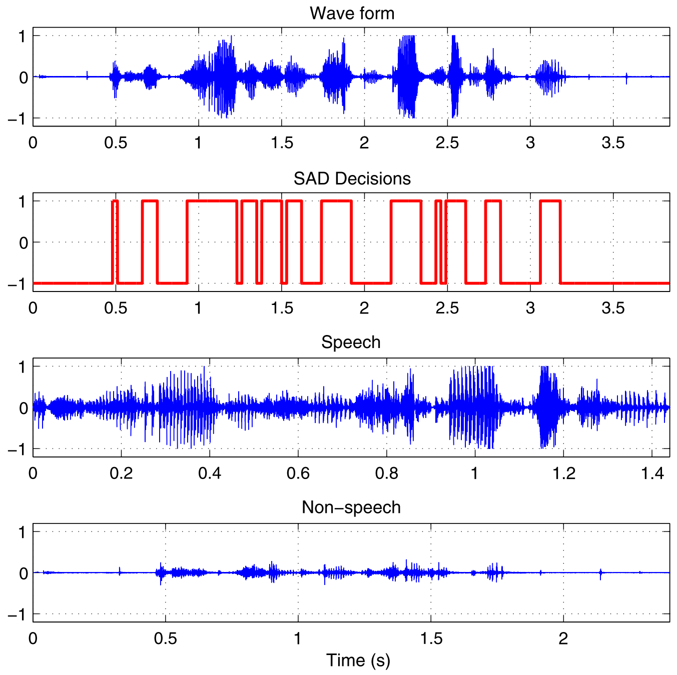

Fig. 3. Example of speech/non-speech detection using SAD (SAD decision with the label +1 corresponds to speech segment and −1 corresponds to non-speech segment. $\alpha = 0.125$).

where $s_i(k)$ is the kth sample of ith speech frame and $\bar{s}_i$ is the sample mean of ith frame and S is the total number of samples in a frame. The threshold is then calculated as:

$$\Lambda = E_{\min} + \alpha(E_{\max} - E_{\min}). \quad (2)$$

Here, $E_{\min}$ and $E_{\max}$ are the minimum and maximum frame energy values over all frames and $0 \le \alpha \le 1$ is a constant. The ith frame is deemed as speech if $E_i \ge \Lambda$ and non-speech otherwise. Note that when $\alpha = 0$ no SAD is used. Fig. 3 shows an example of a speech signal, SAD decision labels and the signal after the non-speech parts are removed. It can be seen that around 65% of this speech signal is determined as non-speech so the length of the speech signal reduced from 4 seconds to 1.4 seconds after SAD.

In this study, we use the energy SAD method for the simulations for reasons of simplicity. In practical forensic applications this would be usually replaced by manual segmentation carried out by the forensic analyst.

2.3. Mutual information in MFCCs

The use of information theory, mutual information concept in particular, has been studied in several speaker recognition studies [19–22], where the mutual information is used in selecting features for speaker recognition with less redundancy. Features which are capable of holding higher mutual information between the input and the speaker models yield higher recognition rate. In this paper, the mutual information is similarly used to select the most discriminating features with the most appropriate size, and hence decide on what part of the recorded signal and what number of features to use to represent the device-specific information most adequately.

The mutual information between two random variables λ and C, representing the statistical model of a cell-phone and the feature vector extracted from recorded speech signals, respectively, is defined as [23]

$$I(\lambda, C) = h(\lambda) - h(\lambda|C) = h(C) - h(C|\lambda) \quad (3)$$

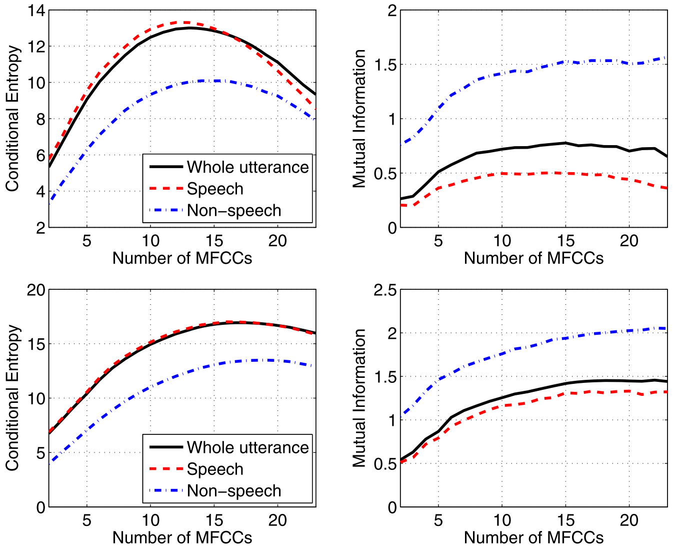

Fig. 4. An estimate of the conditional entropy $h(C|\lambda)$ and the mutual information $I(C, \lambda)$ in (3) as a function of the number of features employing the TIMIT (first row) and the LIVE RECORDS (second row) databases.

where $h(\lambda)$ and $h(\lambda|C)$ are the entropy of λ and the conditional entropy of λ given C, respectively. Suppose that $\{\mathbf{c}_t^j\}_{t=1}^T$, $j = 1, \ldots, N$, where N is the number of cell-phone devices in the database, are the set of features extracted from the speech samples recorded by the jth cell-phone. Under the assumption that each cell-phone in the database has equal prior probability of $1/N$ and adjacent feature vectors are independent, the entropy of a feature set for a given cell phone is the sum of the entropies over all frames. In [19,20] it is shown that an estimate of the mutual information $\hat{I}(\lambda, C)$ (shown to be always less than or equal to the true value, $\hat{I}(\lambda, C) \le I(\lambda, C)$) can be computed as,

$$\begin{aligned}\hat{I}(\lambda, C) &= -E\left[\log_2\left(\sum_{j=1}^{N}\frac{1}{N}p(C|\lambda_j)\right)\right] + E\left[\log_2 p(C|\lambda)\right]\\ &\approx -\frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\log_2\left(\frac{1}{N}\sum_{k=1}^{N}p\left(\mathbf{c}_t^j|\lambda_k\right)\right)\\ &\quad + \frac{1}{NT}\sum_{j=1}^{N}\sum_{t=1}^{T}\log_2 p\left(\mathbf{c}_t^j|\lambda_j\right), \end{aligned} \quad (4)$$

where λ_j represents the jth cell-phone in database. Due to the need of probability density functions (pdfs) to compute the entropy of features for each cell-phone model, we use the GMMs with 32 components as explained in the next section.

An estimate of the conditional entropy $h(C|\lambda)$ and the mutual information $I(C, \lambda)$ in (3) using (4) is shown in Fig. 4 for both TIMIT and LIVE RECORDS databases (details of the datasets will be described in the next section) as a function of the number of MFCC features extracted from the whole utterance, speech parts only, and the non-speech parts only, by varying it from 2 to 24. In Fig. 4, the computation of the mutual information and conditional entropy are done in closed set condition, the training data of each cell-phone is used for model training and mutual information computation, in order to perform mutual information analysis independent from test data, similar to [20]. We assume that the device-specific information is contained in the MFCCs, where the amount of useful information is coded in some form, and its amount depends on the number of MFCCs used. This is clearly seen from Fig. 4. Here, the mutual information increases with the number of MFCCs up to some extent (at most to its true value), but further increase in the number of MFCCs may result in a decrease in mutual information. The turning point may be interpreted as the point at which the number of features (the smoothness of the signal spectrum) is matched to the structure of the discriminatory information coded in recorded speech.

The conditional entropy is a concave function of the number of MFCCs, and shows the amount of uncertainty in C given the model λ. Therefore, smaller conditional entropy corresponds to higher mutual information across C and λ. Clearly, features extracted from the non-speech parts posses the least conditional entropy and therefore attain the highest mutual information whereas features obtained from the speech parts only or the whole utterance yield smaller mutual information (as intuitively expected and discussed above). Thus, we expect to obtain higher recognition rates when the features are extracted from the non-speech parts of the recorded signal. These observations about the MI and conditional entropy hold for the LFCC features, as well.

2.4. Gaussian mixture model classifier

Gaussian mixture model (GMM) is a probabilistic classification method used in many applications including speaker, language and face recognition [24–26]. GMM represents each class as a weighted sum of M multivariate Gaussian components as,

$$f_{\mathbf{X}|\lambda}(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i f_{\mathbf{X}_i}(\mathbf{x}_i), \tag{5}$$

where w_i is the mixture weight and $f_{\mathbf{X}_i}(\mathbf{x}_i)$ is a D-variate Gaussian density function with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$:

$$f_{\mathbf{X}_i}(\mathbf{x}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_i)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_i)\right\}. \tag{6}$$

A complete GMM model is parameterized by the mean vectors, covariance matrices and weights of all component densities, $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{M}$.

Training a GMM consists of estimating the model parameters, w_i, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ by maximizing the log-likelihood function of training vectors $X = \{\mathbf{x}_1, \dots, \mathbf{x}_{\mathrm{T}}\}$ with respect to model λ, defined as,

$$\mathcal{L}(X|\lambda) = \frac{1}{T}\sum_{t=1}^{T} \log f_{\mathbf{X}|\lambda}(\mathbf{x}_t|\lambda). \tag{7}$$

Iterative *expectation maximization* (EM) algorithm [24] is used to estimate the model parameters by maximizing (7) on the training set. In the recognition phase of an identification system, the decision is based on the maximum average log-likelihood of test vectors $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_{\mathrm{T}}\}$ and GMMs stored in the database. Namely,

$$\delta = \arg\max_{1\le j\le N} \mathcal{L}(Y|\lambda_j). \tag{8}$$

The ML training of GMM aims to maximize the overall likelihood of training data for a cell-phone.

Besides generative training of a GMM with ML criterion, we consider discriminative training of device-specific models using Maximum Mutual Information (MMI) criterion [27–29]. The advantage of MMI training is that it maximizes the probability of correct decision by taking all the training samples of each class into account. Another advantage of training GMM with MMI criterion is that it does not suffer from the shortcoming of ML method with short training data.

It is important to note that training GMM with MMI is different from computing the mutual information of MFCC features described in the previous section. Thus, except for the names, there is no direct link between MI of MFCCs and GMM-MMI methods. Training a GMM with MMI corresponds to estimating its parameters (weights, means and covariances) by maximizing the posterior probability of all training features. Formally, the objective function for MMI training is [27–30],

$$F^{\mathrm{MMI}} = \sum_{j=1}^{N}\sum_{r=1}^{R} \log \frac{p(X_r^j|\lambda_j)P(\lambda_j)}{\sum_{k=1}^{N} p(X_r^j|\lambda_k)P(\lambda_k)}, \tag{9}$$

where X_r^j is the collection of MFCC vectors extracted from the rth training utterance recorded by the jth cell-phone, $p(X_r^j|\lambda_j)$ is the likelihood of the r-th training utterance, given the correct cell-phone model, λ_j. R is the number of speech utterances used in training, and the denominator represents the unconditional probability density, $p(X_r^j)$. By assuming the prior probabilities of all classes (cell-phones) to be equal, the prior terms $P(\lambda_j)$ and $P(\lambda_k)$ in (9) can be ignored. In practice, (9) is maximized using a so-called *extended Baum-Welch* (EBW) algorithm [27]. It is an iterative procedure that requires an initial set of models. To this end, we use the ML-trained models. We have used STK toolkit[1] to construct our GMM-MMI based classification system, and trained the models using 20 EBW iterations. Generally the number of EBW iteration is selected between 10 and 20 [29,31–34]. We found that 20 EM and EBW iterations are sufficient for the convergence. In the recognition phase, decision is made according to,

$$\delta = \arg\max_{1\le j\le N} \log\left\{\frac{p(Y|\lambda_j)^{1/T}}{\sum_{k=1}^{N} p(Y|\lambda_k)^{1/T}}\right\}, \tag{10}$$

where T is the number of feature vectors in Y.

2.5. Support vector machine classifier

Support vector machine (SVM) is another powerful classification method. SVM is originally a binary classifier which models the decision boundary (*separating hyperplane*) between two classes. Training an SVM consists of finding the separating hyperplane between two classes with maximum margin. SVM has become a *de facto* reference classification method in many applications. Since speech is a dynamic signal, i.e., its amplitude and frequency content change over time, features obtained from a speech signal are variable-length sequences of D-dimensional vectors rather than a single vector. Thus *sequence kernel* approach was proposed for speech applications of SVM [35–37].

One of the simplest sequence kernel methods is *generalized linear discriminant sequence kernel* (GLDS-SVM) [35,36]. In GLDS-SVM, spectral features are mapped to higher dimensional space by polynomial expansion with monomials (each combination of feature vector components) up to a certain degree m. For example, given a 2-dimensional feature vector $\mathbf{x} = [x_1\ x_2]^{\mathrm{T}}$, its expansion of order 2 is computed as $\mathbf{b}(\mathbf{x}) = [1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2]^{\mathrm{T}}$. For a D-dimensional feature vector, the dimensionality of expanded feature vector is $\binom{D+m}{m} = \frac{(D+m)!}{D!m!}$ where m denotes the maximum monomial order. In practice, the dimensionality of the expanded feature vectors becomes too large when $m > 3$; for example, when $D = 24$ and $m = 4$, the polynomial expansion leads to a vector of dimension 20475. Therefore, $m = 3$ is generally used [31,38]. Given a training or test feature sequence of a cell-phone, $X = \{\mathbf{x}_1, \dots, \mathbf{x}_{\mathrm{T}}\}$, it is represented by its average expanded vector,

$$\mathbf{b} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{b}(\mathbf{x}_{\mathrm{t}}). \tag{11}$$

[1] http://speech.fit.vutbr.cz/software/hmm-toolkit-stk.

Table 1
The brands and models of cell-phones used in the experiments and their class names.

Class name	Brand and model
H1	HP IPAQ514
L1	LG KE970
M1	Motorola Q
N1	Nokia 2730
N2	Nokia 3600
N3	Nokia 3600
N4	Nokia 6500
N5	Nokia 6670
SA1	Samsung E250
SA2	Samsung E250
SA3	Samsung D900
SO1	Sony K750I
SO2	Sony W880
SO3	Sony W880

Each cell-phone in the database is represented by its characteristic vector (**b**) and SVM model for each cell-phone is trained using linear kernel. More details of SVM-GLDS can be found in [35,16,36,37].

In our experiments we use the LibSVM package [39] with 3rd order polynomial expansion. In our preliminary experiments, expansion orders 1, 2, 3 and 4 were studied and $m = 3$ was found to give the highest accuracy. The advantage of LibSVM package is that, in addition to hard decisions, it also provides class probability estimates. Another advantage of this package is that it supports multi-class training in which *one-against-one* approach is used to train $N(N-1)/2$ binary classifiers for N classes. Details about the implementation of multi-class SVM with probability outputs can be found in [40].

In the recognition stage, given a test vector $\mathbf{b}_{\text{test}}$ the probability that $\mathbf{b}_{\text{test}}$ comes from the ith class: namely, $p_i = p(i|\mathbf{b}_{\text{test}})$, $i = 1, \ldots, N$ is computed and the cell-phone model which produces the maximum probability is determined as the decision.

3. Cell-phone recognition setup

In the experiments, we have used $N = 14$ models of cell-phones. Comparing with the source camera identification studies in [6,7,41,42], where 9, 4, 16, and 6 camera models were used, respectively, we consider $N = 14$ cell-phones to be adequate, at least for the purpose of presenting preliminary results in this emerging field. The collection of brands include Nokia, Samsung, Sony Ericsson, LG, Motorola and HP. Five of Nokia, three of Samsung, one of LG, three of Sony Ericsson, one of Motorola, and one of HP models have been used in the experiments. The brands and models of cell-phones are listed in Table 1. Note that we have three pairs of cell-phones representing exactly the same model and brand (N2–N3, SA1–SA2 and SO2–SO3 in Table 1). Source devices of the same model and brand are expected to be more difficult to discriminate and therefore these three pairs of devices are included in our dataset to establish the performance on such devices.

Source cell-phone recognition in forensic applications must be text- and speaker-independent by virtue of its nature. We have used two different databases to investigate the performance of our cell-phone recognition system. The first database, **TIMIT**, is a popular speech/speaker recognition database which consists of 630 speakers from different dialects of American English (192 females and 432 males). Each speaker reads ten utterances each of which is approximately 3 seconds long. We have randomly selected 24 speakers from the test portion of the database, and 240 sentences of these 24 speakers are played back with PC loud speaker and recorded by each cell-phone in an office environment. With this, we have 240 utterances for each cell-phone, in total 3360 speech recordings. For each cell-phone, we have used 120 recordings for training and the remaining 120 utterances for testing (120 individual testings for each cell-phone and total of 1680 identification trials).

Apart from TIMIT, we have built a second database by recording speech spoken by the same speaker for both training and test sessions and refer it to as **LIVE RECORDS** in the following. The reason of using second database is to test the device discriminating capability of our recognition system under different conditions. For each cell-phone, speech data is recorded in the same room (as was the case in TIMIT recordings), which is about 10 minutes long spoken by the same speaker. Half of the recording (5 minutes) is used to train each phone, and the remaining 5 minutes portion is segmented into 3s long chunks for testing (100 test sentences for each phone, total of 1400 tests). The text content used in two sessions are different. It is seen that in the TIMIT database speakers are different accross the training and test portions whereas the same speaker is used in both portions in the LIVE RECORDS database. However, the text content of speech samples are different.

The recorded speech signals are in the adaptive multi-rate (AMR) compression format for all phones with 8 kHz sampling frequency and 12.2 kbps bit rate. Recordings of each cell-phone are processed in different sessions (14 different recording sessions) but in the same office. During the recording sessions, each cell-phone was located on the same spatial point in turn (at the same distance from the loudspeakers). Our set-up considers cell-phones as ordinary voice recorders; recording over a wireless connection (while a call is in progress) is outside the scope. Therefore, our data does not include transmission channel or speech coding effects; instead, some environmental variability is introduced by controlled additive noise degradation described below.

In the cell-phone detection experiments on TIMIT database, 120 test samples of each cell-phone were scored against each cell-phone. This yields 1680 trials of which 120 are positive (same phone) trials and the remaining 1560 trials are negative (different phone) trials per each cell-phone. Thus, we have a total of 1680 (120 × 14) positive and 21840 (120 × 13 × 14) negative trials. In the LIVE RECORDS database, we have a total of 1400 (100 × 14) positive and 18200 (100 × 13 × 14) negative trials.

For additive noise contamination, we use *Filtering and Noise Adding Tool* (FaNT).[2] It is an open-source tool that follows ITU recommendation for noise adding and filtering. Specifically, it uses psychoacoustic speech level computation based on the ITU recommendation P.56 (*objective measurement of active speech level*). *White* and *babble* noises selected from the NOISEX-92 database[3] are used with 3 different signal-to-noise ratio (SNR) levels, 0, 5 and 10 dB. Let us briefly motivate the selection of our two noise types. Firstly, white noise has constant power spectral density and it strongly masks especially the lower-amplitude higher formants of human speech. Even if not representing a typical real-world case, it is often included as a difficult-to-handle [43] case in both speech and speaker recognition studies. Secondly, babble noise [44], representing an unintelligible mixture of multiple speakers, occurs frequently in our daily life: trains, restaurants, school lobbies and family celebrations to name a few. These are examples of sites where one could illicitly record another person's voice with a voice recorder.

One would claim that in a real scenario, the test data would come from a recording made in an unknown acoustical environment with an unknown orientation between the talker and the phone's microphone, with unknown background noise, with unknown automatic gain settings in the handset's input stage, and unknown effects of particular speech coding algorithm used in the

[2] http://dnt.kr.hs-niederrhein.de/download.html.

[3] http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html.

handset. However, the reason for adding noise digitally is that the parameters of the scenario under which the data is collected must be controllable in order to make a plausible performance assessment for their various and critical values, by altering the value of only the one variable and holding those of the remaining constant.

In the noisy experiments, each cell-phone is trained using its original training samples and noise is added to the test speech samples, inducing a mismatch between training and test conditions. In the additive noise experiments, each cell-phone model is trained using its original training recordings and noise is added to test recordings only.

In the noisy experiments, we first apply power spectral subtraction (as described in [45]) as a pre-processing step to reduce additive noise effect on the signal domain. Then, 24 feature coefficients (MFCCs and LFCCs) are extracted from the enhanced signal. After applying RASTA filtering [46] to the features, their first and second order time derivatives (Δ and Δ^2) are appended which finally yields 72 dimensional feature vectors. The last step is cepstral mean and variance normalization (CMVN) which helps suppressing the effect of noise on the feature level. The SAD labels of original recordings are used to locate speech and non-speech frames for all the recordings, including those with digitally added noise.

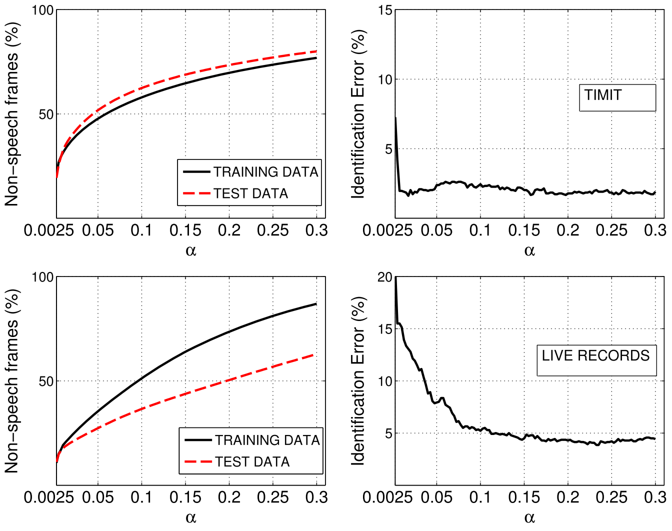

Fig. 5. Number of detected non-speech frames (in %) as a function of the α parameter used in the energy SAD for TIMIT (first row) and LIVE RECORDS (second row) databases.

Such an "oracle" SAD ensures that we use the exact same (number of) frames across the clean and noisy experiments, allowing comparable results across different noise types and SNRs. It is well known that the performance of any energy-based SAD is severely impacted by the presence of additive noise. In forensic casework, the speech/non-speech boundaries would in any case be hand-marked by the forensic analyst, rather than being automatically derived.

We used identification error rate as the performance criterion in the identification experiments. In the detection experiments, we used equal error rate (EER), which is the error rate at which false alarm rate (P_{FA}) and miss rates (P_{FR}) are equal. In the detection experiments, in addition to EER, detection error trade-off (DET) curves [47] are also presented, which is a graphical representation of error rates illustrating the tradeoff between P_{FA} vs P_{FR}.

4. Experimental results

4.1. Effect of SAD parameters

We first optimize the α parameter of the energy SAD separately for TIMIT and LIVE RECORDS. We consider 120 different values of α between $0.0025 \le \alpha \le 0.3$ and compute the number of detected non-speech frames for training and test data of each dataset and the corresponding identification error rates. We used GMM classifier trained with ML (GMM-ML) criterion using 12 MFCCs in this preliminary experiment. Fig. 5 shows the number of non-speech frames (in %) and identification error rates (in %) as a function of α. Recall that $\alpha = 0$ implies *not using any frames*, and is therefore not considered. As α increases, the amount of detected non-speech frames increases. The number of non-speech frames for both training and test data of TIMIT have similar trends and both increase when α increases. For LIVE RECORDS, larger number of non-speech frames are detected in the training data compared to test data. The lowest error rates are obtained with $\alpha = 0.0175$ and $\alpha = 0.2350$ for the TIMIT and LIVE-RECORDS datasets, respectively. We have used these values for the remaining experiments.

4.2. Effect of the number of features

Next, we analyze the source cell-phone identification performance as a function of the number of MFCCs. Fig. 6 shows the identification error rate for the GMM-ML, GMM-MMI and SVM classifiers on both datasets. The lowest identification error rates

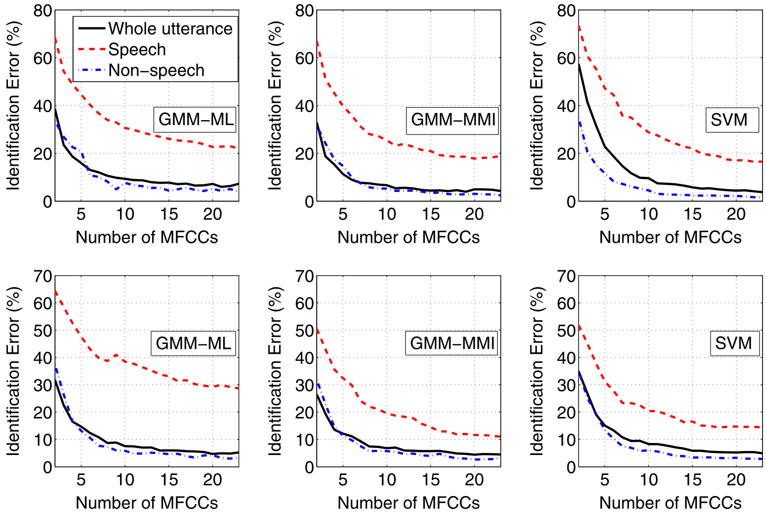

Fig. 6. Identification error rates (in %) using GMM-ML, GMM-MMI and SVM classifiers for TIMIT (first row) and LIVE RECORDS (second row) databases.

Table 2
Confusion table for GMM-MMI-based cell-phone identification on TIMIT database using speech parts only.

	H1	L1	M1	N1	N2	N3	N4	N5	SA1	SA2	SA3	SO1	SO2	SO3
H1	120	0	0	0	0	0	0	0	0	0	0	0	0	0
L1	0	119	0	0	0	0	0	1	0	0	0	0	0	0
M1	0	0	119	0	0	0	0	0	0	0	0	1	0	0
N1	0	0	0	110	1	5	2	0	1	0	0	1	0	0
N2	0	0	0	1	76	9	17	2	2	0	4	1	8	0
N3	0	0	0	3	26	57	22	0	1	2	1	4	4	0
N4	0	0	0	8	3	2	94	0	2	5	0	2	3	1
N5	0	4	0	0	0	0	0	104	6	1	1	3	0	1
SA1	0	0	0	0	0	0	0	0	117	2	1	0	0	0
SA2	0	1	0	0	0	0	0	0	0	119	0	0	0	0
SA3	0	0	0	0	0	1	0	0	19	0	98	0	2	0
SO1	0	0	0	0	0	0	0	0	1	0	0	115	1	3
SO2	0	0	0	0	0	0	0	0	0	1	0	0	78	41
SO3	0	0	0	0	0	0	0	18	0	1	0	23	14	64

Table 3
Confusion table for GMM-MMI-based cell-phone identification on TIMIT database using non-speech parts only.

	H1	L1	M1	N1	N2	N3	N4	N5	SA1	SA2	SA3	SO1	SO2	SO3
H1	120	0	0	0	0	0	0	0	0	0	0	0	0	0
L1	0	120	0	0	0	0	0	0	0	0	0	0	0	0
M1	0	0	120	0	0	0	0	0	0	0	0	0	0	0
N1	0	0	0	120	0	0	0	0	0	0	0	0	0	0
N2	0	0	0	0	112	8	0	0	0	0	0	0	0	0
N3	0	0	0	0	10	110	0	0	0	0	0	0	0	0
N4	0	1	0	2	0	0	115	0	0	1	1	0	0	0
N5	0	0	0	0	0	0	0	120	0	0	0	0	0	0
SA1	0	0	0	0	0	0	0	0	120	0	0	0	0	0
SA2	0	2	0	0	0	0	0	0	0	118	0	0	0	0
SA3	0	0	0	0	0	0	0	0	0	0	120	0	0	0
SO1	0	0	0	0	0	0	0	0	0	0	0	120	0	0
SO2	0	0	0	0	0	0	0	0	0	0	0	0	119	1
SO3	0	0	0	0	0	0	0	0	0	0	0	0	9	111

Table 4
Identification rates (in %) for GMM and SVM classifiers using 24 MFCCs and LFCCs. For a given classifier, all the differences between the features extracted from the speech-only parts and those extracted from the whole utterance or non-speech parts are statistically significant according to McNemar's test with 95% confidence.

Classifier	TIMIT						LIVE RECORDS					
	Whole		Speech		Non-Speech		Whole		Speech		Non-Speech	
	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC
GMM-ML	92.61	94.16	78.03	76.07	95.18	96.79	94.77	95.07	72.85	65.93	96.12	94.57
GMM-MMI	95.65	95.06	82.73	81.67	97.91	94.88	95.21	94.36	89.35	85.21	98.21	95.29
SVM	96.42	96.43	83.63	83.51	98.39	98.27	95.14	93.36	85.93	81.07	97.03	94.93

are obtained when features are extracted from the non-speech parts whereas extracting MFCCs from the speech parts yields systematically the highest identification error rates. For the SVM classifier, the relative difference on identification error is larger when small number of features are used. The SVM classifier outperforms GMM-ML on both datasets. As expected from the results of mutual information graphs (Fig. 4), in most cases, the lowest identification error rate is obtained using 24 MFCCs.

The confusion matrices obtained using speech parts and non-speech parts for TIMIT database using GMM-MMI classifier are given in Tables 2 and 3, respectively. In general, misclassification usually occurs within the same brand of cell-phones rather than across brands, as expected. Clearly, features obtained from the non-speech parts yield considerable improvement in comparison to using the whole utterance or the speech-only parts. When using the speech parts only, especially the pairs with the same brand and model are often misidentified (N2–N3, SA1–SA2 and SO2–SO3 pairs). However, when using the non-speech parts, the recognizer shows fewer confusions in recognizing them.

4.3. Comparison of MFCC and LFCC features

Identification performance comparison of the MFCC and LFCC features using different classifiers is given in Table 4. Twenty-four feature coefficients are used for both feature sets. Regarding the two types of feature sets, MFCC features outperform LFCC features in a vest majority of cases. Further, for both feature sets, features extracted from the non-speech parts yield higher identification rates than features extracted from the whole utterance or the speech parts. For each classifier, the performance differences between the baseline features (extracted from the speech parts) and those extracted from the whole utterance or the non-speech parts were found to be significantly different according to McNemar's test [48]. McNemar's test tabulates the correlation of the correct and incorrect decisions between two systems and counts the number of trials that two systems disagree and then uses the chi-square test statistics to compute the p-values. When $p < 0.05$ the performance difference between two systems are said to be significant with 95% confidence [49].

As seen, both speech and non-speech segments contain discriminatory information for cell-phone recognition. Since speech

Table 5
Ratio of fusion weights ($|w_n|/|w_s|$) computed with logistic regression and identification rates (in %) obtained by fusing the scores of speech and non-speech parts.

Classifier	$\|w_n\|/\|w_s\|$				Fused identification rates (%)				Whole utterance identification rates (%)			
	TIMIT		LIVE RECORDS		TIMIT		LIVE RECORDS		TIMIT		LIVE RECORDS	
	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC	MFCC	LFCC
GMM-ML	2.56	2.16	15.43	12.65	97.97	98.21	96.64	95.57	92.61	94.16	97.77	95.07
GMM-MMI	0.46	0.68	32.24	26.04	97.91	98.33	96.78	95.85	95.65	95.06	95.21	94.36
SVM	1.76	1.81	2.81	1.98	98.92	99.04	96.21	95.28	94.42	96.43	95.14	93.36

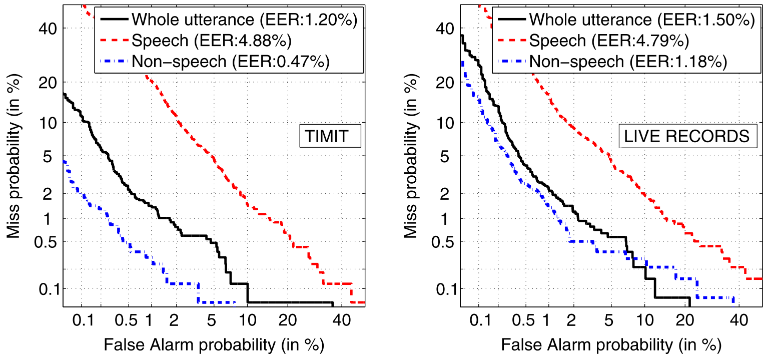

Fig. 7. DET curves for TIMIT and LIVE RECORDS datasets.

and non-speech parts are spectrally very different, one should expect them to contain complementary cues. This motivates us to combine the two classifier output scores (one constructed for the speech parts, the other one from the non-speech parts). To this end, we adopt a linear score fusion of the form $s = w_0 + w_s \times s_s + w_n \times s_n$ where w_0 represents a bias, w_s and w_n are the weights of the speech and non-speech classifiers and s_s (speech) and s_n (non-speech) are the corresponding scores of the classifiers. Linear fusion has the benefit that the weights can be interpreted as the relative importance of each classifier and provide insights to the problem at hand. Linear score fusion, in fact, often produces the most competitive results in state-of-the-art speaker verification [50,51]. We do *not* hand-tune the weights using adhoc grid-search but optimize them using a logistic regression model which provides better generalization. Speaker recognition community has developed useful open-source tools to achieve this task; we utilize *FoCal Multi-class toolkit* [52] to train the fusion weights. It is important to note that the purpose of fusion experiment is to analyze the relative importance of speech and non-speech parts rather than trying to prove that it improves the identification rates.

The ratio of optimized fusion weights $|w_n|/|w_s|$ and the corresponding identification rates obtained through fusion are shown in Table 5. It is important to note that $|w_n|/|w_s| > 1$ indicates that non-speech scores are more discriminative than the speech scores. The last two columns of the table show the identification rates when features are extracted using the whole utterance for TIMIT and LIVE RECORDS databases selected from Table 4. From the table, it is clear that non-speech score weights (w_n) are generally higher than that of speech score weights (w_s), the only exception being GMM-MMI classifier on TIMIT. Fusing the scores of speech and non-speech parts improves the identification rates for each classifier compared with the whole utterance case independent of classifier or feature set. Identification rates after score fusion are higher than the accuracies of best individual rates on TIMIT database (Table 4). This shows that speech and non-speech parts contain complementary device information. However, for LIVE RECORS fusion fails to improve identification rate in comparison to best individual accuracy (e.g. identification rate decreased from 97.03% to 96.21% after score fusion on LIVE RECORDS database using MFCC features with SVM classifier).

4.4. Detection experiments

The DET curves for the source cell-phone detection experiments using SVM classifier with 24 MFCCs are shown in Fig. 7. Similar to the results of the identification experiments, using only the non-speech parts of the recorded speech signal yields higher accuracy compared to those of using the speech only part or the whole utterance for both datasets. Using only the non-speech parts for feature extraction provides 60.83% and 21.33% relative improvements on the performance compared to that of using the whole utterance, on TIMIT and LIVE RECORDS databases, respectively.

4.5. Effect of additive noise

In the noisy experiments, first, the effect of pre-processing (spectral subtraction) and feature normalization (CMVN) *under mismatched conditions* are compared. To this end, identification rates for white and babble noises with 5 and 10 dB SNR levels, as an example, are displayed in Table 6 (similar conclusions hold for other SNR levels considered). Spectral subtraction reduces the identification rate of original test recordings considerably. However, identification rate obtained by feature normalization (CMVN) is slightly lower and identification rate when both pre- and post-processing are applied is higher than the accuracy of baseline MFCCs (without any pre- and post processing). Identification rates reduce dramatically under additive noise contamination, as expected. In noisy case for TIMIT database, spectral subtraction reduces the identification rate for white noise whereas it improves the accuracy for babble noise. Similar observations hold for other cases (when CMVN is applied and when both spectral subtraction and CMVN are applied together). Relative improvement achieved by applying pre-processing before feature extraction and then feature normalization as post-processing is higher than the improvement achieved by spectral subtraction or CMVN. For LIVE RECORDS database in turn, applying pre- and post-processing together yields the best identification rate for white noise whereas feature extraction followed by CMVN achieves the highest performance for babble noise. In general, applying both spectral subtraction and feature normalization yields the highest identification accuracies for both databases (except for the babble noise case on LIVE RECORDS).

Table 6
Comparison of identification rates (in %) under additive noise using GMM-MMI classifier with MFCC features extracted from whole utterance with and without pre- and post processing (SS: spectral subtraction, and CMVN: cepstral mean and variance normalization).

Noise type	TIMIT				LIVE RECORDS			
	Pre-processing / Post-processing				Pre-processing / Post-processing			
	None	SS	CMVN	SS + CMVN	None	SS	CMVN	SS + CMVN
Original	94.70	89.72	94.34	95.41	91.85	74.71	78.14	95.35
White (0 dB)	11.19	8.63	8.92	14.10	7.71	11.85	10.35	12.57
White (5 dB)	13.63	9.70	11.96	16.25	15.21	14.35	10.64	16.00
Babble (0 dB)	7.20	9.22	10.35	11.54	14.92	13.85	18.35	15.50
Babble (5 dB)	7.38	13.69	15.53	16.07	25.28	17.50	32.64	25.42

Table 7
Identification rates (in %) under additive noise using GMM-MMI classifier when spectral subtraction, CMVN and RASTA filtering are applied on feature vectors (Wh., Sp. and NS. correspond to whole utterance, speech and non-speech, respectively.)

	SNR (dB)	TIMIT						LIVE RECORDS					
		MFCCs			LFCCs			MFCCs			LFCCs		
		Wh.	Sp.	NS.	Wh.	Sp.	NS.	Wh.	Sp.	NS.	Wh.	Sp.	NS.
	Original	92.67	66.54	**94.94**	91.66	64.10	**92.20**	76.78	32.57	**84.28**	70.71	35.00	**81.78**
White	10	33.75	**38.75**	16.78	35.11	**44.82**	16.01	**39.21**	21.00	25.85	21.50	**24.64**	18.92
	5	27.91	**27.85**	15.17	26.66	**33.27**	14.82	**27.35**	14.57	17.14	15.85	**18.50**	14.07
	0	**19.22**	18.21	12.85	19.40	**22.50**	13.69	**16.35**	12.14	12.00	**16.21**	15.64	12.00
Babble	10	21.19	**42.55**	12.44	26.07	**48.69**	14.34	21.14	**42.57**	12.42	**34.50**	32.57	20.14
	5	18.39	**30.35**	10.35	22.32	**34.82**	13.21	**27.35**	14.57	17.14	15.85	**18.50**	14.07
	0	15.17	**23.33**	12.32	18.51	**27.97**	12.61	**17.42**	12.00	13.85	15.50	**20.07**	12.78

The previous experiments have focused on the effects of speech enhancement, feature normalization and choice of classifier on a limited set of data conditions. Let us now fix both the feature processing chain (spectral subtraction, delta features, RASTA and CMVN) and the back-end classifier (GMM-MMI) and focus on noise type, SNR, feature set and frame selection principle (whole utterance, speech or non-speech). These results, for both corpuses, are displayed in Table 7. Identification accuracies reduce dramatically independent of the noise type or features in the case of added noise. When features are extracted from the speech parts, white noise represents a more challenging case from the two types of noises. In the cases of non-speech parts and whole utterance babble noise is more challenging than the white noise, in general.

In contrast to the results of clean recordings presented in the previous subsection, extracting features from the speech parts yield the highest identification rates independent of the feature type or the SNR (except for the MFCC features on LIVE RECORDS). The reason for obtaining smaller identification rates by using features extracted from the non-speech parts is probably because, non-speech parts have smaller energy levels thus these parts are corrupted by noise more than the speech parts. With the feature extraction setup used for additive noise experiments (spectral subtraction, RASTA filtering and CMVN), the recognition accuracy of clean recordings are lower than the identification rates given in Table 4. This might be because spectral subtraction and RASTA filtering techniques aim to reduce additive noise effect in the signal and feature levels, respectively. However, they reduce relevant information when these methods are applied on clean signals, as expected.

Comparing the two sets of features, in contrast to the clean tests results (Table 4), under additive noise contamination LFCC features outperforms MFCCs in general. For example, for white noise (0 dB SNR) LFCCs yield approximately 23% relative improvement over MFCCs (18.21% → 22.50%).

5. Discussion

In speaker, language, and speech recognition, non-speech frames are generally not used in the recognition system since they do not contain relevant information for such applications. However, different from the previous microphone or handset identification studies that used whole signals to extract features [10–12,15], in this study we show that non-speech portions of the signal are more important for identifying the recording device. The results on clean data indicate that features extracted from non-speech parts of the signal contain higher mutual information than those from speech parts or whole utterance (Fig. 4). Twenty-four MFCCs are found to give the highest MI when features are extracted from non-speech parts. These results are further supported by the classification results in Fig. 6 and Table 4.

Similar to the findings for mutual information, identification rates of features extracted using the whole utterance is higher than that of extracted from the speech-only parts but lower than that of extracted from the non-speech parts. This could be because the whole utterance is a mixture of the two extreme parts and naturally has the performance lying in between but closer to the whole utterance. We believe that the performance difference between whole utterance and non-speech cases would be larger if the SAD was perfect. The imperfection mixes some fragments of speech signal into the noise-like non-speech parts and some fragments of non-speech segments into the speech parts.

SVM and GMM-MMI classifiers outperform GMM-ML method in all cases. The performance improvement of GMM-MMI over GMM-ML is expected because generative training of a GMM (ML) aims to optimize the model parameters so that the estimated model reproduces the training data with the greatest probability whereas discriminative training (MMI) focuses on learning the boundaries between classes by considering all classes. In [16], we obtained 96.42% and 92.28% identification rates using 12 MFCCs with SVM classifier for TIMIT and LIVE RECORDS databases, respectively. In that study the features were extracted from whole utterances. In the same study, vector quantization (VQ) classifier yielded recognition rates of 92.56% (TIMIT) and 92.57% (LIVE RECORDS). A com-

Table 8
Comparison of identification rates (in %) obtained in [16] and in this study.

Classifier	TIMIT	LIVE RECORDS
VQ [16]	92.56	92.57
SVM [16]	96.42	92.28
GMM-ML [this study]	96.79	96.12
GMM-MMI [this study]	97.91	98.21
SVM [this study]	98.39	97.03

parison of the highest identification rates obtained in [16] and this work are summarized in Table 8.

When linear logistic regression based score fusion is applied to the scores of speech and non-speech segments (Table 5) the optimum estimated weight of the non-speech scores are higher than that of speech segments, suggesting that non-speech segments are more discriminative than the speech segments. Besides, they contain complementary information and fusing the scores improves the identification rates, in general, in comparison to the accuracies of whole utterance. MFCC features outperforms LFCCs in general independent of classifier for both TIMIT and LIVE RECORDS databases on clean data. However, when score fusion is applied, identification rates obtained with LFCCs are slightly higher than that of MFCCs for TIMIT database.

Recognition rates reduce dramatically under additive noise, as expected (Tables 6 and 7). Interestingly, in contrast to the findings on clean data, the features extracted from the non-speech parts of the signal yield smaller identification rates than the features obtained from the speech-only parts or whole utterance. This is probably because non-speech parts are corrupted by the noise more than the speech-only parts since they have smaller segmental signal-to-noise ratio (SNR). Signal pre-processing by spectral subtraction improves identification rates under babble noise contamination but reduces the accuracy for white noise on TIMIT database (Table 6) compared to the accuracy of baseline MFCCs without any pre- or post-processing. However, for LIVE RECORDS database it reduces the identification rates for both white and babble noises but relative degradation under white noise is considerably smaller than babble noise.

From Table 6, applying post-processing (CMVN) reduces the identification accuracy under white noise case. However, it considerably improves the accuracy for babble noise (e.g. identification rate improves from 7.38% to 15.53% for TIMIT database). In general, the highest identification rates under additive noise are obtained when both spectral subtraction and CMVN are applied. The identification rates with the full feature extraction setup which consists of applying pre- and post-processing, applying RASTA filtering and appending delta features which are summarized in Table 7 are considerably higher than that of baseline features with pre- and/or post-processing given in Table 6.

6. Conclusion

We proposed source cell-phone recognition system that utilizes features extracted from non-speech parts of the signal. We have shown that extracting features using non-speech parts yields higher mutual information and hence higher recognition rates in comparison to features extracted from speech-only parts or the whole utterance. Experiments conducted on two different datasets (TIMIT and LIVE RECORDS) using three classifiers (SVM, GMM-ML and GMM-MMI) and two sets of features (MFCC and LFCC) indicate that non-speech parts are more representative of the source device than the speech parts when the data is relatively clean and the highest identification rates and mutual information were obtained using 24 features (MFCC and LFCC) in most cases. However, under additive noise, since speech parts have higher segmental signal-to-noise ratio, extracting features from the speech parts was found more successful. Because of the considerable reduction on the recognition accuracy under additive noise, addressing noise-robust feature extraction methods (e.g. features extracted from phase rather than magnitude spectra) for this challenging task are necessary in future work. Since the speech/non-speech detection used in the experiments has a major effect on the performance of source identification using clean data and it is not error-free, comparison of different SAD methods using more up-to-date devices such as iPhone and Samsung Galaxy would be interesting as well.

Acknowledgments

The work was partially funded by Academy of Finland (projects 253120 and 283256). The authors would like to thank the anonymous reviewers for their detailed feedback.

References

[1] R. Poisel, S. Tjoa, Forensics investigations of multimedia data: a review of the state-of-the-art, in: Proc. Int. Conf. IT Security Incident Management and IT Forensics, IMF, 2011, pp. 48–61.

[2] N. Khanna, A.K. Mikkilineni, A.F. Martone, G.N. Ali, G.T.-C. Chiu, J.P. Allebach, E.J. Delp, A survey of forensic characterization methods for physical devices, Digit. Investig. 3 (2006) 17–28.

[3] G.C. Langelaar, I. Setyawan, R.L. Lagendijk, Watermarking digital image and video data: a state-of-the-art overview, IEEE Signal Process. Mag. 17 (2000) 20–46.

[4] I. Avcibas, N.D. Memon, B. Sankur, Steganalysis using image quality metrics, IEEE Trans. Image Process. 12 (2003) 221–229.

[5] M.W.M. Stamm, K.J.R. Liu, Information forensics: an overview of the first decade, IEEE Access 1 (2013) 167–200.

[6] J. Lukáŝ, J. Fridrich, M. Goljan, Digital camera identification from sensor pattern noise, IEEE Trans. Inf. Forensics Secur. 1 (2006) 205–214.

[7] A.E. Dirik, H.T. Sencar, N. Memon, Digital single lens reflex camera identification from traces of sensor dust, IEEE Trans. Inf. Forensics Secur. 3 (2008) 539–552.

[8] A.E. Dirik, H.T. Sencar, N.D. Memon, Flatbed scanner identification based on dust and scratches over scanner platen, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'09, 2009, pp. 1385–1388.

[9] H. Malik, H. Zhao, Recording environment identification using acoustic reverberation, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'12, 2012, pp. 1833–1836.

[10] C. Krätzer, A. Oermann, J. Dittmann, A. Lang, Digital audio forensics: a first practical evaluation on microphone and environment classification, in: Proc. Int. Workshop on Multimedia & Security, MM&Sec'07, 2007, pp. 63–74.

[11] R. Buchholz, C. Krätzer, J. Dittmann, Microphone classification using Fourier coefficients, in: Proc. Information Hiding, 2009, pp. 235–246.

[12] H. Malik, J.V. Miller, Microphone identification using higher-order statistics, in: Proc. AES Int. Conf. Audio Forensics, 2012, pp. 2–5.

[13] D. Garcia-Romero, C. Espy-Wilson, Automatic acquisition device identification from speech recordings, J. Acoust. Soc. Am. 125 (2009) 2530.

[14] D. Garcia-Romero, C.Y. Espy-Wilson, Speech forensics: automatic acquisition device identification, J. Acoust. Soc. Am. 127 (2010) 2044.

[15] D. Garcia-Romero, C.Y. Espy-Wilson, Automatic acquisition device identification from speech recordings, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'10, 2010, pp. 1806–1809.

[16] C. Hanilçi, F. Ertaş, T. Ertaş, Ö. Eskidere, Recognition of brand and model of cell-phones from recorded speech signals, IEEE Trans. Inf. Forensics Secur. 7 (2012) 625–634.

[17] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech Commun. 52 (2010) 12–40.

[18] C. Hanilçi, F. Ertas, Optimizing acoustic features for source cell-phone recognition using speech signals, in: Proc. ACM Workshop on Information Hiding and Multimedia Security, IH&MMSec '13, 2013, pp. 141–148.

[19] T. Eriksson, An information-theoretic perspective on feature selection in speaker recognition, IEEE Signal Process. Lett. 12 (2005) 500–503.

[20] G. Garcia, T. Eriksson, Study of mutual information for speaker recognition features, in: Proc. European Signal Processing Conference, EUSIPCO 2010, 2010, pp. 601–605.

[21] D. Ellis, J. Bilmes, Using mutual information to design feature combinations, in: Proc. Int. Conf. Spoken Language Processing, ICSLP 2000, 2000, pp. 79–82.

[22] R. Fernández, J.-F. Bonastre, D. Matrouf, J.R. Calvo, Feature selection based on information theory for speaker verification, in: Proc. Iberoamerican Conf. on Pattern Recognition, CIARP 2009, 2009, pp. 305–312.

[23] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley-Interscience, 1991.

[24] D.A. Reynolds, R. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1995) 72–83.

[25] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, J.R. Deller Jr., Approaches to language identification using Gaussian mixture models and shifted delta cepstral features, in: Proc. Int. Conf. Spoken Language Processing (ICSLP 2002), 2002, pp. 89–92.

[26] C. Sanderson, S. Bengio, Y. Gao, On transforming statistical models for non-frontal face verification, Pattern Recognit. 39 (2006) 288–302.

[27] P. Matějka, L. Burget, P. Schwarz, J. Černocký, BRNO University of Technology system for NIST 2005 language recognition evaluation, in: Proc. Odyssey: The Speaker and Language Recognition Workshop, 2006, pp. 57–64.

[28] L. Burget, P. Matějka, J. Černocký, Discriminative training techniques for acoustic language identification, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'06, 2006, pp. 209–212.

[29] V. Hubeika, L. Burget, P. Matějka, P. Schwarz, Discriminative training and channel compensation for acoustic language recognition, in: Proc. Interspeech, vol. 9, 2008, pp. 301–304.

[30] V. Hubeika, I. Szöke, L. Burget, J. Černocky, Maximum likelihood and maximum mutual information training in gender and age recognition system, in: Proc. Int. Conf. Text, Speech and Dialogue, TSD'07, 2007, pp. 496–501.

[31] T. Kinnunen, R. Saeidi, J. Leppänen, J.P. Saarinen, Audio context recognition in variable mobile environments from short segments using speaker and language recognizers, in: Proc. The Speaker and Language Recognition Workshop, Odyssey, 2012, pp. 301–311.

[32] G.F. Choueiter, G. Zweig, P. Nguyen, An empirical study of automatic accent classification, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'08, 2008, pp. 4265–4268.

[33] P.A. Torres-Carrasquillo, D.E. Sturim, D.A. Reynolds, A. McCree, Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition, in: Interspeech, 2008, pp. 723–726.

[34] W. Shen, D.A. Reynolds, Improved GMM-based language recognition using constrained MLLR transforms, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'08, 2008, pp. 4149–4152.

[35] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, P.A. Torres-Carrasquillo, Support vector machines for speaker and language recognition, Comput. Speech Lang. 20 (2006) 210–229.

[36] W.M. Campbell, Generalized linear discriminant sequence kernels for speaker recognition, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'02, 2002, pp. 161–164.

[37] K. Daoudi, J. Louradour, A comparison between sequence kernels for SVM speaker verification, in: Proc. IEEE Int. Conf. Audio, Speech and Signal Processing, ICASSP'09, 2009, pp. 4241–4244.

[38] J. Louradour, K. Daoudi, F. Bach, SVM speaker verification using incomplete Cholesky decomposition sequence kernel, in: Proc. The Speaker and Language Recognition Workshop, Odyssey, 2006, pp. 1–5.

[39] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 27:1–27:27, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[40] T.-F. Wu, C.-J. Lin, R.C. Weng, Probability estimates for multi-class classification by pairwise coupling, J. Mach. Learn. Res. 5 (2004) 975–1005.

[41] O. Çeliktutan, B. Sankur, İ. Avcıbaş, Blind identification of source cell-phone model, IEEE Trans. Inf. Forensics Secur. 3 (2008) 553–566.

[42] C.-T. Li, Source camera identification using enhanced sensor pattern noise, IEEE Trans. Inf. Forensics Secur. 5 (2010) 280–287.

[43] M.S. Zilovic, R.P. Ramachandran, R.J. Mammone, Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions, IEEE Trans. Speech Audio Process. 6 (1998) 260–267.

[44] N. Krishnamurthy, J.H.L. Hansen, Babble noise: modeling, analysis, and applications, IEEE Trans. Audio Speech Lang. Process. 17 (2009) 1394–1407.

[45] P.C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, 2007.

[46] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech Audio Process. 2 (1994) 578–589.

[47] A.F. Martin, G.R. Doddington, T. Kamm, M. Ordowski, M.A. Przybocki, The DET curve in assessment of detection task performance, in: Proc. EUROSPEECH, 1997.

[48] B. Bostanci, E. Bostanci, An evaluation of classification algorithms using McNemar's test, in: BIC-TA'12, 2012, pp. 15–26.

[49] D.A. van Leeuwen, A.F. Martin, M.A. Przybocki, J.S. Bouten, Nist and nfi-tno evaluations of automatic speaker recognition, Comput. Speech Lang. 20 (2006) 128–158.

[50] A. El Hannani, D. Petrovska-Delacrétaz, G. Chollet, Linear and non-linear fusion of ALISP-based and GMM systems for text-independent speaker verification, in: Proc. The Speaker and Language Recognition Workshop, Odyssey, 2004.

[51] V. Hautamäki, T. Kinnunen, F. Sedlak, K.-A. Lee, B. Ma, H. Li, Sparse classifier fusion for speaker verification, IEEE Trans. Audio Speech Lang. Process. 21 (2013) 1622–1631.

[52] N. Brümmer, Focal multi-class toolkit, @Available https://sites.google.com/site/nikobrummer/focalmulticlass, 2014.

Cemal Hanilçi received the B.Sc., M.Sc. and Ph.D. degrees from Uludağ University, Turkey, in 2005, 2007 and 2013, respectively, all in Electronic Engineering. From March to December 2011, he was visiting School of Computing, University of Eastern Finland (UEF), Finland, as a visiting researcher. Currently he is an assistant professor in the Department of Electrical and Electronic Engineering, Bursa Technical University, and his research interests include speaker recognition, speech signal processing and audio forensics.

Tomi Kinnunen received the M.Sc., Ph.Lic. and Ph.D. degrees in Computer Science from the University of Joensuu (now University of Eastern Finland, UEF), Finland, in 1999, 2004 and 2005, respectively. From 2005 to 2007, he worked as an associate scientist at the Institute for Infocomm Research (I2R), Singapore. Since 2007, he has been with UEF. From 2010 to 2012, he was funded by a post-doc grant from Academy of Finland and he currently holds position of university researcher. He serves as an associate editor in Digital Signal Processing. He was the chair of Odyssey 2014: the Speaker and Language Recognition Workshop. His primary research interests include speaker recognition, speech analysis, pattern recognition and biometrics.