Multimodal Speaker Authentication – Evaluation of Recognition Performance of Watermarked References

C. Vielhauer[1], T. Scheidat[1], A. Lang[1], M. Schott[1], J. Dittmann[1], T.K. Basu[2], P.K. Dutta[2]
[1] *Department of Computer Science, ITI Research Group on Multimedia and Security, Universitätsplatz 2, 39106 Magdeburg, Germany*
{claus.vielhauer, tobias.scheidat, andreas.lang, jana.dittmann}@iti.cs.uni-magdeburg.de
mschott@cs.uni-magdeburg.de
[2] *Indian Institute of Technology, Kharagpur, India*
{tkb|pkd}@ee.iitkgp.ernet.in

Abstract

In this paper a new approach for combining biometric authentication and digital watermarking is presented. A digital audio watermark method is used to embed metadata into the reference data of biometric speaker recognition. Metadata in our context may consist of feature template representations complementary to the speech modality, such as iris codes or biometric hashes, ancillary information about the social, cultural or biological context of the originator of the biometric data or technical details of the sensor.

We suggest a well-known watermark embedding technique based on LSB (least significant bit) modulation for this purpose, perform experiments based on a database taken from 33 subjects and 5 different utterances and a known cepstrum based speaker recognition algorithm in verification mode. The goal is to perform a first evaluation of the recognition precision for our selected algorithm. The first tests show that the recognition precision is not significantly deteriorated by the embedding of the information, as in three out of five cases, no degradation was observed at all and in the worst case the relative increase in false recognition was limited to 1%.

1. Introduction

In the last years the necessity for user authentication rose strongly. The automatic authentication of persons is an important function in many areas of the everyday life (e.g. in e-commerce applications). Biometric authentication becomes more and more important besides or in combination to the traditional techniques basing on knowledge or possession. Compared to the later two techniques, biometric modalities are firmly connected to the body or the behavior of the owner. With these physiological and behavioral characteristics a biometric system identifies the person itself rather than some information or objects, which can be lost, stolen or handed over.

Digital watermarking technology is widely used in many application fields like copyright owner identification, integrity recognition or annotation of digital content. In general, watermarking is an embedding and retrieval process, where hidden or secret information is embedded into or retrieved from digital content like music, image or video [1]. However to date, in general only few approaches of applying watermarking techniques for embedding information in biometric data have been published and in particular no analysis of the effects of watermarking of speech media can be found, as discussed later in our paper.

Since the content of the medium is modified by marking, it is important to examine the influence of the change on the biometric authentication. In this publication we investigate the impact of these changes to the authentication performance of the whole biometric system.

In our scenario, speech data shall represent the only medium for an authentication system and the goal is to integrate ancillary information, denoted as **metadata**, in this medium. In practical applications, such metadata may consist of **additional multimodal biometric information** or **additional properties of subjects and the technical environment**. Multimodal speaker authentication is therefore defined as watermarking speech data with multiple modes of an information channel. In the first case, metadata may include compact feature template representations of a complementary biometric trait, such as biometric hashes for online handwriting [2] or iris codes [3]. By embedding such multimodal metadata into the original

speech data, multimodal recognition can thus be achieved while utilizing only one single media stream: audio.

Secondly, it has been shown recently that information retrieved from biometric data is not only related to the characteristics of individuals, but also depend strongly on the social, ethnical or technical environment of the application. From offline analysis of hand-written documents, for example, it is possible to derive group discriminatory information such as gender or ethnicity [4] and also it has been shown that a specific language of a spoken sequence can be identified by biometric features [5]. On the other side, it has been shown that the recognition performance of biometric algorithms depend strongly on technical specifics such as temporal and spatial resolution of sensor devices for handwriting [5] and on cultural properties of the user population such as the language written [6]. Consequently, by modeling such non-biometric data into ancillary metadata and making such metadata available to the biometric system, the recognition process will be enabled to perform a local optimization with respect to particular metadata groups.

Soutar et al. suggested in [7] a method that hides cryptographic keys into biometric reference data by using a bit replacement algorithm. If the biometric system authenticates the user successfully, this algorithm extracts the key and returns it to the system. In [8] Jain and Uludag described a steganography and watermarking approach of hiding fingerprint minutiae in images. The suggested amplitude modulation-based watermarking method provides image adaptivity and watermark strength controller. The authors' goal is a secure exchange of the biometric data. They point out that the method is robust to different attacks on the cover image, like cropping or compression. Jain at al. [9] presented a watermarking method that embeds biometric facial data into a fingerprint image. In addition to the use of the fingerprint for authentication, the face can be consulted to confirm the authenticity of the fingerprint image and its owner. The authors report that the algorithm does not degrade the recognition rate of the fingerprint images substantially. Another approach of combining biometrics and watermarking was presented by Namboodiri and Jain in [10]. Their method embeds an online signature as fragile watermark in digital documents in order to verify the integrity of the documents and the identity of the author.

With respect to **speech biometrics**, a multimodal biometric scenario was described by Schimke et al. in [11] where speech and handwriting was combined. The authors' idea is to embed the online handwriting data (i.e. time dependent X-, Y-position or pressure), captured by a graphical tablet, into the speech data using a LSB watermarking algorithm. Based on this watermarking algorithm the authors have presented a capacity list for mono channel audio files in wav-format, as shown in Table 1.

Table 1. LSB watermarking capacity of mono channel audio files in wav-format [11]

Sampling rate	Capacity per second
48,000 Hz	48,000 bits = 6,000 bytes
44,100 Hz	44,100 bits ≈ 5,512 bytes
32,000 Hz	32,000 bits = 4,000 bytes
22,050 Hz	22,050 bits ≈ 2,756 bytes
16,000 Hz	16,000 bits = 2,000 bytes
11,025 Hz	11,025 bits ≈ 1,378 bytes
8,000 Hz	8,000 bits = 1,000 bytes
6,000 Hz	6,000 bits = 750 bytes

Since the cover medium, here a speech audio file (wav-format, mono channel, 16 bit, 44,100 Hz), constitutes the reference information for a biometric authentication, independent of the embedded content, it is also important, that the **degradation in recognition performance is limited**. However, this trade-off effect of watermarking embedding capacity and recognition performance has not been studied in the original work. In order to study this effect, we have examined the impact of embedded metadata to the authentication performance of the speaker recognition system in this work. For our experimental evaluation we selected a well known speaker authentication approach and widely used audio watermarking algorithm.

This paper is structured as follows: In section 2, we give a short description of the three approaches: our selected speaker authentication system, the metadata approach, as well as the chosen watermarking technique. Section 3 provides an overview of the test database and the evaluation methodology. We present experimental results of the combination of all three selected approaches in section 4. In section 5, we summarize this article, draw the most relevant conclusions for our research and suggest further activities in this area.

2. Biometrics and digital watermarking

In this section we describe the techniques for speaker authentication, for watermarking speech signals using a LSB technique and our concept of metadata. In the last sub-section we specify a combination of all three selected approaches.

2.1. Speaker authentication

Our speaker authentication system is based on Mel-Frequency Cepstrum Coefficients (MFCC), currently being one of the most popular and widely used feature extraction methods. MFCC's are a model of the human perception of sounds. On the one hand, this is aspired by using a mel-frequency scale rather than frequencies themselves. The mel scale was proposed by S. Stevens, J. Volkman and E. Newman in 1937 as a measure of the perceived pitch which is nearly linear for frequencies below 1,000 Hz and logarithmic above [12]. On the other hand the cepstrum of signals are used, whereby the cepstrum is the Fourier transform of the log Fourier transform or the spectrum of the log spectrum [13].

In our system, each of the input wave files has a sampling frequency of 44,100 Hz and a sampling precision of 16 Bit. To reduce the influence of the textual content of the utterances to how it was spoken, the algorithm first blocks the input signal in frames of 30ms length, using a hamming window function with an overlapping shift of 10ms. Then the total frame energy is compared against a threshold to discard frames with silence or low noise only. A filter bank with *L=20* mel-spaced triangle bandpass filters *l*, thus more narrow aligned at low frequencies, ranging up to 8,000 Hz was applied to the spectrum of every remaining frame to get the corresponding mel-frequency wrapped spectrum Ψ. As described in [14], with an adoption that our implementation is using "simple" MFCCs instead of the proposed T-MFCCs using a Teager Energy Operator, the frame's acoustic vector was calculated according the following equation for each cepstrum coefficient *k*:

$$MFCC_k = \sum_{l=1}^{L} \log \Psi(l) \cos\left[\frac{k(l-0{,}5)}{L}\pi\right], k = 1,2,...,L$$

Each acoustic vector is then added to the frame's acoustic vector set.

In enrollment mode for each enrollment's reference model the LBG algorithm [15] selects 32 reference vectors (centroids) out of the enrollment's acoustic vectors. In verification mode the verifications vector set's score is the minimum of all Euclidean distances between each verification vector and each reference vector.

2.2. Metadata in biometric context

In this work we have embedded metadata based on individual user information and technical settings into biometric reference data for speaker verification. In our previous work [6] we already examined the influence of the consideration of biological, cultural and conditional aspects to the biometric authentication. Based on online handwriting verification, we have shown the impact of these metadata to biometric user authentication. In context of the collection of biometric data both handwriting and speech were captured. In our first tests to determine the recognition precision the following information are embedded into the speech reference audio files, including description: The *SampleID* is the ascending internal number of the speech files in the database. An event (*EventID*) describes a collection of samples belonging together because of originator, semantic and action (enrollment, verification or forgery). The internal identification number of the user is stored in the *PersonID*. The *SemanticID* encodes the semantic of a speech task. Semantics represent utterances with different content and duration, which have been collected from the test subjects according to a predetermined task list. They are divided in individual, creative and predefined tasks (see section 3.1). The hardware device of the voice recording is defined in the *DeviceID*. Further, *Date* and *Time* of recording is stored as metadata. In the *LanguageID* the spoken language of an utterance is encoded and the environment of the capturing (e.g. soundproof cabin) of the speech is stored as *EnvironmentID*.

As shown in row two of Table 1, watermarking payload of audio files, approximately 5,500 bytes per second are available for our metadata. The metadata as described above, which we have used in our first tests, have an average payload of 215 bytes, which will be embedded repeatedly in the speech data during watermarking.

2.3. Watermarking using LSB

Digital watermarking is used to embed and retrieve a hidden (in some cases also secret) information into digital content like images, audio or video data. As stated earlier, we set our focus on audio signals, which are recorded from the speaker recognition system. For a general introduction to watermarking see for example from the variety [16] or [17]. In general the most important properties of digital watermarking techniques are robustness, security, imperceptibility/ transparency, complexity, capacity and possibility of verification (detection) as well as invertibility (see [18]). For our first evaluation of recognition precision the most relevant parameter for the first tests are blind verification (detection), capacity for the metadata and high transparency. Note that in this initial work, we do

not consider any optimization towards robustness or fragility of the resulting watermark. More detailed discussions on these aspects can be found for example in [19]. Our used watermarking algorithm operates in a *blind verification method*, where the original media is *not required* for retrieval of the payload. To ensure high transparency and the required capacity we chose a known watermarking approach operating in time domain by embedding the watermark information (our metadata) into the least significant bits (LSB) of the audio signal by using the sequence of least significant bits of sample values [18], thus methods of this kind are typically referred to as *LSB watermarking*. In our implementation, a single bit of the watermark information is embedded into one sample value of the audio signal. The implementation provides two embedding modes: with or without a secret key (k). If no key is used, the message is embedded into each LSBs of each of the samples of the audio signal. If the key is used, then the watermark is not embedded in all LSBs. This is due to the fact that the key initializes a Pseudo-Random Number Generator (PRNG), generating values, which are used to scramble the embedding position and to select the embedding positions of marked LBSs. Both modes can be combined with an Error Correction Code (ECC) based on the Viterbi algorithm [20]. If ECC is used, then the whole watermarking message size is doubled. By employment of ECCs, the algorithm includes a mechanism to correct errors that can occur during transmitting or attacking the audio signal. To provide a reliable verification (detection and retrieval) of the entire watermark, the embedding process embeds the watermarking message multiple times into the audio signal. Here, if the size of the secret message is smaller than the possible embedding positions, the secret message is repetitively embedded until the audio file end is reached. As already mentioned the overall maximal capacity is approximately 5,500 bytes per second where we embed the metadata payload of 215 bytes repeatedly.

In particular the watermarking parameters with their type are defined as: secret key k (integer), embedding message m (string), ECC c (Boolean), maximal scrambling size j (integer).

In the embedding mode, m our metadata is embedded into the audio signal. If k is used, then j specifies the maximal distance between two marking positions. This decreases the potential embedding capacity (since only parts of all LSBs are used) and it increases the transparency (minor signal manipulations occur). If no k is used, then the embedding capacity increases and the transparency decrease.

For detection (verification) of our watermark, only k and j (if used) are required and the knowledge of if error correction was used during the embedding process.

2.4 Combining biometrics, watermarking and metadata

For the tests we divide the analysis into two scenarios: Firstly, we examine the authentication with watermarked reference data and unmarked authentication data. In the second scenario the watermarked reference data are compared with watermarked authentication data.

2.4.1 Watermarked reference data and unmarked authentication data. Figure 1 shows the general process of authentication based on enrollment data and test data including watermark embedding and retrieving. During the enrollment process the metadata and speech data are captured. In the next step the watermarking algorithm embeds the metadata information into the speech file repetitively until the full capacity for each audio signal is reached. This watermarked file is stored in the reference database. In the authentication process the metadata information is retrieved from the reference data. Then, the reference data pass through the preprocessing and feature extraction and will be compared with the captured speech data after the same preprocessing and feature extraction procedure in the matching process. The matching delivers a value of similarity or dissimilarity (*matching score*) between reference and authentication data. Based on this score the decision module will be able to make a decision upon the authenticity of the speaker.

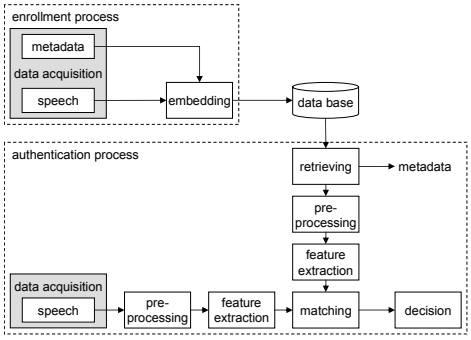

Figure 1. Enrollment and authentication using marked reference data and unmarked authentication data

2.4.2 Watermarked reference and authentication data. Figure 2 presents an authentication process with

the difference to Figure 1 that in addition to reference data the verification data were watermarked with the (same) metadata, as well.

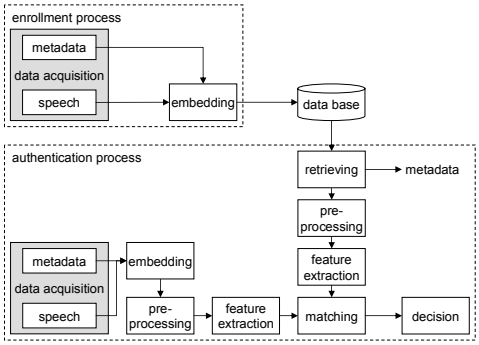

Figure 2. Enrollment and authentication using marked reference and authentication data

In our tests, described in the following section 3 and evaluated in section 4, we study the impact of the watermark embedding on the overall authentication performance of our biometric speaker recognition system (MFCC approach), by studying the recognition errors for both scenarios for our experimental data collection. Embedding information is a special case of additional noises within the audio data. In this state of our work we focus on aspects of quality loss in terms of biometric measurements and due to watermarking embedding because of the MFCC approach's noise sensitivity. On the other side there are some approaches for the MFCC method to improve the authentication performance in noisy environments ([21], [22]). However, in future applications, retrieval of such metadata information may be utilized to optimize the recognition accuracy of the authentication process or can support binning strategies in case of identification.

3. Experimental setup

In this chapter the examined data and the methodology used for the tests are described. Furthermore we explain our measurement of error rates in the biometric context.

3.1. Test database

The audio data were captured from German and Indian test persons within the scope of the cross-cultural CultureTech project (see [23]). For the tasks in this project we aquired speech and handwriting data from each of our test persons. In addition, the acquisition of individual metadata of each participant was necessary. We refrained from using an existing database, like the NIST speaker verification data [24], due to the insufficient amount of metadata for our purposes. Consequently, we collected the captured speech and handwriting data in our proprietary database. For the speech acquisition a test plan with 47 different semantics in two languages (English, German) was developed, where the semantics are based on individual, creative and predefined tasks. A single task was captured with 10 iterations, where the first 5 are used as reference data and the remaining 5 as authentication data. Table 2 shows an overview of the different tasks and their classification. The audio files are recorded with a sampling frequency of 44,100 Hz and a sampling precision of 16 Bit using a headset microphone in a laboratory environment for a uniform data collection.

For our first initial tests we selected five semantics in English from the set of 47. The audio samples are captured from 16 Indian test persons for each of the five semantics and from 16 German test persons for three semantics and for 17 Germans for the remaining two out of the five semantics.

Table 2. Classification of speech tasks

Number of task	Task	Classification of task
1, 32 – 39	Numbers	Given
2	Pass phrase	Individual, creative
3	Number 0 - 9	Given
4	Latin alphabet	Given
5 – 10	Questions to answer	Individual
11 – 13, 17 – 32	Words	Given
14 – 16, 40 – 46	Sentences	Given
47	Passage	Given

The sentences "*She sells sea shells on the shore."* and "*Hello, how are you?"* are representative for predefined tasks with a length at the average of 3.08/2.61 (Indians/Germans) seconds (average duration) and 1.83/1.35 seconds (average duration), respectively. A predefined semantic with a short duration is the word "*Communication"* (1.54/1.22 seconds). The questions "*What is your good name?"* and "*Where are you from?"* are tasks with individual answers from the test persons. They have a short duration at an average of 1.40/1.09 seconds and 1.33/1.10 seconds, respectively.

In our test environment we use the verification mode for authentication. During the verification a claimed user identity is confirmed by the biometric system. The person is verified if the confirmation is successful, in the other case the person is rejected from the system.

3.2. Methodology

For each semantic and each user's nationality the tests are divided into three parts, each composed of

verification and random forgery tests. Firstly, we started the procedure without embedded metadata in the reference speech files. Secondly, the embedding capacity of the LSB watermarking algorithm was fixed at the lowest value by setting the maximal scrambling size *j* to highest value that will most likely only embed one instance of the message. For the third test we switched the LSB capacity to maximum. The tests on different capacities are running under the following conditions twice: In the first setup only the enrollments are marked. In the second scenario both, enrollments and verifications are watermarked with the same metadata information. Furthermore the LSB watermarking algorithm operates with a fixed secret key *k* and without any ECC.

For the comparison of the impact of embedded metadata in different semantics we use the well known biometric error rates: The False Rejection Rate (FRR) indicates, how frequently authentic users are rejected from the system. The calculation of FRR is based on comparison of the enrollment and verification of each user. The False Acceptance Rate (FAR) specifies the acceptance of non-authentic persons and is determined based on the relation of the enrollment and verification data of different users. In order to compare the different semantics and the different strengths of watermark embedding we use the Equal Error Rate (EER) measurement. The Equal Error Rate is found at the point of the intersection of the characteristics of FRR and FAR, i.e. where FRR and FAR yield the same value. The EER is not necessarily the optimal operating point in every biometric system and measurements such as Receiver Operating Characteristics (ROC) may provide more detailed information about the system's characteristics, but it is an initial clue for comparing recognition capability of biometric systems.

4. Experimental results

Table 3 and Table 4 summarize the results from our experiments for each of the five semantic classes (five columns from left). For each semantic class, the first (No. of Speakers) and second row (Total No. Samples) indicate the number of test subjects and the total number of samples taken from all persons respectively. Rows three to five outline the average speech duration of each semantic class, and the number of tests performed to analyze false rejections (No. FRR Tests), false acceptances (No. FAR Tests) in our scenario. Row six (EER without WM) shows the results for tests without embedded metadata. Finally, the lower six lines present the EER divided into tests with watermarked enrollments vs. unmarked verifications and watermarked enrollments vs. watermarked verifications. The results here are obtained for the two scenarios: metadata embedded with low capacity (EER low capacity) and high capacity (EER High Capacity) respectively.

In Table 3 we observe, that in all five cases of the English test samples of the Indian participants, the EER remains unchanged, regardless, whether a metadata watermark is present in the speech media or not. This result can be confirmed both for the watermarked verifications and for the unmarked verifications.

With an EER of 0.171 the best authentication results for the Indian test persons are reached by the individual answer to the question "*Where are you from?*" (see column "Where are …?"). On the other side the given sentence "*She sells sea shells on the shore.*" has the worst authentication results with an EER of 0.333 (see column "She sells …").

Table 3. Equal Error Rates for Indians obtained from different semantic classes

	She sells …	Hello, how …?	Com-muni-cation	What is …?	Where are …?
No. of Speakers	16	16	16	16	16
Total No. Samples	161	160	160	160	161
Avg. Speech Duration	3,08	1,83	1,54	1,40	1,33
No. FRR Tests	405	400	400	400	405
No. FAR Tests	6075	6000	6000	6000	6075
EER without WM	0,333	0,223	0,255	0,255	0,171
Watermarked enrollments and unmarked verifications					
EER Low Capacity	0,333	0,223	0,255	0,255	0,171
EER High Capacity	0,333	0,223	0,255	0,255	0,171
Watermarked enrollments and watermarked verifications					
EER Low Capacity	0,333	0,223	0,255	0,255	0,171
EER High Capacity	0,333	0,223	0,255	0,255	0,171

Table 4. Equal Error Rates for Germans obtained from different semantic classes

	She sells …	Hello, how …?	Com-muni-cation	What is …?	Where are …?
No. of Speakers	16	16	16	17	17
Total No. Samples	160	161	160	171	171
Avg. Speech Duration	2,61	1,35	1,22	1,09	1,10
No. FRR Tests	400	405	400	430	430
No. FAR Tests	6000	6075	6000	6880	6880
EER without WM	0,388	0,277	0,298	0,310	0,326
Watermarked enrollments and unmarked verifications					
EER Low Capacity	0,388	0,277	0,298	0,310	0,326
EER High Capacity	0,388	0,277	0,300	0,310	0,326
Watermarked enrollments and watermarked verifications					
EER Low Capacity	0,388	0,277	0,298	0,310	0,326
EER High Capacity	0,388	0,277	0,300	0,311	0,326

In Table 4 the results of five semantics of the German test persons are presented. Here we observe that in three out of five cases, the EERs yield identical values in each of the three scenarios (without WM, Low Capacity and High Capacity) for both setups, watermarked and non watermarked verifications.

Apparently, this holds true for those two semantic classes having rather middle speech duration ("*She sells sea shells on the shore."* (see column "She sells ...") and "*Hello, how are you?"* (see column "Hello, how …?") and one with short duration ("*Where are you from?"*, see "Where are …?"), see second, third and rightmost column. For the remaining semantic classes, "*Communication"* and "*What is your good name?"* (see *"What is ...?"*) we observe a slight increase in the EER, at least when embedding at a high capacity. In the worst case, EER decreased from 29.8% to 30.0% for the semantic class "*Communication"*, thus a relative increase less than 1%. On the other side, for semantic "*What is your good name?"* with watermarked enrollments and watermarked verifications, the EERs of low (EER=0.310) and high (EER=0.311) capacity deviate by only 0.01% points from each other.

For the German test persons the best authentication results (EER=0.277) are reached by the predefined sentence "*Hello, how are you?"* (see column "Hello, how …?"). In this scenario the worst authentication are determined for the same semantic as for Indians: the given sentence "*She sells sea shells on the shore."* (see column "She sells ..."). Here the EER has a value of 0.388.

5. Conclusions and future work

The test results have shown that for the selected MFCC approach combined with LSB watermarking it is possible to embed metadata in speech based biometric reference data without decreasing the authentication performance considerably. We have shown that the differences between non-watermarked data, watermarked data with low and high capacity are marginal or not existent. These results have been reconfirmed for biometric speaker verification with watermarked reference vs. watermarked verification data as also with watermarked reference, as well as for unmarked verification data. Consequently, our approach may be applied for implementing multi-modal biometric authentication systems based on a single audio media carrier and metadata containing complementary biometric references such as biometric hashes or iris codes.

We know that the used data are not sufficient in order to achieve statistic significance. On the other side this is an initial investigation, which examines the influence of embedded data on biometric recognition performance. In the further process of our work we will acquire further biometric data including the corresponding metadata and include these into our investigations. Additionally also the use of other data bases is considered for cross-validation, however the absence of metadata in such alternative databases such as NIST speaker verification data [24] will require additional strategies do simulate such metadata.

As these first tests are performed only with an LSB watermarking technique, our future work is focused on the evaluation of the impact of different watermarking approaches like frequency or wavelet domain techniques. In case that robustness is required for the metadata embedded in speech audio, for example due to errors or noise on transmission channels or for ensuring owner protection, these alternative methods might be more appropriate than LSB. The effect on the biometric user data of such different watermarking algorithms then needs to be further evaluated. Further also tests should be accomplished in order to determine the extent of audible changes of the audio data by embedding information and the impact of lower quantization (e.g. 8 bit) of the audio samples.

In our current research, the payload of embedded metadata information is not fully exploited, we used in average 215 bytes repeatedly. Therefore it is possible to hide further biometric information such as other modalities as payload into the metadata (see [11]). Through this a multimodal authentication is feasible. Another possibility is to embed a knowledge based hash (i.e. password hash) as metadata, in order to use it in a multi-factor authentication. Multi-factor means a combination of biometric based (e.g. handwriting) and non-biometric based (e.g. knowledge, possession) user authentication. In this case the input knowledge can be confirmed by the knowledge retrieved from biometric reference data in addition to the biometric authentication. Further, with a maximum theoretical LSB watermarking capacity of 5.500 Bytes per second, our approach provides additional capacity for ancillary metadata regarding the personal and technical background, as described in this paper. Here the multimodal system can parameterized depending on the payload of the embedded metadata, e.g. user depending weights for each of the subsystems. Another application for the embedded metadata is to consider different hardware devices during capturing of enrollment and test data. Here, since both the type of the enrollment device is encoded in the metadata and the device used at time of authentication is known, it appears possible to parameterize or use alternative preprocessing and feature extraction modules for both, reference data and test data, in order to increase the authentication performance.

Based on the promising work on the combined LSB and MFCC technique in summary the future work will focus on the one hand on the evaluation of the impact of different watermarking approaches as well of

different speaker authentication approaches. On the other hand we will investigate the potential application fields and the required metadata to determine the watermarking parameter capacity, robustness and transparency as well as the required recognition precision.

6. Acknowledgments

This publication has been produced partly with the assistance of the European Union (project CultureTech, see [23]). The work on the general test methodology described in this paper has been additionally supported in part by the European Commission through the IST Programme under Contract IST-2002-507634 BIOSECURE. The content of this publication is the sole responsibility of the University Magdeburg and their co-authors and can in no way be taken to reflect the views of the European Union.

7. References

[1] J. Dittmann, Digitale Wasserzeichen, Xpert.press, Springer Berlin, ISBN 3-540-66661-3, 2000

[2] C. Vielhauer, Biometric User Authentication for IT Security: From Fundamentals to Handwriting, Springer, New York, ISBN: 0-387-26194-X, 2005

[3] J. Daugman, "How Iris Recognition works", IEEE Trans. CSVT 14(1), 2004, pp.21-30

[4] C. I. Tomai, D. M. Kshirsagar, S. N. Srihari, "Group Discriminatory Power of Handwritten Characters", Proceedings of SPIE-IS&T Electronic Imaging, 2004, pp. 116-123

[5] C. Vielhauer, T. Basu, J. Dittmann, P.K. Dutta, "Finding Metadata in Speech and Handwriting Biometrics", Proc. of SPIE-IS\&T Electronic Imaging, Vol. 5681, ISBN 0-8194-5654-3, 2005, pp. 504-515

[6] T. Scheidat, F. Wolf, C. Vielhauer, "Analyzing Handwriting Biometrics in Metadata Context", to appear in: SPIE Proceedings - Electronic Imaging, Security and Watermarking of Multimedia Contents VIII, 2006

[7] C. Soutar, D. Roberge, A. Stoianov, R. Gilroy, B.V.K. Vijaya Kumar, "Biometric Encryption", R.K. Nichols (Ed.), ICSA Guide to Cryptography, McGraw-Hill, 1999

[8] A.K. Jain, U. Uludag, "Hiding fingerprint minutiae in images", Proc. Automatic Identification Advanced Technologies (AutoID), New York, March 14-15, 2002, pp. 97-102.

[9] A.K. Jain, U. Uludag, R.L. Hsu, "Hiding a face in a fingerprint image", Proc. International Conference on Pattern Recognition (ICPR), Canada, August 11-15, 2002

[10] A.M. Namboodiri, A.K. Jain, "Multimedia Document Authentication using On-line Signatures as Watermarks", Security, Steganography and Watermarking of Multimedia Contents VI, San Jose California, June 22, 2004, pp. 653-662

[11] S. Schimke, T. Vogel, C. Vielhauer, J. Dittmann, "Integration and Fusion Aspects of Speech and Handwriting Media", Proceedings of the Ninth International Conference Speech and Computer, SPECOM'2004, ISBN 5-7452-0110-x, 2004, pp. 42-46

[12] S.S. Stevens, J. Volkman, E.B. Newman, "A scale for the measurement of the psychological magnitude of pitch", Journal of the Acoustical Society of America, 8, 1937, pp. 185-190

[13] J. W. Tukey, B. P. Bogert, J. R. Healy, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autovariance, cross-cepstrum and saphe cracking", Proceedings of the Symposium on Time Series Analysis, 1963, pp. 209-243

[14] H. A. Patil, P. K. Dutta, T. K. Basu, "The Teager Energy Based Mel Cepstrum for Speaker Identification in Multilingual Environment", Journal of Acoustical Society of India, Nov. 2004

[15] Y. Linde, A. Buzo, R. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, 1980, pp.84-95

[16] J. Bloom, M. Miller, I. Cox, Digital Watermarking: Principles & Practice, Morgan Kaufmann Publishers, San Francisco, ISBN 1-558-60714-5, 2001

[17] Y. Duan, I. King, "A Short Summary of Digital Watermarking Techniques for Multimedia Data", Department of Computer Science & Engineering, The Chinese University of Hong Kong. Shatin, N. T., Hong Kong, China, 1999

[18] J. Dittmann, P. Wohlmacher, K. Nahrstedt, "Multimedia and Security – Using Cryptographic and Watermarking Algorithms", IEEE MultiMedia, October-December 2001, Vol. 8, No. 4, ISSN 1070-986X, 2001, pp. 54-65

[19] A. Lang, J. Dittmann, "Profiles for evaluation and their usage in Audio-WET," in IS&T/SPIE's 18th Annual Symposium, Electronic Imaging 2006: Security and Watermarking of Multimedia Content VIII, Vol. 6072, P. W. Wong and E. J. Delp, eds., SPIE Proceedings, (San Jose, California USA), 2006

[20] I.A. Glover, P.M. Grant, Digital Communications, Prentice Hall, 1997

[21] Z. Wu, Z. Cao, "Improved MFCC-Based Feature for Robust Speaker Identification", Tsinghua Science & Technology, Volume 10, Issue 2, April 2005, pp. 158-161

[22] Q. Zhu, A. Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise", Computer Speech & Language, Volume 17, Issue 4, October 2003, pp. 381-402

[23] The Culture Tech Project, Cultural Dimensions in digital Multimedia Security Technology, a project funded under the EU-India Economic Cross Cultural Program, http://amsl-smb.cs.uni-magdeburg.de/culturetech/, requested July 2005

[24] NIST Speech Group Home, http://www.nist.gov/speech/index.htm, requested January 2006