Abstract- This paper presents a double-digit system for text-dependent speaker verification and text validation that can be accessed with security via a telephone network. A speech recognition and a user authentication scheme are proposed that utilize concatenated phoneme HMMs and operate in a sound-prompted mode. The recognition performance of the system is evaluated through extended simulations on Hidden Markov Models (HMMs) using Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC), as well as Cepstral Mean Subtraction (CMS). The effect of various factors such as the length of the training data, the number of embedded re-estimations, and Gaussian mixtures in training of the HMMs, the use of world models, bootstrapping, and user-dependent thresholds on the performance of speech recognition and speaker verification is also examined.

I. INTRODUCTION

Due to the growth of transaction-type telephone applications, such as telephone-banking and mobile-commerce, higher security protection should be provided to customers to further increase their respect and interest. Authentication methods using biometrics can replace or complement conventional authorization mechanisms, namely passwords and personal identification numbers (PINs), for higher security applications [1]. Speech is the commonest means of user- interaction with a telephone service and therefore voice biometrics could be regarded as a promising low cost solution to be deployed. Their clear advantage is that the biometric authentication can be performed by the service creating no further need of additional equipment than a telephone headset.

The current contribution proposes such a voice biometric solution for secure access to a telephone service based on a Double-Digit (DD) speaker verification system that uses speech over landline and mobile networks. The system incorporates and tests a speech recognition mechanism to validate whether the speaker pronounced the prompted sequence correctly. Speech validation prohibits false rejections due to mispronunciation of the prompted DD sequence. The well established concatenated phoneme HMM [2] technique for modeling the inherent speaker variability was used and various parameters were examined against speech validation and speaker verification performance. Implementation, training, and testing of the HMM models were achieved through the use of the Entropic HTK toolkit [3] utilizing both the Perceptual Linear Prediction coefficients (PLP) [4] and Mel Frequency Cepstral Coefficients (MFCC) [5], [6].

II. OVERVIEW OF DOUBLE-DIGIT SPEAKER AUTHENTICATION AND VERIFICATION SYSTEM

The proposed voice biometric system comprises of two parts, namely a text-dependent concatenate phoneme HMM-based speaker verifier, and a concatenate phoneme HMM-based speech recognizer. Both parts are working in a sound-prompted mode that is used for both the enrollment and verification procedures. It has been demonstrated in [7] that sound prompts leaves the user a more difficult task than text prompts, and more speaking errors are therefore produced. In text-dependent HMM-based speaker verification, a speaking error results to a defective enrollment or incorrect authentication. As a countermeasure to this problem, a speech recognition procedure has been incorporated into the system in order to validate the sound-prompted password.

A. System Architecture

Fig.1 shows the structure of the proposed double-digit speaker authentication system. The first step involves the capturing of speech samples as the user is voice-prompted by the system for utterances. This procedure is repeated both in the enrollment phase where the HMM models for each user are created and the verification phase where the system verifies that the captured speech matches the models of the verified user. The second step involves the calculation of voice features, a task performed by a front-end feature extractor. These features are used for both text validation and speaker verification. Once the utterances have been validated, the authentication procedure begins.

In the enrollment phase, the system creates speaker-specific phoneme models for each reference speaker. In the speaker verification phase, the phoneme-concatenation model corresponding to the prompted double-digit sequence is constructed, and the accumulated likelihood of the input speech frames for the model is compared with a threshold to decide whether to accept or reject the speaker. In the case of successful speaker verification, the features of the speech signal are stored for updating the HMM models of the specific speaker. The vocabulary that is used by the system consists of two digit numbers spoken continuously in sequences (e.g. “29-34-52-76”). This vocabulary is advantageous since HMM models can be trained for all the double-digits combinations, and random sequences can be generated for authentication, thus increasing the robustness of the system against impostors.
B. Data Collection

For the evaluation part of this contribution, three double-digit sequence speech databases were used, namely the In-house database, the YOHO-PSTN, and the YOHO-GSM Database. The In-house Database comprises of data that were collected over a period of four months over the GSM and PSTN networks. It contains speech samples from 23 speakers, that are categorized into groups, one for the enrolment and another for the verification procedure. These two databases are exact replicas of the YOHO corpus [8] recorded (using an analogue voice modem) over the PSTN and GSM networks. The same databases were used for initial training of the HMM models.

III. EVALUATION OF TEXT VALIDATION SCHEME

Experiments were conducted for HMM double digit sentence recognition in order to evaluate the performance of the text validation over the two telephone channels. The performance on text validation is evaluated against various parameters such as the effect of embedded re-estimations and bootstrapping in training, the number of Gaussian mixtures of the HMM models, as well as the use of PLP and MFCC coefficients.

A. Effect of Embedded Re-estimations and Gaussian Mixtures (GMs)

The first experiment has been conducted to examine the effect of the number of embedded re-estimations of the Baum-Welch Algorithm [9] on DD recognition performance. The test was performed by training continuous-density single Gaussian mono-phone HMM models (18 mono-phones) modeled by 3 left-to-right states [10]. A silence model was also trained for modeling the beginning and ending of an utterance, and the intermediate pauses. The features used were 12-order MFCCs and the normalised energy coefficient augmented by the corresponding delta and delta-delta coefficients. Fig.2 shows the monophone occurrence of the YOHO-PSTN database that was used for the training of the speaker independent mono-phone models. The aim of the test was the verification of the In-house database comprising of 23 users, each one pronouncing 50 times the double-digits ‘twenty-nine’, ‘fifty-two’, ‘seventy-six’ and ‘eighty-one’.

Fig.3 shows the percentage of DD recognition against the number of embedded re-estimations of the Baum Welch algorithm. It can be observed that the performance of DD recognition is increasing when using up to 4 embedded re-estimations. By increasing the number of re-estimations, the performance asymptotically converges to the maximum performance for the specific 1 GM system. Although the DD recognition was shown to converge after 4 embedded re-estimations, the overall recognition performance does not exceed 70%. This is due to the fact that the single GM HMMs were not able to provide a good parametric modeling of the acoustic space. Therefore, the experiments were repeated so as to investigate whether more GMs per state can provide a better DD recognition. For each set of the new experiments, the GMs were split by a factor of 2 and re-trained using 4 iterations of the Baum-Welch algorithm. Fig.4 shows that using four embedded re-estimations of the Baum-Welch algorithm, the overall DD recognition performance increases with the number of GMs. The computational complexity was observed to increase exponentially with the number of GMs. Since the increase in performance from four to eight GMs per HMM state did not compensate for the computational complexity which almost doubled, a decision was taken to use only four GMs for the remaining tests. Hence, it can be concluded that by using only four GMs and four embedded re-estimations of the Baum-Welch algorithm, satisfactory performance can be achieved with reasonable computational complexity.
B. Training Procedure with YOHO-PSTN Database for Bootstrapping

The previous experiments investigated the DD recognition performance of the system using the American accent speech data of YOHO database performing tests using the Greek accent speech data of the In-house database. In this section, it is examined whether using a pre-trained HMM prototype instead of the simple prototype of zero means and unitary variances (used in the experiments of previous section), can result in a better performance of DD recognition and whether additional training will adapt the models to the Greek accent and pronunciation of the speakers in the In-house database.

The experiments of Section III.B were repeated using the YOHO-PSTN trained HMM models for bootstrapping additional training using the enrolment files of the In-house database, and testing using the verification files of the database as above. To compare the effect of the bootstrapping procedure on the DD recognition performance of the system, identical experiments were repeated using a simple prototype of zero means and unitary variances for HMM training.

Experiments showed that bootstrapping the HMM training procedure results in a better DD recognition. Fig. 5 shows that using the YOHO-trained HMM to bootstrap the training on the In-house database increases the sentence recognition performance by approximately 2-4%.

C. PLP and MFCC coefficients on DD recognition

In this section, the use of PLP coefficients was compared with the classical MFCC speech recognition parameterization that was utilized in all previous experiments, so as to evaluate the effect of the former on DD recognition performance. Tests were conducted using 12-order MFCCs and PLPs, the normalised energy coefficient augmented by the corresponding delta and delta-delta coefficients. The HTK toolkit was used for implementing the front-end mechanism. Initial single Gaussian HMM models were trained using session III of both YOHO-GSM and YOHO-PSTN databases and were used to bootstrap additional HMM training on the 80% of the In-house database. A total of 2277 speech files of the In-house database were used for training. We used the remaining 20% of the database (570 speech files) for testing. Identical experiments were performed so as to investigate the performance of the two kinds of coefficients using 2 and 4 GMs per state. The CMS technique [11] was also incorporated in the tests to examine the effect of channel normalization on DD recognition when the recording channel is changing (PSTN, GSM). Identical tests were performed using PLPs and MFCCs, with and without CMS. It was demonstrated that PLP coefficients outperform the classical MFCCs under the specific conditions increasing the DD recognition by 2-3%. In addition, the introduction of CMS was found to improve the performance of the system by approximately 2%. Figure 6 illustrates the comparison of the proposed DD recognizer performance when the two sets of features, different number of Gaussian mixtures, and embedded re-estimations are used. It can be seen that the 8 GM DD recognizer which uses PLPs and CMS, trained with 10 embedded re-estimations results to a 98.4% sentence recognition performance. Of course, the 10 embedded re-estimations in training and 8 GMs require relatively higher computational time than other combinations, but this is not an issue for the proposed system since the training of the models will be performed offline. However, computational complexity is an important issue in Speaker Verification when training is performed online.

IV. EVALUATION OF SPEAKER VERIFICATION SCHEME

This section evaluates the speaker verification performance of the system using both the MFCC and PLP coefficients.
A. MFCCs and PLPs for HMM Speaker Verification

In the training part, single Gaussian mixture HMM models [12] were employed for each speaker using the five enrolment sessions (each session contains 10 DD utterances) of the In-house database. Thirty authentication sessions (each session contains 5 DD utterances) from each of the 23 speakers were used for the evaluation of the speaker verification performance. Each speaker was authenticated against all 23 HMM speaker models using his/her 150 double-digit authentication utterances. Fig.8 was created by averaging speaker-dependent HMM scores over the 150 double digits for each speaker. Axis X shows the speakers attacking each model (impostor) while axis Y shows the speaker-dependent HMM models. Axis Z represents the averaged HMM scores for each impostor-model combination. Shifting a horizontal plane along the Z axis and each time taking the point of intersection with Z axis, we calculate the False Acceptance Rate (FAR), False Rejection Rate (FRR) and hence the Equal Error Rate (EER) [9]. In Fig.7, the horizontal plane represents the threshold for which FAR equals FRR. It can be seen that the prominent diagonal represents speaker identification for the 23 sets of In-house database speakers.

Speaker-dependent, three-state left to right single Gaussian HMM models were trained using five and six sessions, respectively. The six session data set contains data recorded over the GSM network. For each of the two sets of enrollment data, four speaker-specific HMM models with one Gaussian mixture were trained using the following feature combinations: MFCCs, MFCCs with CMS, PLPs, and PLPs with CMS. The following tests evaluated the speaker authentication performance against FAR, FRR and EER through use of thirty authentication sessions from each speaker. As it can be seen in Fig.8-9, the CMS method significantly increases speaker verification performance, when applied either on MFCC, or PLP feature sets. This is because CMS is able to compensate for the effect of different recording channels and conditions. The use of PLP coefficients was found to improve the speaker verification performance by 1-4% when compared to the MFCCs.

B. HMM Speaker Authentication Decision Threshold

Applying the threshold which yields a FAR equal to FRR, estimated over all tests as shown in Fig.9, the individual FAR and FRR for each speaker can be estimated. Repeating the tests shown in Fig.9 where six sessions are used for training and examining the FAR and FRR for each individual speaker, we obtain Fig.10 and Fig.11, respectively. An important observation derived from these figures is that FAR is not equal to FRR for each speaker and at some cases the deviation is considerably high. This observation motivated the investigation whether a different verification threshold for each speaker could achieve a much better speaker authentication performance. Repeating the same tests using PLPs and CMS but calculating the EER as the mean of the individual EER of each speaker, it was found that the EER is significantly dropped from 3.52% to 1.14%. The decision threshold estimated as the average individual threshold was found to produce a much better EER when compared to the one estimated by averaging the utterance scores.
C. Normalization of HMM Scores

The normalization of the HMM scores of each individual speaker using world [13] or cohort models [14] is an important factor in estimating an appropriate threshold for speaker authentication. The cohort model approach involves the search for a set of speakers whose characteristics in speech are similar to a specific speaker. However, due to the difficulty of the achievement of this task in a real environment application and because of the heavy computation complexity it involves, this investigation focused on evaluating the effect of a world model. The world model approach relies on the development of a universal speaker model from a pool of speech utterances produced by various speakers. Our evaluations were based on pre-training a world model using all the speakers in the enrollment data and use this model to evaluate speaker authentication using the verification part of the In-house database. The feature set used was the PLPs with CMS. Following the same testing procedure as in Section IV.B, and using the average individual threshold, we evaluated the overall speaker verification performance. The EER calculated over all individual EER for each speaker using a world model was 0.094%, while 1.14% was obtained without using the world model.

V. CONCLUSIONS

The current contribution has proposed a voice biometric system which combines a text-dependent concatenate phoneme HMM-based speaker verifier and a concatenate phoneme HMM-based speech recognizer which operate in a sound-prompted mode over the PSTN and GSM networks. The overall performance of the system has been evaluated using a custom In-house database and two versions of YOHO recorded over the PSTN and GSM networks respectively. During the evaluation of the text validation scheme, four Gaussian mixtures were found to produce high recognition accuracy, while retaining low complexity. The DD speech recognition performance was found to converge asymptotically after four embedded re-estimations of the Baum-Welch algorithm. The bootstrapping of HMM training, and the incorporation of CMS, improved system’s performance. Moreover, the utilization of PLPs was found to give better results compared to MFCCs. Finally, speaker dependent thresholds and the world model could further improve verification performance, resulting in an EER of 0.094%.

ACKNOWLEDGEMENT

This research was supported by the Cyprus Research Foundation, partly by research contract PLHRO/0603/01, and partly by the contract KY-EL/0603/77.

REFERENCES