
EURECOM
Department of Security
Campus SophiaTech
CS 50193
06904 Sophia Antipolis cedex
FRANCE

Research Report RR-16-322

# Further Optimisations of Constant Q Cepstral Processing for Integrated Utterance Verification and Text-Dependent Speaker Verification

Submitted to the IEEE Workshop on Spoken Language Technology (SLT) 2016
on July 22nd 2016

Héctor Delgado[1], Massimiliano Todisco[1], Md Sahidullah[2], Achintya K. Sarkar[3],
*Nicholas Evans*[1], *Tomi Kinnunen*[2], *Zheng-Hua Tan*[3]

[1]Department of Digital Security, EURECOM, France
[2]Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland
[3]Signal and Information Processing, Department of Electronic Systems, Aalborg University, Denmark

July 21st, 2016
Last update August 23rd, 2016

Tel : (+33) 4 93 00 81 00
Fax : (+33) 4 93 00 82 00
Email : {delgado,todisco,evans}@eurecom.fr

[0]EURECOM's research is partially supported by its industrial members: BMW Group Research and Technology, IABG, Monaco Telecom, Orange, Principauté de Monaco, SAP, ST Microelectronics, Symantec.

# Further Optimisations of Constant Q Cepstral Processing for Integrated Utterance Verification and Text-Dependent Speaker Verification

Héctor Delgado[1], Massimiliano Todisco[1], Md Sahidullah[2], Achintya K. Sarkar[3], *Nicholas Evans[1], Tomi Kinnunen[2], Zheng-Hua Tan[3]*

[1]Department of Digital Security, EURECOM, France
[2]Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland
[3]Signal and Information Processing, Department of Electronic Systems, Aalborg University, Denmark

**Abstract**

Many authentication applications involving automatic speaker verification (ASV) demand robust performance using short-duration, fixed or prompted text utterances. Text constraints not only reduce the phone-mismatch between enrolment and test utterances, which generally leads to improved performance, but also provide an ancillary level of security. This can take the form of explicit utterance verification (UV). An integrated UV + ASV system should then verify access attempts which contain not just the expected speaker, but also the expected text content. This paper presents such a system and introduces new features which are used for both UV and ASV tasks. Based upon multi-resolution, spectro-temporal analysis and when fused with more traditional parameterisations, the new features not only generally outperform Mel-frequency cepstral coefficients, but also are shown to be complementary when fusing systems at score level. Finally, the joint operation of UV and ASV greatly decreases false acceptances for unmatched text trials.

**Index Terms**

speaker verification, utterance verification, text dependent, constant Q transform.

# Contents

# List of Figures

# 1 Introduction

Automatic speaker verification (ASV) [1] technology has matured over recent years to become a low-cost and reliable approach to person recognition. Example applications include smart-phone log-in, telephone banking, logical and physical access control [2]. In these and indeed in any other scenarios, both user convenience and reliability are usually dependent on text constraints.

At one end of the spectrum of possible text constraints is text-independent ASV. Here, both enrolment and testing are performed with free-text utterances. In some sense, this approach is the most convenient, but the use of free text usually requires long-duration utterances in order to marginalise mis-matching text content and thus to ensure reliable performance.

At the other end of the spectrum is text-dependent ASV. This implies the use of the same fixed-text phrase for both enrolment and test. The use of fixed-text phrases may be less convenient but usually provides for better ASV performance with short utterances on account of matching text content.

Text-dependent ASV can be addressed with an elementary ASV system, such as a Gaussian mixture model system with a universal background model (GMM-UBM) [3] or an i-vector system with probabilistic linear discriminant analysis (PLDA) [4]. These systems on their own capture only implicitly the time sequence information of the text content. Other approaches, such as those based on hidden Markov models [HMM] [5], can capture this content explicitly but, being usually more complex, typically require more data to train.

With user convenience being often a priority, alternative approaches to verify the text content of short spoken utterances have been investigated. This approach is referred to as utterance verification (UV). UV is the task of determining whether or not a given utterance corresponds to a given text. The combination of ASV and UV systems can then verify both the claimed speaker identity and text content of a given utterance. Some works have addressed the tasks of UV and ASV jointly by combining separate systems [6–8]. In [9] a number of different UV and ASV strategies and their combination are compared using the RedDots database [10] and specially designed protocols.

Recently, features based on the constant Q transform (CQT) [11] have been successfully applied to a number of speech-related applications, including ASV [12, 13]. In these features, CQT is used to obtain variable-resolution spectra which provide a greater frequency resolution at low frequencies and a greater time resolution at high frequencies. However, the frequency scale of such spectra is geometric. This poses difficulties when it is coupled with traditional cepstral analysis, where some post-processing is usually required to yield a linear frequency scale [12]. This multi-resolution analysis together with further post-processing may impose a high computational load.

This work proposes to replace the CQT algorithm in [11] with the infinite impulse response constant Q transform (IIR-CQT) proposed in [14] as a more efficient alternative. It delivers multi-resolution time-frequency analysis in a linear

scale spectrum which is ready to be coupled with traditional mel-cepstral analysis. The resulting features of combining IIR-CQT and cepstral analysis are called *infinite impulse response - constant Q, Mel-frequency cepstral coefficients* (ICMC).

This paper reports the authors' subsequent work on UV and text-dependent ASV [9] with the new ICMC features to fully expose the potential. Specifically, the contributions are as follows:

- **new features for UV and ASV** – the paper introduces ICMC features which are used to improve the performance of both ASV and UV systems;
- **UV optimisation** – the paper presents an assessment of UV performance using an HMM-UBM system and the dependence of performance on its configuration;
- **ASV optimisation** – the paper presents an assessment of GMM-UBM, HMM-UBM and i-vector approaches for short utterance, text-dependent ASV, and
- **stand-alone and combined assessment** – UV and ASV systems are assessed independently and when combined with a decision-based fusion in order to determine an optimal operating point.

# 2 Infinite Impulse response - constant Q mel-frequency cepstral coefficients

Recent research [12, 13] has shown that better performance for a range of speaker modelling and classification tasks can be achieved by replacing the traditional short-time Fourier transform (STFT) with an alternative approach to time-frequency analysis known as the constant Q transform (CQT) [11]. These findings provided the stimulus behind its application to UV and ASV. Starting with a treatment of the limitations of the STFT, we present the specific approach as follows.

## 2.1 Short-time Fourier transform

The classical STFT spectrogram is a visual representation of the spectro-temporal composition of a signal through regularly spaced time intervals and frequency bands. Different signals, such as speech, music or noise, give rise to different spectro-temporal structure.

As a result, spectro-temporal analysis requires a resolution adapted to the signal in question. For example, a higher frequency resolution may be prefered for the analysis of low-frequency content of voiced speech signals where the harmonic density is typically high. Conversely, a higher time resolution may be required to capture rapid modulation at high frequencies. As a consequence of competing requirements, multi-resolution spectro-temporal representations are appealing for the analysis of speech signals.

## 2.2 Constant Q transform

There are a number of alternatives to the constant resolution of the STFT [15–17]. Rather than a constant resolution, some of these alternatives offer instead a constant Q factor. The Q factor is a measure of the selectivity of each filter and is defined as the ratio between the center frequency $f_k$ and the bandwidth $\delta f$:

$$Q = \frac{f_k}{\delta f} \tag{1}$$

The human perception system is known to approximate a constant Q factor between 500Hz and 20kHz [18]. This is the main motivation for the constant Q analysis of audio signals [19–21].

The constant Q transform (CQT) was introduced in 1978 by Youngberg and Boll [17] and refined some years later in 1991 by Brown [11]. The CQT employs geometrically distributed octaves and center frequencies. As a result, the CQT provides for a higher frequency resolution at low frequencies and, conversely, a higher temporal resolution for high frequencies.

Multi-resolution processing, however, carries a penalty in computation time. In addition, the use of a geometric frequency scale can necessitate still more processing to linearise the scale for decorrelation and modelling purposes [12, 13].

The infinite impulse response-CQT (IIR-CQT) algorithm proposed in [14] provides a compromise between computational cost and design flexibility. The authors of [14] propose a direct method to approximate a time-varying IIR (TV IIR) filterbank to accomplish the constant Q behavior in which the pole varies with frequency ($p = p[n]$)

$$Y(k) = X(k) + X(k+1) + p(k)Y(k-1) \tag{2}$$

where $X$ is the discrete Fourier Transform of the signal computed after centering the signal at time 0. Finally, a forward-backward TV IIR filtering is performed to obtain zero-phase distortion.

The location of the pole varies for each frequency band along the real axis in order to obtain different time window widths (wider for low frequencies and narrower for high frequencies). Further details and the algorithm implementation can be found in [14].

Unlike the CQT algorithm described in [11], frequency scale of spectrum derived by the IIR-CQT algorithm is linear. This allows the direct coupling with traditional cepstral analysis without further post-processing [12].

Figure 1 shows the difference between STFT (top) and IIR-CQT (bottom) derived spectrograms for an arbitrary utterance from the RedDots database [10]. As a result of multi-resolution analysis, harmonics at lower frequencies are better defined in the IIR-CQT-derived spectrogram than in the STFT-derived spectrogram. In addition, time resolution is improved at higher frequencies.

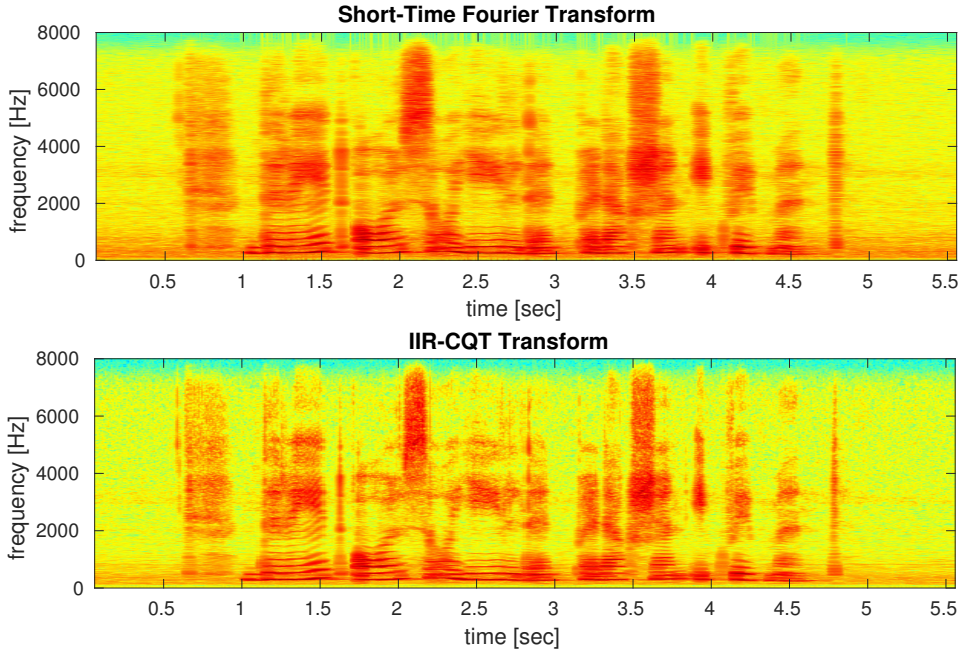


Figure 1: *Spectrograms of the utterance 'the rate also yielded production equipment' for an arbitrary speaker in the RedDots database [10]. Spectrograms computed with the STFT (top) and with the IIR-CQT (bottom).*

### 2.3 Mel-cepstral analysis

As is the case for traditional STFT derived spectro-temporal estimates, cepstral processing can be applied to individual spectral magnitude frame estimates derived with the IIR-CQT. The cepstrum of a time sequence $x(n)$ is obtained from the inverse transformation of the logarithm of the spectrum.

The inverse transformation is normally implemented with the discrete cosine transform (DCT). The cepstrum is then a (usually truncated) orthogonal decomposition of the log spectrum. It maps $N$ Fourier coefficients onto $q \ll N$ independent cepstrum coefficients which capture the most significant and relevant information contained within the spectrum.

Based upon auditory critical bands [22], Mel-scaling is normally applied prior to cepstral analysis. Mel-scaling is commonly employed in a range of speech processing taks and is typically extracted according to:

$$MFCC(q) = \sum_{m=1}^{M} \log\left[MF(m)\right] \cos\left[\frac{q\left(m-\frac{1}{2}\right)\pi}{M}\right] \tag{3}$$

where the Mel-frequency spectrum is defined as

$$MF(m) = \sum_{k=1}^{K} \left|X(k)\right|^2 H_m\left(k\right) \tag{4}$$

where $k$ is the DFT index, $H_m(k)$ is the triangular-shaped weight function for the $m$-th Mel-scaled bandpass filter. Normally, the number of coefficients $q$ is less than the number of Mel-filters $M$. Typically, $M = 25$ and $q$ varies between 13 and 20.

This paper investigates the combination of the IIR-CQT with Mel-scaling and cepstral analysis. This is achieved by replacing $X(k)$ in Equation 4 with $Y(k)$ from Equation 2. The resulting features are referred to as **I**nfinite impulse response **C**onstant Q **M**el-frequency **C**esptral coefficients (ICMC).

# 3 UV system

A number of different approaches to UV were reported in [9], including GMM-UBM, HMM-UBM, dynamic time warping and a forced alignment system. The HMM-UBM system was found to outperform the alternatives and was thus adopted for all work repoted here.

The HMM-UBM system reported in [9] is a 2-layer model similar in nature to the so-called HiLam approach to text-dependent ASV introduced in [5]. The model is a left-to-right, utterance-dependent HMM with continuous observation densities modeled with GMMs adapted from an utterance-independent UBM pre-trained with external data.

Utterances are first split into $N$ equal-sized segments, where $N$ is the number of HMM states. Each state is a GMM and is estimated by adapting the UBM to the corresponding utterance segment using maximum *a posteriori* (MAP) adaptation. A number of Viterbi realignment and readaptation sequences are then applied to optmise the model.

UV scores are the likelihood ratio given the data and either the utterance-dependent HMM or the utterance-independent UBM. The number of HMM states and the number of Gaussian mixtures per state are empirically optimised and are the same for each utterance.

In UV experiments different numbers of GMM components of 8, 16, 32 and 64 were evaluated. In addition, different number of HMM states of 14, 24 and 34 were assessed. The UBM was trained on male speech from the TIMIT database.

Two different score normalisation approaches are also investigated. Mean-Norm subtracts from the utterance score the mean score produced by all alternative utterance models. MaxNorm subtracts from the utterance score the maximum score produced by all the alternative utterance models. Note that these normalization techniques can only be applied in a practical scenario if the universe of pass-phrases is limited (10 in this case).

# 4 ASV systems

This section describes the three ASV systems used for the experimental work reported in this paper.

### 4.1 GMM-UBM

The GMM-UBM system is the de facto standard approach to ASV. The UBM represents the speaker-independent acoustic space [3] and is trained with an expectation maximisation algorithm on a large quantity of external data. Speaker-specific models are then learned from the UBM using MAP adaptation. Only the UBM means are adapted. The UBM was trained using speech data from the TIMIT corpus[1] and all models have $512$ components.

### 4.2 HMM-UBM

The HMM-UBM system is described in [23]. A universal, text and speaker-independent HMM [24] is learned with the data of 157 speakers from the RSR2015 database (approximately 30 phrases/speaker over 9 sessions) without any speech transcriptions and with several iterations of the Baum Welch algorithm. Speech transcriptions are not utilized for HMM training, thus model parameters reflect general temporal information only.

Speaker dependent models are derived from the HMM-UBM using enrollment data with MAP adaptation [25]. Three MAP iterations are used with a relevance factor of $10$ and only Gaussian mean parameters are adapted. The number of HMM-UBM states and Gaussian components per state are optimized to minimize the equal error rate (EER) on the impostor-correct condition (see Section 5.1) of the development set. Test utterance scores are obtained from their forced alignment to the claimed target model and the universal HMM-UBM and then the corresponding log-likelihood ratio.

### 4.3 i-vector

The i-vector system is based on original work in [4].

i-vectors are extracted using a GMM-UBM of 512 components with diagonal co-variance matrices which are learned using the same data as that used to learn the universal HMM-UBM as described above.

Each target is represented by an average i-vector computed over the phrase-wise i-vectors of their enrollment data. Test utterance i-vector are extracted in the same way and then compared to those of the claimed target in the usual way. We consider an i-vector dimension of 400.

Before scoring, i-vectors are post-processed using the iterative conditioning algorithm with spherical normalization (Sph) described in [26] in order to compensate for session variability. The normalisation procedure is trained using the same data as that used for GMM-UBM learning. Scores are then calculated using probabilistic linear discriminant analysis (PLDA) in which Gaussian priors are assumed for speaker and channel factors. Scores between the claimed target ($w_1$)

[1] `https://catalog.ldc.upenn.edu/LDC93S1`

and test ($w_2$) i-vectors are then calculated according to:

$$score(w_1, w_2) = \log \frac{p(w_1, w_2|\theta_{tar})}{p(w_1, w_2|\theta_{non})} \quad (5)$$

where $\theta_{tar}$ defines the hypothesis that i-vectors $w_1$ and $w_2$ are from the same speaker, whereas $\theta_{non}$ represents the alternative hypothesis. For PLDA training, same-speaker utterances are considered to come from different speakers, thereby resulting in the order of 4710 utterances for PLDA learning (157 speakers, 30 pass-phrases/speaker over 9 sessions). For more details about the PLDA and *Sph* algorithm are available in [26–29].

# 5 Experimental setup

This section describes the experimental setup, including metrics, databases, protocols and feature extraction.

## 5.1 Metrics and evaluation

UV and ASV performance are assessed in terms of the EER. In contrast, and in order to illustrate more clearly difference in performance with and without UV, combined performance is expressed in terms of false acceptance rate (FAR) and false rejection rate (FRR). As illustrated in Table 2, in addition to the one target correct (TC) condition in which both the utterance and speaker labels match, there are three types of impostor trial where either the utterance or speaker do not match. They are the target wrong (TW), impostor correct (IC) and impostor wrong (IW) [10] conditions. Accordingly, FAR performance is furthermore illustrated independently for each impostor trial, namely FAR(TW), FAR(IC) and FAR(IW). The operation point for ASV is when FRR and FAR(IC) are equal. We selected this operation point to tune the system to give balanced performance when the text content matches. Then, FAR(TW) and FAR(IW) are expected to be lowered by the joint operation of the UV module. Lastly, performance is evaluated for UV and ASV systems in isolation and when combined. Combination is achieved through score level fusion by means of logistic regression and is performed with the BOSARIS toolkit[2].

## 5.2 Database and protocols

Experiments are conducted with speech data collected in connection with the RedDots challenge[3] [10]. Since the challenge relates exclusively to ASV, new protocols are created to support UV and ASV experiments. Due to the limited number of female subjects in the RedDots corpus, only male speakers are included

[2] https://sites.google.com/site/bosaristoolkit/

[3] https://sites.google.com/site/thereddotsproject/home

Table 1: *Database description for UV experiments.*

| | Development | Evaluation |
|---|---|---|
| Test Utterances | 1049 | 1536 |
| Matched-Text trials | 1049 | 1536 |
| Unmatched-Text trial | 9441 | 13824 |

Table 2: *Database description for ASV experiments.*

| | Development | Evaluation |
|---|---|---|
| Number of Targets | 96 | 152 |
| Target Correct (TC) | 1011 | 1108 |
| Target Wrong (TW) | 9099 | 9972 |
| Impostor Correct (IC) | 9059 | 22220 |
| Impostor Wrong (IW) | 81535 | 200172 |

in the protocols. They are formed from a subset of part 01 of the evaluation subset which contains utterances of 10 common phrases.

Data from 9 different speakers are used for training utterance models. This results in a total of 1485 utterances used for training (roughly 148 files per phrase). The development set is formed with data from 10 speakers whereas the evaluation set contains data from a different set of 30 speakers. Table 1 gives details of the UV development and evaluation protocols: number of utterances, number of matched-text (target) trials, and number of unmatched-text (nontarget) trials. As regards ASV, each speaker-and-passphrase dependent model is enrolled with 3 utterances. Table 2 shows details of the ASV development and evaluation protocols: number of target speakers-passphrase models, number of target trials (target-correct), and number of nontarget trials (target-wrong, impostor-correct and impostor wrong).

Note that, while part 04 of the evaluation, namely the text-prompted condition, may at first seem better suited to the development and assessment of UV systems, it relates to the verification of speaker-sentence pairs. As such, it is not suited to both the independent and combined assessment of UV and ASV, hence the approach adopted here.

### 5.3 Feature extraction

Both UV and ASV experiments are performed independently and when combined using two feature extraction methods. Mel-frequency cepstral coefficients (MFCC) serve as the baseline for comparisons with performance when using the new ICMC features. Except for differences in the underlying approach to spectro-temporal analysis (STFT for MFCC versus IIR-CQT for ICMC), the two configurations share an identical configuration.

The common processing is as follows. Pre-emphasised speech signals are frame-blocked using a sliding window of 20 ms with a 10 ms shift. The power spectrum is obtained using either the STFT or the IIR-CQT from Hamming win-

dowed frames before 19th order static coefficients (excluding the 0-th coefficient) are extracted using the discrete cosine transform (DCT) of 20 log-power, Mel-scaled filterbank outputs. For IIR-CQT, a Q factor of 96 was empirically determined.

RASTA filtering is then applied before delta and delta-delta coefficients are computed from the static parameters thereby resulting in feature vectors of dimension 57. Speech activity detection (SAD) based on energy modelling is applied to discard low-energy content. Finally, cepstral mean and variance normalization are applied to compensate for channel variation.

### 5.4 Integration of UV and ASV

In this scenario, we simultaneously verify the spoken content as well as the speaker identity and accept the claim only if both are correct. In this paper, the UV and ASV system are combined in two different methods. In the first strategy, score level fusion is performed on the scores obtained from two systems. The fusion is performed using linear regression method using BOSARIS toolkit where the trials from TC condition are used as target and the rests as non-target. In the second integration method, decision level fusion is performed on the binary decision (i.e., accept/reject) available from the UV and ASV systems. For this, the decision thresholds are computed first separately on two individual tasks to produce the binary decision labels. Then the decision labels are combined by 'AND' operation.

## 6 Experimental results

Performance is first assessed for UV and ASV in independence and then when combined.

### 6.1 Standalone UV

UV results for the development set are illustrated in Table 3 in terms of EER for both MFCC and ICMC features. Results are shown for different numbers of HMM states and GMM components. Different rows show performance with and without normalisation and for two fusion strategies.

For smaller GMMs (8 and 16 components), ICMC features outperform MFCC, while MFCC gives better performance for more complex models (32 and 64 components). Without normalisation, ICMC features generally outperform MFCC, while MFCC features mostly outperform ICMC features for MeanNorm and MaxNorm, with the latter providing the best results. The best performance achieved with MFCC features is 0.20% EER with 14 HMM states and 32 Gaussian components per state. The best performance for ICMC features is 0.45% EER with 24 HMM states and 32 Gaussian components per state. In these optimum configurations, relative improvements of 93% and 83% are achieved for MFCC and ICMC features,

Table 3: *Utterance verification performance using the standalone UV protocol, measured in EER (%) on the development set for different number of HMM states (14, 24 and 34) and different number of GMM components (8, 16, 32 and 64).*

| | | 14 HMM states | | | | 24 HMM states | | | | 34 HMM states | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Norm. method | Feature | 8 | 16 | 32 | 64 | 8 | 16 | 32 | 64 | 8 | 16 | 32 | 64 |
| None | MFCC | 5.09 | 4.20 | 3.19 | 3.25 | 4.01 | 3.36 | 2.63 | 2.70 | 3.47 | 2.92 | **2.11** | 2.26 |
| | ICMC | 3.77 | 3.48 | 2.97 | 3.38 | 3.35 | 2.86 | **2.67** | 2.94 | 3.04 | 2.83 | 2.81 | 2.83 |
| Mean | MFCC | 1.56 | 1.28 | 0.94 | 0.96 | 1.31 | 1.25 | 1.07 | 0.94 | 1.22 | 1.27 | 1.04 | **0.90** |
| | ICMC | 1.48 | 1.48 | 1.51 | 1.39 | 1.32 | 1.25 | 1.22 | **1.17** | 1.30 | 1.27 | 1.27 | 1.28 |
| Max | MFCC | 0.59 | 0.36 | **0.20** | 0.35 | 0.38 | 0.35 | 0.37 | 0.41 | 0.40 | 0.39 | 0.48 | 0.48 |
| | ICMC | 0.66 | 0.63 | 0.63 | 0.58 | 0.58 | 0.54 | **0.45** | 0.48 | 0.48 | 0.56 | 0.51 | 0.46 |
| Max | Fused | 0.34 | 0.33 | **0.19** | 0.25 | 0.28 | 0.24 | 0.27 | 0.31 | 0.26 | 0.35 | 0.28 | 0.32 |
| Max | Best MFCC and ICMC fused | **0.19** | | | | | | | | | | | |

Table 4: *Utterance verification results using the standalone UV protocol, measured in EER (%), on evaluation set. Best classifier configurations based on the results on development set are selected.*

| System | noNorm | meanNorm | maxNorm |
|---|---|---|---|
| MFCC | 4.38 | 2.43 | 0.71 |
| ICMC | 4.16 | 2.09 | 0.78 |
| Fusion | 3.42 | 1.87 | 0.51 |

Table 5: *ASV performance of i-vector system with MFCC and ICMC features, in terms of EER (%), for IC condition, on the development set for different speaker factors in PLDA (channel factor is kept full rank i.e. equal to the dimension of i-vector).*

| | Speaker factors | | | | |
|---|---|---|---|---|---|
| Feature | 200 | 250 | 300 | 350 | 400 |
| MFCC | 5.60 | 5.30 | 4.99 | 5.07 | **4.98** |
| ICMC | 6.31 | 5.69 | 5.33 | 5.27 | **5.21** |

with respect to un-normalized scores. The fusion of max-normalised MFCC and ICMC scores delivers an EER of 0.19%. Finally, the fusion of the two best MFCC and ICMC max-normalised scores obtains the same performance.

Results for the evaluation set using the two best configurations for MFCC and ICMC features are reported in Table 4. Here the trend is reversed, with ICMC feature producing slightly better performance than MFCC. Once again, MeanNorm and MaxNorm are effective. The best EER of 0.51% is achieved with the fusion of max-normalised MFCC and ICMC scores. Taking results for the evaluation set as a whole, MFCC and ICMC features are consistently complementary.

## 6.2 Standalone ASV

Fig. 2 illustrates the performance of the **GMM-UBM system** against relevance factor for the TW, IC and IW conditions of the development set. For the IC condition, optimal performance is obtained with relevance factors of 4 and 2 for MFCC and ICMC features respectively, corresponding to EERs of 3.19% and 2.26%. Furthermore ICMC features are shown to universally outperform MFCC features.

Results for the **HMM-UBM system** are illustrated in Figure 3 for the IC condition on the development set and for MFCC and ICMC features, using different number of HMM states and GMM components for a fixed relevance factor of 10. Performance varies in the intervals $5.28 \pm 3.18\%$ (mean$\pm$variance of EER) and $4.51\% \pm 1.32\%$ for MFCC and ICMC, respectively. Best performance is obtained in the case of MFCC features with 4 states and 64 components and with 14 states and 32 components in the case of ICMC features.

In the **i-vector system**, channel factors are kept full rank (i.e. equal to the dimension of i-vector) and the value of speaker factor is varied to find the optimal

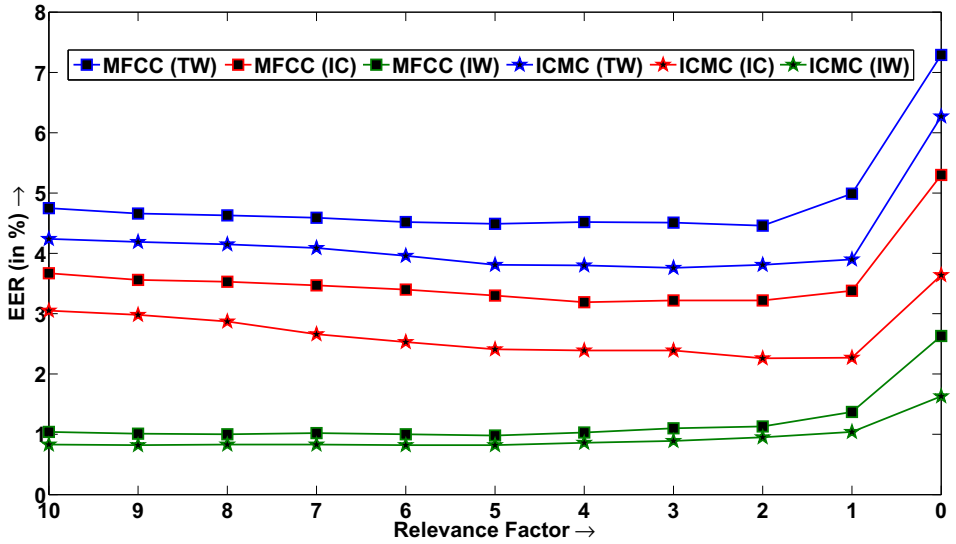


Figure 2: *ASV performance of GMM-UBM system with MFCC and ICMC features, in terms of EER (%), for TW, IC and IW conditions on the development set for various values of the relevance factor.*

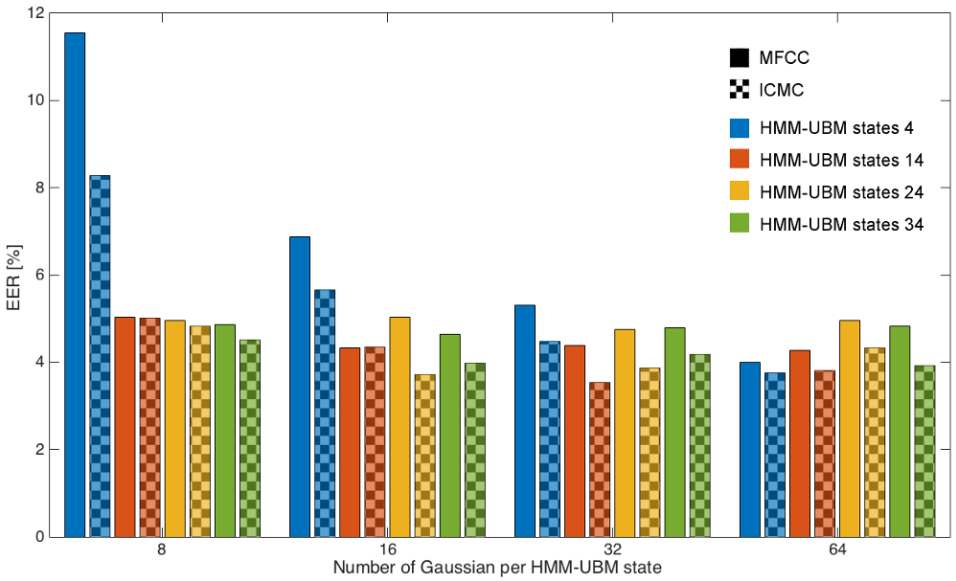


Figure 3: *ASV performance of HMM-UBM system with MFCC and ICMC features, in terms of EER (%), for IC condition, on the development set for various number of states and Gaussian components in HMM-UBM.*

Table 6: *Text-dependent speaker recognition performance (in terms of EER %) for different tasks on development and evaluation set using different ASV systems and fusion.*

| | Development | Evaluation |
|---|---|---|
| GMM-UBM (MFCC) | 3.19 | 2.14 |
| GMM-UBM (ICMC) | 2.26 | 1.56 |
| HMM-UBM (MFCC) | 3.99 | 2.26 |
| HMM-UBM (ICMC) | 3.54 | 1.67 |
| i-vector (MFCC) | 4.98 | 2.78 |
| i-vector (ICMC) | 5.21 | 3.83 |
| Fusion | **1.68** | **1.03** |

speaker verification performance on the development set (for IC condition in terms of lowest EER value) as presented in Table 5. EER decreases with an increasing value of speaker factor on both features. Best performance is achieved when both number of channel and speaker factors are equal to the full dimension of i-vector.

A comparative summary of ASV performance for each of the three independent systems and for their **score-level fusion** is presented in Table 6 for the IC condition of both development and evaluation sets. Most likely due to the short-duration nature of the RedDots database [10], the simplest GMM-UBM system is the best performing. With the exception of the i-vector system, results for ICMC features are better than those for MFCC features, for both development and evaluation sets. The reason of the inverse performance trend of MFCC and ICMC on i-vector has to be further investigated.

Results for fused ASV systems are illustrated in the last row of Table 6. Fusion results stem from the combination of scores produced by each of the three systems and with each of the two different features configurations (six systems) using logistic regressions. In contrast to previous work in [9], fusion weights are optimised with TC trial scores used as positives and IC trial scores used as negatives. Fusion results in the lowest EERs for both development and evaluation sets.

### 6.3 Effect of combined UV in text-dependent ASV performance

First, results for ASV in isolation, in terms of FAR and FAR for the selected operation points (using IC trials as impostors), are illustrated in Table 7. Then, two different UV + ASV integration strategies are illustrated in the final two rows. The penultimate row shows results for score fusion, whereas the last row illustrates results for decision fusion.

ASV system in fusion in isolation does not help in reducing the FAR for the TW condition. This is not unexpected since, without UV, ASV on its own offers little potential to reject incorrect pass-phrases. However, when UV and ASV are combined, FAR(TW) is greatly decreased for the two proposed combination schemes (from 2.85% to 0.79% and 0.00% in the development set for the two combinations, respectively). Nevertheless, this has the cost of increasing FRR slightly. Decision

Table 7: *Performance with joint-protocol on development and evaluation sets using different ASV systems.*

| | Development | | | | Evaluation | | | |
|---|---|---|---|---|---|---|---|---|
| | FRR | FAR (TW) | FAR (IC) | FAR (IW) | FRR | FAR (TW) | FAR (IC) | FAR (IW) |
| ASV GMM-UBM (MFCC) | 3.26 | 6.73 | 3.25 | 0.09 | 2.08 | 8.41 | 2.57 | 0.03 |
| ASV GMM-UBM (ICMC) | 2.37 | 6.91 | 2.34 | 0.04 | 1.26 | 7.38 | 2.30 | 0.05 |
| ASV HMM-UBM (MFCC) | 4.06 | 5.87 | 4.05 | 0.07 | 1.26 | 11.78 | 4.92 | 0.13 |
| ASV HMM-UBM (ICMC) | 3.56 | **1.21** | 3.58 | 0.03 | 0.99 | **3.97** | 4.31 | 0.04 |
| ASV i-vector (MFCC) | 4.95 | 7.13 | 5.03 | 0.10 | 1.53 | 14.62 | 6.97 | 0.18 |
| ASV i-vector (ICMC) | 5.34 | 3.40 | 5.35 | 0.09 | 3.52 | 5.97 | 4.61 | 0.08 |
| ASV Fusion | **1.78** | 2.85 | **1.74** | **0.00** | **0.72** | 4.94 | **2.24** | **0.01** |
| Score Fusion of UV (Fused) + ASV (Fused) | **2.18** | 0.79 | 2.17 | **0.00** | **0.72** | 1.22 | 2.59 | **0.00** |
| Decision Fusion of UV (Fused) + ASV (Fused) | 2.28 | **0.00** | **1.73** | **0.00** | 1.35 | **0.03** | **2.22** | **0.00** |

fusion outperforms score level fusion, which further degrades FRR and FAR(IC). This is expected since, in score fusion, UV scores are raising the overall ASV score, and therefore increasing FAR when text matches. Results nonetheless indicate that both combined approaches lead to considerably lower FARs. In the evaluation set, similar systems' behavior is found. Compared to our prior work in [9], errors related to matched text trials (TC and IC) are significantly lower, while keeping unmatched text errors (FAR(TW) and FAR(IW)) virtually to 0%.

# 7 Conclusions

This paper has presented a new feature for utterance verification (UV) and automatic speaker verification (ASV). Referred to as infinite impulse response - constant Q Mel-frequency cepstral coefficients (ICMC), the new multi-resolution approach is better adapted than the short-term Fourier transform (STFT) to the spectro-temporal analysis and parameterisation of speech signals. The use of ICMC features improves the performance of a UV system based on spectro-temporal modelling and also the performance of three different approaches to text-dependent ASV. The fusion of UV with the three different ASV systems leads to the best overall performance, decreasing false acceptances related to unmatched text to 0% while just slightly increasing false rejections.

The work demonstrates the potential of UV to improve text-dependent ASV performance. Even so, UV is still a research field in its relative infancy. One can thus readily expect significant developments in the coming years.

# 8 Acknowledgements

The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position on the European Commission.

# References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] K. Lee, B. Ma, and H. Li, "Speaker verification makes its debut in smartphone," *IEEE signal processing society speech and language technical committee newsletter*, February 2013.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[4] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[5] A. Larcher, J.-F. Bonastre, and J. S. D. Mason, "Reinforced temporal structure information for embedded utterance-based speaker recognition," in *Proceedings of INTERSPEECH*, Brisbane, Australia, 2008, pp. 371–374.

[6] Y. Liu, P. Ding, and B. Xu, "Using nonstandard SVM for combination of speaker verification and verbal information verification in speaker authentication system," in *Proceedings of ICASSP*, vol. 1, 2002, pp. I–673–I–676.

[7] L. Rodriguez-Linares and C. Garcia-Mateo, "A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task," in *Proceedings of ICSLP*, vol. 2, 1998, pp. 213–216.

[8] Q. Li, B.-H. Juang, and C.-H. Lee, "Automatic verbal information verification for user authentication," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 585–596, 2000.

[9] T. Kinnunen, M. Sahidullah, I. Kukanov, H. Delgado, M. Todisco, A. Sarkar, N. B. Thomsen, V. Hautamäki, N. Evans, and Z.-H. Tan, "Utterance verification for text-dependent speaker recognition: a comparative assessment using the RedDots corpus," in *Proceedings of INTERSPEECH (to appear)*, 2016.

[10] K. A. Lee, A. Larcher, W. Wang, P. Kenny, N. Brummer, D. A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proceedings of Interspeech*, 2015.

[11] J. Brown, "Calculation of a constant Q spectral transform," *Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, January 1991.

[12] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Speaker Odyssey Workshop*, Bilbao, Spain, 2016.

[13] ——, "Articulation rate filtering of CQCC features for automatic speaker verification," in *INTERSPEECH*, 2016.

[14] P. Cancela, M. Rocamora, and E. López, "An efficient multi-resolution spectral transform for music analysis," in *Proceedings of ISMIR*, 2009, pp. 309–314.

[15] F. C. C. B. Diniz, I. Kothe, S. L. Netto, and L. W. P. Biscainho, "High-selectivity filter banks for spectral analysis of music signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–12, 2007.

[16] K. Dressler, "Sinusoidal extraction using an efficient implementation of a multi-resolution FFT," in *Proceedings of International Conference on Digital Audio Effects (DAFx-06)*, 2006, pp. 247–252.

[17] J. Youngberg and S. Boll, "Constant-Q signal analysis and synthesis," in *Proceedings of ICASSP*, vol. 3, Apr 1978, pp. 375–378.

[18] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. BRILL, 2003.

[19] G. Costantini, R. Perfetti, and M. Todisco, "Event based transcription system for polyphonic piano music," *Signal Processing*, vol. 89, no. 9, pp. 1798–1811, Sep. 2009.

[20] R. Jaiswal, D. Fitzgerald, E. Coyle, and S. Rickard, "Towards shifted nmf for improved monaural separation," in *24th IET Irish Signals and Systems Conference (ISSC 2013)*, June 2013, pp. 1–7.

[21] C. Schorkhuber, A. Klapuri, and A. Sontacch, "Audio pitch shifting using the constant-Q transform," *Journal of the Audio Engineering Society*, vol. 61, no. 7/8, pp. 425–434, July/August 2013.

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[23] A. K. Sarkar and Z.-H. Tan, "Text dependent speaker verification using unsupervised HMM-UBM and temporal GMM-UBM," in *Proceedings of INTERSPEECH (to appear)*, 2016.

[24] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–285, 1989.

[25] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[26] P. M. Bousquet, A. Larcher, D. Matrouf, J. F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Proceedings of Odyssey Speaker and Language Recognition Workshop*, 2012.

[27] S. J. D. Prince, "Computer Vision: Models Learning and Inference," in *Cambridge University Press, 2012, In press.*

[28] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of ICCV*, 2007, pp. 1–8.

[29] M. Senoussaoui, P. Kenny, N. Brümmer, E. de Villiers, and P. Dumouchel, "Mixture of plda models in i-vector space for gender-independent speaker recognition," in *Proceedings of INTERSPEECH*, 2011.