Speaker Identification with Short Sequences of Speech Frames

Giorgio Biagetti, Paolo Crippa, Alessandro Curzi, Simone Orcioni and Claudio Turchetti
DII – Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche,
Via Brecce Bianche 12, I-60131 Ancona, Italy

Keywords: Speaker Identification, Speaker Recognition, Classification, Speech, Speech Frames, Short Sequences, DKLT, GMM, EM Algorithm, MFCC, Cepstral Analysis, Feature Extraction, Digitized Voice Samples.

Abstract: In biometric person identification systems, speaker identification plays a crucial role as the voice is the more natural signal to produce and the simplest to acquire. Mel frequency cepstral coefficients (MFCCs) have been widely adopted for decades in speech processing to capture the speech-specific characteristics with a reduced dimensionality. However, although their ability to de-correlate the vocal source and the vocal tract filter make them suitable for speech recognition, they show up some drawbacks in speaker recognition. This paper presents an experimental evaluation showing that reducing the dimension of features by using the discrete Karhunen-Love transform (DKLT), guarantees better performance with respect to conventional MFCC features. In particular with short sequences of speech frames, that is with utterance duration of less than 1 s, the performance of truncated DKLT representation are always better than MFCC.

1 INTRODUCTION

Biometric person identification systems based on human speech are increasingly being used as a means for the recognition of people. Among the most popular measurements for identification, voice is the more natural signal to produce and the simplest to acquire, as the telephone system provides an ubiquitous network of sensors for delivering the speech signal (Jain et al., 2004; Bhardwaj et al., 2013). Typical applications are access control, telephone services for transaction authorization in place of password or PIN, speaker diarization.

Speaker recognition is the key research area in developing speaker recognition technologies which utilize speech to recognize, identify or verify individuals (Togneri and Pullella, 2011; Kinnunen and Li, 2010; Reynolds, 2002) and can be categorized into two fundamental modes of operations: identification and verification. In identification systems, the issue is to detect which speaker from a given pool the unknown speech is derived from, while in verification systems the speech of the unknown person is compared against both the claimed identity and against all other speakers (the imposter or background model) (Gish and Schmidt, 1994; Campbell, 1997; Bimbot et al., 2004). Both tasks fall into the general problem of statistical pattern recognition, in which a given pattern is to be assigned to one of a set of different categories (Jain et al., 2000). From this point of view, the main difference between speaker identification and speaker verification is that in the first one the classification is based on a set of S models (one for each speaker), while, in the second case, a total of two models (one for the hypothesized speaker and one for the background model), have to be derived during training.

This paper addresses the problem of speaker identification with short sequences of speech frames, that is with utterance duration of less than 1 s. In particular, as this is a very severe test for speaker identification, we want to investigate for feature representations of voice sample that guarantees the best performance in terms of classification accuracy. This is motivated by the fact that although Mel frequency cepstral coefficients (MFCCs) have demonstrated particularly suitable for speech recognition, they present some drawbacks in speaker recognition. In particular, the speaker variability due to pitch mismatch, that is a specific characteristic that distinguishes different speakers, is greatly mitigated by smoothing property of the MFCC filter bank (Zilca et al., 2006). Besides, with reference to the accuracy of dimensionality reduction techniques and their application to speaker identification, the MFCC linear transform does not guarantees any convergence properties as the dimension of subspace tends to the dimension of the frame.

It is well known that among linear transforms that

Biagetti G., Crippa P., Curzi A., Orcioni S. and Turchetti C..
Speaker Identification with Short Sequences of Speech Frames.
DOI: 10.5220/0005191701780185
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 178-185
ISBN: 978-989-758-077-2

can be used for feature extraction and dimensionality reduction, the best known linear feature extractor is the discrete Karhunen-Love transform (DKLT) expansion. In addition, as robust speaker recognition remains an important problem in speaker identification (Zhao et al., 2012; Maina and Walsh, 2011; Zhao et al., 2014; McLaughlin et al., 2013; Sadjadi and Hansen, 2014), in a recent paper (Patra and Acharya, 2011) it has been shown that principal component analysis (PCA) transformation minimizes the effect of noise and improves the speaker identification rate as compared to the conventional MFCC features.

In this work we want to show that the truncated version of DKLT, that is with a subset of components, exhibits good performance in terms of classification accuracy, without affecting speaker variability as in MFCCs filtering approach occurs. In a comparison with standard approach, experimental results clearly show that truncated DKLT behaves better than MFCC features.

2 SINGLE FRAME SPEAKER IDENTIFICATION

2.1 Bayesian Classification

Let us refer to a frame $y[n]$, $n = 0,\ldots,N-1$, representing the power spectrum of speech signal, extracted from the time domain waveform of the utterance under consideration, through a pre-processing algorithm including pre-emphasis, framing and log-spectrum. Typical duration values for frames ranges from 20 ms to 30 ms (usually 25 ms) and a frame is generated every 10 ms (thus consecutive 25 ms frames generated every 10 ms will overlap by 15 ms).

The problem of classification is in general stated as: *Given a set $\mathcal{W}$ of tagged data (training set), such that each of them is known to belong to one of S classes, and a set $\mathcal{Z}$ of data (testing set) to be classified, determine a decision rule establishing which class an element* $\mathrm{y} \in \mathcal{Z}$ *belongs to.*

Thus in the context of spectrum identification we assume that the speech from each known, verified speaker, for all speakers that need to be identified, is acquired and divided in two sets, $\mathcal{W}$ for training and $\mathcal{Z}$ for testing.

For Bayesian speaker identification, a group of S speakers is represented by the pdf's

$$p_s(\mathrm{y}) = p(\mathrm{y} \mid \theta_s)\,, \qquad s = 1,\cdots,S \tag{1}$$

where θ_s are the parameters to be estimated during training, $\mathrm{y} \in \mathcal{W}$. Thus we can define the vector,

$$\mathrm{p} = [p_1(\mathrm{y}),\cdots,p_S(\mathrm{y})]^T\,. \tag{2}$$

The objective of classification is to find the speaker model θ_s which has the maximum a posteriori probability for a given frame $\mathrm{y} \in \mathcal{Z}$. Formally:

$$\hat{s}(\mathrm{y}) = \underset{1\le s\le S}{\operatorname{argmax}}\{p_r(\theta_s|\mathrm{y})\} = \underset{1\le s\le S}{\operatorname{argmax}}\left\{\frac{p(\mathrm{y}|\theta_s)p_r(\theta_s)}{p(\mathrm{y})}\right\} \tag{3}$$

Assuming equally likely speakers (i.e. $p_r(\theta_s) = 1/S$) and noting that $p(\mathrm{y})$ is the same for all speakers models, the Bayesian classification is equivalent to

$$\hat{s}(\mathrm{y}) = \underset{1\le s\le S}{\operatorname{argmax}}\{p(\mathrm{y}|\theta_s)\}\ , \tag{4}$$

or in a more compact form to

$$\hat{s}(\mathrm{y}) = \arg\{\|\mathrm{p}\|_\infty\}\ , \tag{5}$$

where

$$\|\mathrm{p}\|_\infty = \max_{1\le s\le S}\{p_s(\mathrm{y})\} \tag{6}$$

is the maximum or infinity norm. Thus speaker Bayesian identification reduces to solving the problem stated by (5).

2.2 GMM Model Estimation

The most generic statistical speaker modeling one can adopt is the Gaussian mixture model (GMM) (Reynolds and Rose, 1995). The GMM for the single speaker, is a weighted sum of F components densities and given by the equation

$$p(\mathrm{y}|\theta) = \sum_{i=1}^{F} \alpha_i\, \mathcal{N}(\mathrm{y} \mid \mu_i, \mathrm{C}_i) \tag{7}$$

where α_i, $i = 1,\ldots,F$ are the mixing weights, and

$$\mathcal{N}(\mathrm{y}|\mu_i,\mathrm{C}_i) = \frac{(2\pi)^{-\frac{N}{2}}}{\sqrt{|\mathrm{C}_i|}} \exp\left\{-\frac{(\mathrm{y}-\mu_i)^T\mathrm{C}_i^{-1}(\mathrm{y}-\mu_i)}{2}\right\} \tag{8}$$

represents the density of a Gaussian distribution with mean μ_i and covariance matrix C_i. It is worth noting that α_i must satisfy $0 \le \alpha_i \le 1$ and $\sum_{i=1}^{F}\alpha_i = 1$. θ (the index s is omitted for the sake of notation simplicity) is the set of parameters needed to specify the Gaussian mixture, defined as

$$\theta = \{\alpha_1,\mu_1,\mathrm{C}_1,\ldots,\alpha_F,\mu_F,\mathrm{C}_F\}\,. \tag{9}$$

As the *maximum likelihood* (ML) estimate of θ,

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta}{\operatorname{argmax}}\{\log p(\mathcal{W} \mid \theta)\} \tag{10}$$

with training data $\mathcal{W}$ is difficult to find analytically due to the log of the sum in (10), the usual choice for solving ML estimate of the mixture parameters is the expectation maximization (EM) algorithm. This algorithm is based on a set $\mathcal{H} = \{\mathrm{h}^{(1)},\ldots,\mathrm{h}^{(L)}\}$ of

L labels associated with the L observations, each label being a binary vector $\mathrm{h}^{(\ell)} = [\mathrm{h}_1^{(\ell)}, \ldots, \mathrm{h}_F^{(\ell)}]$, where $\mathrm{h}_i^{(\ell)} = 1$ and $\mathrm{h}_l^{(\ell)} = 0$ for all $l \neq i$, means that the vector $\mathrm{y}^{(\ell)} \in \mathcal{W}$ was generated by the i-th Gaussian component $\mathcal{N}(\mathrm{y}|\mu_i, \mathrm{C}_i)$. The EM algorithm is based on the interpretation of $\mathcal{W}$ as incomplete data and $\mathcal{H}$ as the missing part of the complete data $\mathcal{X} = \{\mathcal{W}, \mathcal{H}\}$. The complete data log-likelihood, i.e. the log-likelihood of $\mathcal{X}$ as though $\mathcal{H}$ was observed, is

$$\log\left[p\left(\mathcal{W},\mathcal{H}|\theta\right)\right] = \sum_{\ell=1}^{L}\sum_{i=1}^{F} \mathrm{h}_i^{(\ell)} \log\left[\alpha_i\, \mathcal{N}(\mathrm{y}^{(\ell)}|\mu_i, \mathrm{C}_i)\right] . \tag{11}$$

In general the EM algorithm computes a sequence of parameter estimates $\left\{\hat{\theta}(p)\, , p = 0, 1, \ldots\right\}$ by iteratively performing two steps:

- *Expectation Step:* computes the expected value of the complete log-likelihood, given the training set $\mathcal{W}$ and the current parameter estimate $\hat{\theta}(p)$. The result is the so-called *auxiliary function*

$$Q\left(\theta|\hat{\theta}(p)\right) = E\left\{\log\left[p\left(\mathcal{W},\mathcal{H}|\theta\right)\right] | \mathcal{W}, \hat{\theta}(p)\right\} . \tag{12}$$

- *Maximization Step:* update the parameter estimate

$$\hat{\theta}(p+1) = \underset{\theta}{\operatorname{argmax}} \left\{Q\left(\theta|\hat{\theta}(p)\right)\right\} \tag{13}$$

  by maximizing the Q-function.

Recently, Figueiredo *et al.* (Figueiredo and Jain, 2002) suggested an unsupervised algorithm for learning a finite mixture model from multivariate data, that overcomes the main lacks of the standard EM approach, i.e. sensitiveness to initialization and selection of number F of components. This algorithm integrates both model estimation and component selection, i.e. the ability of choosing the best number of mixture components F according to a predefined minimization criterion, in a single framework. In particular, it is able to perform an automatic component annihilation directly within the maximization step of the EM iterations.

2.3 The Problem of Dimensionality Reduction

For usually 8 kHz (16 kHz) bandwidth speech the vector y has a dimension $N = 128$ (256). Although the Figueiredo's EM algorithm behaves well with multivariate random vectors, a too large amount of training data would be necessary to estimate the pdf $p(\mathrm{y} \mid \theta_s)$ and, in any case, with such a dimension the estimation problem is impractical.

2.3.1 DKLT Truncation

In order to face the problem of dimensionality, the usual choice is to reduce y to a vector k_M of lower dimension by a linear non-invertible transform H (a rectangular matrix) such that

$$\mathrm{k}_M = \mathrm{H}\,\mathrm{y} , \tag{14}$$

$\mathrm{y} \in \mathbb{R}^N$, $\mathrm{k}_M \in \mathbb{R}^M$, $\mathrm{H} \in \mathbb{R}^{M\times N}$, and $M \ll N$. The vector k_M represents the so-called feature-vector belonging to an appropriate subspace of dimensionality M.

It is well known that, among the allowable linear transforms $\mathrm{H} : \mathbb{R}^N \to \mathbb{R}^M$, the DKLT truncated to $M < N$ orthonormal basis functions, is the one that ensures the minimum mean square error (Therrien, 1992).

More formally, let us consider the vector $y[n]$, $n = 0, \ldots, N-1$, as an observation of the $N \times 1$ real random vector $\mathrm{y} = [y_1, \ldots, y_N]^T$ whose autocorrelation function is given by $\mathrm{R_{yy}} = E\left\{\mathrm{y}\mathrm{y}^T\right\}$, where the symbol $E\{\cdot\}$ denotes the expectation.

Once $\mathrm{R_{yy}}$ is estimated, an orthonormal set $\{\phi_1, \ldots, \phi_N\}$, can be derived as a solution of the eigenvector equations

$$\mathrm{R_{yy}} = \Phi\Lambda\Phi^T \tag{15}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, $\Phi = [\phi_1, \ldots, \phi_N] \in \mathbb{R}^{N\times N}$.

The DKLT of y is defined by the couple of equations

$$\mathrm{k} = \Phi^T \mathrm{y} , \tag{16}$$

$$\mathrm{y} = \Phi\, \mathrm{k} , \tag{17}$$

where $\mathrm{k} = [k_1, \ldots, k_N]^T$ is the transformed random vector (Fukunaga, 1990).

In order to evaluate the effect of truncation on DKLT, let us rewrite (17) as:

$$\mathrm{y} = \Phi\, \mathrm{k} = \Phi_M\, \mathrm{k}_M + \Phi_\eta\, \mathrm{k}_\eta = \mathrm{x}_M + \eta_\mathrm{y} , \tag{18}$$

where $\Phi = [\Phi_M\ ,\ \Phi_\eta]$, being $\Phi_M = [\phi_1, \ldots, \phi_M] \in \mathbb{R}^{N\times M}$, $\mathrm{k}_M \in \mathbb{R}^M$, and (16) as:

$$\begin{bmatrix} \mathrm{k}_M \\ \mathrm{k}_\eta \end{bmatrix} = \begin{bmatrix} \Phi_M^T \\ \Phi_\eta^T \end{bmatrix} \mathrm{y} . \tag{19}$$

In (18)

$$\mathrm{x}_M = \Phi_M\, \mathrm{k}_M , \tag{20}$$

is the truncated expansion, and

$$\eta_\mathrm{y} = \Phi_\eta\, \mathrm{k}_\eta , \tag{21}$$

is the error or residual. The truncation is equivalent to the approximations

$$\mathrm{y} \approx \mathrm{x}_M \quad , \qquad \mathrm{k} \approx \mathrm{k}_T = \begin{pmatrix} \mathrm{k}_M \\ 0 \end{pmatrix} , \tag{22}$$

thus, as k_M is given by

$$\mathrm{k}_M = \Phi_M^T\, \mathrm{y}\,, \tag{23}$$

comparing (23) with (14) yields $\mathrm{H} = \Phi_M^T$. This is equivalent to the PCA that extracts the most important features of data.

It can be shown (Therrien, 1992) that the minimum mean square error $\mathcal{E}_M = E\left\{\eta_{\mathrm{y}}^T \eta_{\mathrm{y}}\right\}$, subject to the constraints $\phi_i^T \phi_i = 1$, $i = M+1, \ldots, N$, is given by

$$\begin{aligned}\mathcal{E}_M &= E\left\{\|\mathrm{y}-\mathrm{x}_M\|^2\right\}\\ &= E\left\{(\mathrm{y}-\mathrm{x}_M)^T(\mathrm{y}-\mathrm{x}_M)\right\} = \sum_{i=M+1}^{N} \lambda_i\,,\end{aligned} \tag{24}$$

where λ_i is the eigenvalue corresponding to the eigenvector ϕ_i. Once the λ_i are arranged in decreasing order, the error $\mathcal{E}_M$ decreases monotonically as the index M increases towards N.

2.4 Bayesian Classification by Truncation

Given a group of S speakers, the pdf's

$$p_s(\mathrm{k}_T) = p(\mathrm{k}_T \mid \theta_s)\,, \qquad s = 1, \cdots, S \tag{25}$$

can be derived, where k_T is the truncation of k, and consequently also the vector

$$\tilde{\mathrm{p}} = [p_1(\mathrm{k}_T), \cdots, p_S(\mathrm{k}_T)]^T\,, \tag{26}$$

which represents an approximation of the vector p in (2), is defined. Thus (5) becomes:

$$\hat{s}(\mathrm{y}) = \arg\{\|\tilde{\mathrm{p}}\|_\infty\}\,. \tag{27}$$

However, since

$$\|\tilde{\mathrm{p}}\|_\infty = \max_{1\le s\le S}\{p_s(\mathrm{k}_T)\}\,, \tag{28}$$

and from (22) we have

$$p_s(\mathrm{k}_T) = p_s(\mathrm{k}_M)\,\delta(\mathrm{k}_\eta)\,, \tag{29}$$

it results

$$\|\tilde{\mathrm{p}}\|_\infty = \max_{1\le s\le S}\{p_s(\mathrm{k}_M)\,\delta(\mathrm{k}_\eta)\} = \max_{1\le s\le S}\{p_s(\mathrm{k}_M)\}\,. \tag{30}$$

As you can see comparing (30) with (6), the dimensionality of classification problem is reduced from N to M, with $M < N$.

3 MULTI-FRAME SPEAKER IDENTIFICATION

The accuracy of speaker identification can be considerably improved using a sequence of frames instead of a single frame alone. To this end let us refer to a sequence of frames defined as $Y = \{\mathrm{y}^{(1)}, \ldots, \mathrm{y}^{(V)}\}$ where $\mathrm{y}^{(v)}$ represents the v-th frame. Using (27) and (30) we can determine the class each frame $\mathrm{y}^{(v)}$ belongs to. Thus the S sets

$$\mathcal{Z}_s = \left\{\mathrm{y}^{(v)} \mid \mathrm{y}^{(v)} \text{ belongs to class } S\right\}\,, \quad s = 1, \ldots S\,, \tag{31}$$

are univocally determined.

Given Y, we define the score for each class s as:

$$r_s(Y) = \mathrm{card}\{\mathcal{Z}_s\}\,, \tag{32}$$

where the operator card$\{\cdot\}$ (cardinality) extracts the number of elements belonging to $\mathcal{Z}_s$. Finally the multi-frame speaker identification is based on:

$$\hat{s}(Y) = \operatorname*{argmax}_{1\le s\le S}\{r_s(Y)\}\,. \tag{33}$$

4 EXPERIMENTAL RESULTS

4.1 Data Base

The experiments were carried out on a large identification corpus based on the audio recordings of five different speakers, two females (A, B) and three males (C, D, E) as reported in Table 1. The material was originally extracted from five freely available Italian audiobooks. All recordings are mono, 8 kilosamples per second, 16 bit, particularly suitable for telephone applications.

Figure 1 shows the block diagram of the proposed front-end employed for feature extraction. At the input of the processing chain a voice activity detection block drops all non speech segments from the input audio records, exploiting the energy acceleration associated with voice onset. The signal is then divided into overlapping frames of 25 ms (200 samples), with a frame shift of 10 ms (80 samples). Hence buffering is required for storing overlapping regions among frames. Besides, before computing the DKLT features, each frame is cleaned up by a noise reduction block based on the Wiener filter. Further enhancements are then performed by a SNR-dependent waveform processing phase, that weights the input noise-reduced frame according to the positions of its smoothed instant energy contour maxima. It is worth noting that noise reduction introduces an overall latency of 30 ms (3 frames) due to its algorithm requiring internal buffering.

The consistency of DBT database in terms of number of frames used for each speaker is reported in Table 2.

Table 1: Recordings used for the creation of the identification corpus. Source: *liber liber* (http://www.liberliber.it/). The material was used both for training and testing purposes.

Speaker	Gender	Audiobook	Chapter	Duration [s]
A	F	"Il giornalino di Gianburrasca" by L. Bertelli	I	761
B	F	"I promessi Sposi" by A. Manzoni	I	2593
C	M	"Fu Mattia Pascal" by L. Pirandello	I	251
D	M	"Le tigri di Mompracem" by E. Salgari	I	838
E	M	"I Malavoglia" by G. Verga	I	1162

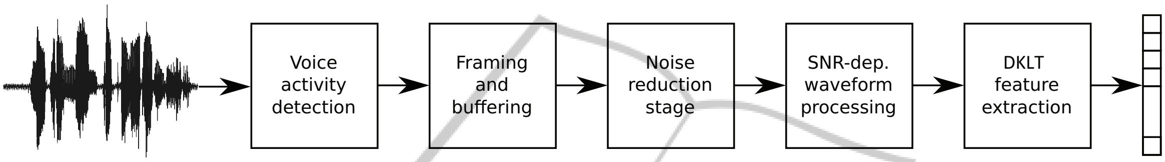


Figure 1: The proposed front-end for feature extraction.

Table 2: Consistency of the databases used for experimental evaluation.

Database	DBT	DB1 (80:20)		DB2 (50:50)		DB3 (20:80)	
Speaker		train	test	train	test	train	test
A	58903	47122	11781	29451	29452	11780	47123
B	195591	156472	39119	97795	97796	39118	156473
C	18867	15093	3774	9431	9434	3773	15094
D	63713	50970	12743	31856	31857	12742	50971
E	91253	73002	18251	45626	45627	18250	73003
Total	428327	342659	85668	214161	214166	85663	342664

From DBT the databases DB1, DB2, and DB3, with different percentage consistency of training and testing subsets, have been derived. More in detail for generating the DB2 database we divided the full DBT database in two datasets containing for each of the five speakers the same proportion of speech frames chosen by considering the first part of them (50%) for training (model evaluation) and the second part (50%) for testing (performance evaluation) purposes. In a similar manner, the DB1 and DB3 databases have been generated by assigning to the testing set / training set ratio the values of 80% / 20% and 20% / 80% respectively.

4.2 Speaker Identification with Truncated DKLT

Several experiments were performed by varying the number of DKLT components retained in the GMM model, with the three different databases in order to evaluate the effect of training data amount on the classification results. An optimum value of 12 DKLT components has been chosen for the GMM model.

With the frames belonging to the testing sets, we ran our classifier and counted the number of occurrences of each recognized type, so as to obtain a confusion matrix for every speaker identification experiment. The resulting confusion matrices are reported in Table 3 for the single-frame, and in Table 4 for the multi-frame ($V = 100$) speaker identification to illustrate in detail the performance of single-frame identification as well as the improvement of the accuracy when 100 consecutive frames (corresponding to a speech sequence of 1s) have been used for the speaker classification.

To gain some insight on the performance of the method, the standard set of performance indices for classification was also extracted from the confusion matrices. To this end we computed the sensitivity, specificity, precision and accuracy, defined as

$$\text{sensitivity} = \text{TP}/(\text{TP}+\text{FN}) \tag{34}$$

$$\text{specificity} = \text{TN}/(\text{TN}+\text{FP}) \tag{35}$$

$$\text{precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{36}$$

$$\text{accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN}) \tag{37}$$

where TP are the true positives (the diagonal elements

Table 3: Single-frame confusion matrices, for the different DB1, DB2, and DB3 databases, obtained by considering 12 DKLT components.

	Recognized				
Input	A	B	C	D	E
	DB1 (80:20)				
A	9885	688	349	312	547
B	1927	35226	570	720	676
C	200	100	2313	573	588
D	812	384	2113	6831	2603
E	1165	429	2443	2648	11566
	DB2 (50:50)				
A	23729	2166	1138	823	1596
B	4027	89214	1401	1810	1344
C	427	270	5444	2185	1108
D	1925	1151	5590	17199	5992
E	2919	1232	6265	8400	26811
	DB3 (20:80)				
A	36552	5135	1682	1395	2359
B	6104	143531	2018	2973	1847
C	708	557	9132	2671	2026
D	3364	2183	10957	24372	10095
E	6186	2587	11127	12613	40490

Table 4: Multi-frame ($V = 100$) confusion matrices, for the different DB1, DB2, and DB3 databases, obtained by considering 12 DKLT components.

	Recognized				
Input	A	B	C	D	E
	DB1 (80:20)				
A	117	0	0	0	0
B	0	391	0	0	0
C	0	0	36	0	1
D	0	0	1	125	1
E	0	0	0	0	182
	DB2 (50:50)				
A	294	0	0	0	0
B	0	977	0	0	0
C	1	0	89	4	0
D	3	0	12	301	2
E	4	0	1	4	447
	DB3 (20:80)				
A	471	0	0	0	0
B	0	1564	0	0	0
C	0	0	148	1	1
D	11	0	31	436	31
E	42	0	10	20	658

of the confusion matrix), FN the false negatives (the sum of the other elements on the same row of the confusion matrix), FP the false positives (the sum of the other elements on the same column of the confusion matrix), and TN the true negatives (the sum of the elements on the other rows and columns of the confusion matrix). Additionally we considered the *overall sensitivity*, also named *correct identification rate* (CIR) by some authors, defined as the ratio of the sum of the diagonal elements (true positives) respect to the sum of all the elements of the confusion matrix.

The results for 12 DKLT components are reported in Table 5 for the multi-frame ($V = 100$) speaker identification. Also in this case the effect of the database consistency has been investigated. The overall sensitivity obtained in the single frame identification is of 76.83%, 75.83%, and 74.15% for DB1, DB2, and DB3 databases, respectively. Significantly greater values have been obtained in the multi-frame (sequence of $V = 100$ consecutive frames) case i.e. 99.65%, 98.55%, and 95.71% for DB1, DB2, and DB3 databases, respectively.

To show the effect of the sequence length on the speaker identification, Figs. 2 and 3 depict the sensitivity as a function of the number V of the frames for two different numbers of DKLT components $M =$ 20 and 15, respectively, retained in the GMM model, using the DB1 database.

4.3 Comparison with MFCC Model

To investigate the relative performance of our method with the state of the art, we conducted for comparison some experiments using MFCC features. In this case, 13 MFCC features have been considered and the performance for all the databases has been reported in Table 6 where sequences of 100 frames have been considered for identification purposes. Additionally, the overall sensitivity obtained in this case is of 93.33%, 94.81%, and 93.52% for DB1, DB2, and DB3 databases, respectively. Comparing these results with those of our method with 12 DKLT components and 100 frames, it is evident that our method behaves better than the MFCC-based one.

In particular, with reference to a sequence of $V =$

Table 5: Truncated DKLT performance analysis for the different databases ($V = 100$ frames, 12 DKLT components).

Speaker	Sens. (%)	Spec. (%)	Prec. (%)	Acc. (%)
	DB1 (80:20)			
A	100.00	100.00	100.00	100.00
B	100.00	100.00	100.00	100.00
C	97.30	99.88	97.30	99.77
D	98.43	100.00	100.00	99.77
E	100.00	99.70	98.91	99.77
	DB2 (50:50)			
A	100.00	99.57	97.35	99.63
B	100.00	100.00	100.00	100.00
C	94.68	99.36	87.25	99.16
D	94.65	99.56	97.41	98.83
E	98.03	99.88	99.55	99.49
	DB3 (20:80)			
A	100.00	98.21	89.89	98.45
B	100.00	100.00	100.00	100.00
C	98.67	98.75	78.31	98.74
D	85.66	99.28	95.40	97.25
E	90.14	98.81	95.36	96.96

Table 6: MFCC performance analysis for the different databases ($V = 100$ frames).

Speaker	Sens. (%)	Spec. (%)	Prec. (%)	Acc. (%)
	DB1 (80:20)			
A	100.00	96.34	81.25	96.84
B	93.35	99.78	99.73	96.84
C	94.59	98.65	76.09	98.48
D	94.49	98.35	90.91	97.78
E	87.91	99.11	96.39	96.72
	DB2 (50:50)			
A	100.00	97.62	86.98	97.94
B	96.72	99.83	99.79	98.41
C	85.11	99.56	89.89	98.92
D	96.23	97.69	87.93	97.48
E	88.38	99.17	96.64	96.87
	DB3 (20:80)			
A	99.36	98.68	92.31	98.77
B	98.02	99.57	99.48	98.86
C	84.67	98.78	76.05	98.16
D	92.14	96.54	82.28	95.88
E	82.88	98.74	94.68	95.36

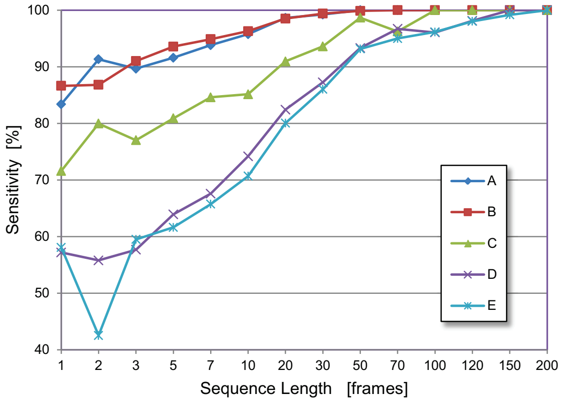


Figure 2: Classifier performance as a function of sequence length, with 20 DKLT components, using DB1 database.

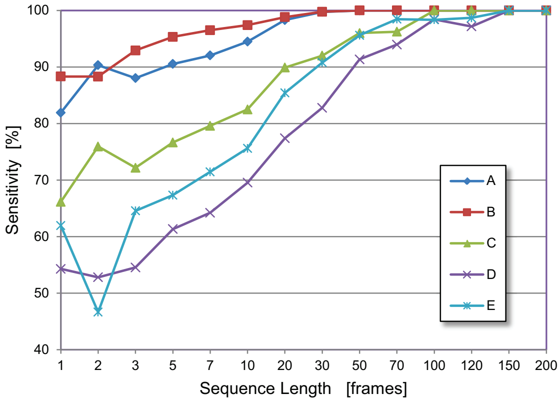
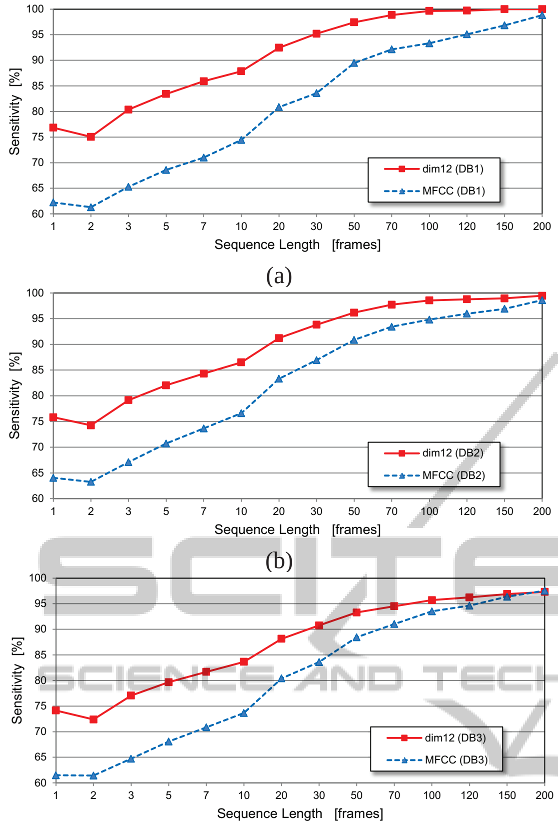


Figure 3: Classifier performance as a function of sequence length, with 15 DKLT components, using DB1 database.

100 frames, Tables 5 and 6 clearly show that all the performance indices for truncated DKLT are better than those for MFCC-based classifier. Similar results are obtained by varying the sequence length, as Fig. 4 points-out.

In order to better compare the two methods, several additional experiments were carried out. Fig. 4 reports, for a more intuitive comparison, the overall sensitivity as a function of speech sequence and database consistency. As you can see, and in particular for short sequences, the truncated DKLT behaves always better than the MFCC-based counterpart.



(a)

(b)

(c)

Figure 4: Overall sensitivity of MFCC features and truncated DKLT for (a) DB1, (b) DB2, and (c) DB3 databases.

5 CONCLUSION

In this paper we have proposed a new speaker identification approach based on truncated DKLT representation, that behaves better than conventional MFCC-based methods. This is motivated by the fact that although MFCCs have demonstrated particularly suitable for speech recognition, they present some drawbacks for speaker recognition.

Several experimental results show that with short sequences of speech frames, that is with utterance duration of less than 1 s, the performance of truncated DKLT are always better than MFCC.

REFERENCES

Bhardwaj, S., Srivastava, S., Hanmandlu, M., and Gupta, J. R. P. (2013). GFM-based methods for speaker identification. *IEEE Trans. Cybernetics*, 43(3):1047–1058.

Bimbot, F. et al. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 2004:430–451.

Campbell, J. P., J. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462.

Figueiredo, M. A. F. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(3):381–396.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Gish, H. and Schmidt, M. (1994). Text-independent speaker identification. *IEEE Signal Processing Magazine*, 11(4):18–32.

Jain, A. K., Duin, R. P. W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(1):4–37.

Jain, A. K., Ross, A., and Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Trans. Circuits and Systems for Video Technology*, 14(1):4–20.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12 – 40.

Maina, C. W. and Walsh, J. M. (2011). Joint speech enhancement and speaker identification using approximate Bayesian inference. *IEEE Trans. Audio, Speech, and Language Processing*, 19(6):1517–1529.

McLaughlin, N., Ming, J., and Crookes, D. (2013). Robust multimodal person identification with limited training data. *IEEE Trans. Human-Machine Systems*, 43(2):214–224.

Patra, S. and Acharya, S. K. (2011). Dimension reduction of feature vectors using WPCA for robust speaker identification system. In *2011 Int. Conf. Recent Trends in Information Technology (ICRTIT)*, pages 28–32.

Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In *2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages IV–4072–IV–4075.

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

Sadjadi, S. O. and Hansen, J. H. L. (2014). Blind spectral weighting for robust speaker identification under reverberation mismatch. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 22(5):937–945.

Therrien, C. W. (1992). *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Togneri, R. and Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, 11(2):23–61.

Zhao, X., Shao, Y., and Wang, D. (2012). CASA-based robust speaker identification. *IEEE Trans. Audio, Speech, and Language Processing*, 20(5):1608–1616.

Zhao, X., Wang, Y., and Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 22(4):836–845.

Zilca, R. D., Kingsbury, B., Navratil, J., and Ramaswamy, G. N. (2006). Pseudo pitch synchronous analysis of speech with applications to speaker recognition. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(2):467–478.