KTH Computer Science
and Communication

5. The PER system

HÅKAN MELIN

KTH/CSC/TMH
Stockholm, Sweden 2007

This is a self-contained re-print of Chapter 5 in:
Melin, H. (2006). Automatic speaker verification on site and by telephone: methods, applications and assessment. Doctoral Thesis, KTH, Stockholm, Sweden 2006. ISBN 978-91-7178-531-2.

Contents

Chapter 5

The PER system

5.1 Introduction

The PER (Prototype Entrance Receptionist) system is an application of speaker verification created at KTH Department of Speech, Music and Hearing. Its primary, on-site version is essentially a voice-actuated door lock that provides staff and students working at the Department on a regular basis with a means to unlock the central gate to their workplace. The speaker verification system is text-dependent. Users are authenticated using a single repetition of their proper name and a visually prompted, random sequence of digits. The automatic collection of enrollment and test utterances is governed by the system through the use of speech recognition, multi-modal speech synthesis and a graphical display.

In addition to the on-site version of PER, a telephone version was created to support the collection of parallel on-site and telephone data. The speaker verification and speech recognition system components are the same for both system versions, including background/acoustic models and the choice of feature representations. They were initially developed using landline telephone data, and were expected to perform better with landline telephone calls than with calls from mobile telephones and in the on-site version of the system.

The first (on-site) version of PER was installed in 1999 as part of a student project (Armerén, 1999), and the system and its components has since then been improved successively until May 2003 when data collection for an evaluation started. The corpus resulting from this collection is described in Section 6.3 and results from the evaluation are presented in Chapter 10. This chapter focuses on the description of the two versions of the PER system.

5.2 On-site version

The on-site version of the system, depicted in Figure 5.1, serves an entrance where users are physically present. It uses a graphical display to prompt claimants visually

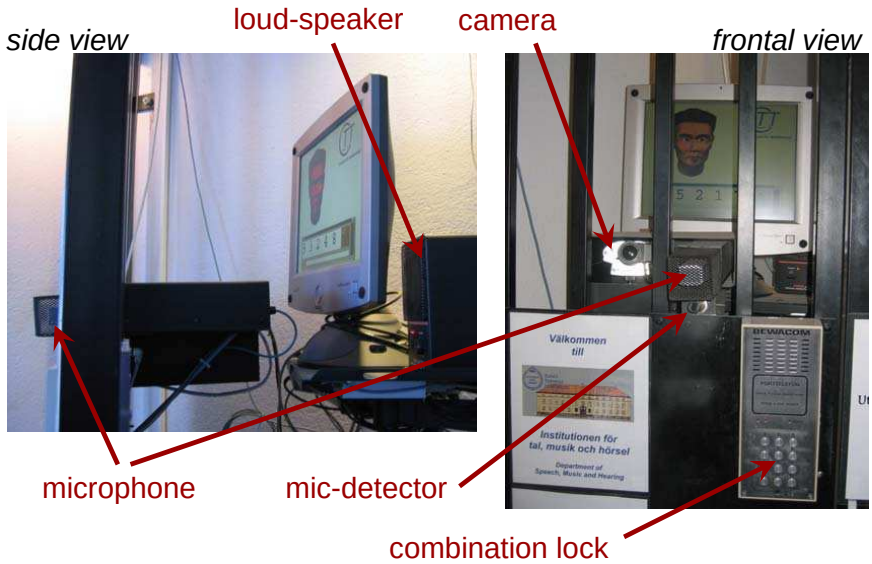

Figure 5.1: The on-site version of PER from side and frontal views. Photos by Botond Pakucs.

for five-digit passphrases. A new passphrase is generated every 10 seconds and the display is updated correspondingly until the system detects the presence of a person. Presence detection is implemented through a diffuse reflection type photoelectric sensor (Diell MS6/00-1H). The sensor is mounted just below the microphone (cf. Figure 5.1) and triggers the start of a session when an object is within approximately 20 cm from the microphone. To allow claimants to start speaking immediately upon arriving at the microphone, the idle system is set in a stand-by mode where it is continually recording audio into a one-second circular buffer. When the presence of a person is detected, input processing starts from the first sample stored in the buffer at that time.

Speech input is recorded with a Sennheiser MD441U directed microphone and sampled at 16 kHz with 16 bits per sample via a SoundBlaster Live! sound card and an external pre-amplifier (M-Audio AudioBuddy). This sample stream is stored to file for future wide-band experiments, while it is decimated to 8 kHz and compressed to 8 bits/sample using A-law coding for use in the on-line processing by system components described below. The decimated and compressed sample stream is also stored to file for off-line experiments. It was used for internal processing because our available development data were telephone data, and both acoustic models for speech recognition and speaker verification background models were trained on such data (cf. Section 10.2.1).

Figure 5.2: Panorama view of the bar gate with the on-site version of PER. Photo by Botond Pakucs.

A video camera (Axis 2120 Network Camera) is installed next to the microphone to capture close-up images of claimants (cf. Figure 5.1). The purpose of the camera is to help annotators decide if a claimant is the target speaker or not, to annotate sessions from the same (unknown) impostor speaker with a single identity label, and to speed up annotation work in general (cf. Section 6.3.3.1). The camera is not used for automatic face recognition, while this would have been an obvious possibility for this on-site application. Figure A.3 shows examples of images captured by the camera.

The gate where PER is installed[1] is an iron-bar gate located in a spacious stairwell just below the two floors housing the Department. The stairwell is a reverberant room with stone floor and bare concrete walls. It spans three floors of the building and contains several potential sources of transient background noise such as doors and talking people. Figure 5.2 shows a picture of the stairwell with the PER system to the right of the bar gate as seen from inside the Department. Figure A.2 shows a closer picture of the gate and PER.

The reverberation time (T60) of the stairwell was measured by Nordqvist and Leijon (2004) to 2.4 s at 500 Hz and 2.1 s at 2000 Hz, while both corresponding values for a typical office in the Department were measured to 0.7 s.

The PER system provided one of three possible ways for employees to unlock the gate, the other two being a combination lock and a regular door lock.

5.3 Telephone version

The main differences introduced in the telephone version of the system with respect to the on-site version are listed in Table 5.1. Except for the choice of authentication

[1] describes the location where evaluation data were collected; the Department (and PER) has moved since then (Figure A.1 shows the new installation)

method during enrollment explained in Section 5.5, differences are all motivated by the limited number of available choices of output modalities in the telephone case, or by the standardization in the telephony system. With (traditional) telephones the speech and audio modality is the only available one, while for on-site applications, any modality could be used. (However, in this work on-site output was limited to the (multi-modal) speech and graphics modalities.)

As shown in the table, the prompting method differs between the two versions of the system. In the on-site case the graphical display is used to prompt the digit string visually. This has several advantages over aural prompting like in the telephone case:

A. No aural prompt is needed to initiate the first attempt from the claimant, allowing for a short time from session start to system decision.

B. Longer digit sequences can be used, allowing for reduced speaker verification error rates. In a previous study (presented in Chapter 8) it was found that using five digits with aural prompting introduced a lot more errors and disfluencies in user responses compared to using only four digits, suggesting that four digits is an upper limit for practical use with aural prompting. Preliminary observations with PER indicated that visual prompts with five digits caused no difficulties for users.

C. With visual prompts it is quite possible to collect the name and the digit sequence in a single utterance, again allowing for a short time to system decision. With aural prompts we believe this would be very difficult for users because of an increased cognitive load and limitations in short-term memory in users, so we chose to collect name and digits separately in a two-step procedure.

5.4 Web interface

A web interface to an SQL database used by PER is provided through the Department's intranet server. The interface serve several purposes. The two first are related to normal use of an access control system:

- Regular PER users (clients) visit their personal page to enable gate or telephone enrollment, and to generate enrollment sheets for telephone enrollment. They can also customize PER's greetings to them.

- The system administrators use a privileged part of the interface to add new users to the database, or delete users. Statistics on system use is also provided, such as enrollment status of users, enrollment durations, the number of sessions, access times, the number of attempts required for access and a list of error messages.

Table 5.1: Differences between the on-site and telephone versions of the system.

Property	On-site	Telephone
Transducer	directed, high-quality microphone	telephone instruments (landline/cellular)
Sampling	16 kHz, 16 bits/sample, then decimated to 8 kHz, 8 bits/sample (A-law)	8 kHz, 8 bits/sample (A-law)
Passphrase prompting method	visual prompts	aural prompts (synthetic speech)
Number of digits	5	4
Collection of name and digit sequence during test	single utterance	separate utterances
Turn-taking	no system prompt before first attempt; graphical indication of when system expects user input	system prompt before every expected user utterance
Session start	optical sensor	telephone call
Authentication during enrollment	30 minute time window from activation via web page	7-digit code and two hour time window

Other purposes are related to the data collection for scientific purposes:

- Client and impostor subjects can see how many sessions are expected from them and how many they have completed, together with statistics on recorded sessions. Impostor subjects are provided with a list of possible target speakers. For further encouragement subjects could also access group statistics of all users (like that provided to system administrators).
- Data collection supervisors are given an overview on what subjects have not yet completed their expected number of sessions, etc., in addition to all the statistics provided to system administrators. This function was very useful during data collection for issuing reminders to subjects. Statistics and results of annotation is also given.

Table 5.2: Digit sequences collected from each subject during enrollment.

Item	Sequence	Item	Sequence
1	3 5 6 0 2	6	6 9 4 1 3
2	7 6 3 2 4	7	2 1 8 5 7
3	9 3 0 4 6	8	0 8 1 7 5
4	8 7 2 9 0	9	4 0 9 6 8
5	1 2 5 8 9	10	5 4 7 3 1

5.5 Enrollment

The use of speaker verification requires clients to enroll. During an enrollment session, the PER system collects speech from the enrolling client (the enrollee) and creates a target model for that client.

The system is set to collect one valid repetition of each of ten items per subject with a proper name and five digits in each item. The digit sequences, listed in Table 5.2, are the same across subjects. They were designed such that each digit occur exactly five times; exactly once in every position within the sequence; and never more than once in a given left-context or right-context. A repetition of an item is deemed *valid* if the on-line speech recognition includes the expected name and digits for that item in its N-best output. The system asks clients to repeat the same item until a valid repetition is found and before moving on to the next item.

To avoid users being held up by repeating a particular item an unreasonable number of times in the event that the speech recognizer repeatedly fails to produce the correct hypothesis, they are offered to skip an item after every sixth attempt on the same item. A user is allowed to skip up to two of the ten items in this way. This skip-possibility was utilized by a few clients as presented in Section 10.3.1, thus the number of collected items per client varied between eight and ten.

Before using PER for the first time, clients have to enable their enrollment via the system's web interface. By doing so, they are given a time window for enrollment of 30 minutes for gate enrollment and 2 hours for telephone enrollment. Access to the intranet is protected by standard user name plus password login that constitutes the main mechanism for authenticating clients at enrollment. For the gate version of the system, it is the only authentication mechanism. It was judged as sufficient since the client also have to be physically present by the gate within the allocated time window. To enroll with the telephone version of PER, clients also have to enter a seven-digit authorization code at the beginning of the call. The code is issued by the web interface and presented to the client on an enrollment sheet that also includes the ten items to speak during enrollment. An enrollment sheet is not needed with gate enrollment since enrollment items are presented on the display.

5.6 Speech recognition

The automatic speech recognition (ASR) component of the system is based on the Starlite decoder (Ström, 1996) and acoustic models trained on the Swedish landline FDB5000 SpeechDat database (Elenius, 2000).

Acoustic models are state-tied triphone models created using the COST249 reference recognizer framework (Lindberg et al., 2000). Each triphone is modeled by three states and a mixture of eight Gaussian terms per state, with a total of 7623 states. The data set used for training the acoustic models is described in Section 10.2.1.

Input features are specified by the reference recognizer framework. They are 12-element MFCCs plus the 0'th cepstral coefficient and their first and second order deltas. MFCCs are similar to those used in the speaker verification system except that the filter bank has 26 filters spaced between 0-4000 Hz (cf. Section 3.2.1). Energy normalization and cepstral mean subtraction are not used.

The decoder uses a two-pass search strategy: a Viterbi beam-search followed by an A* stack-decoding search. A number of class-pair grammars are used to simulate a dialog-state dependent finite-state grammar. Output is an N-best list with up to 10 hypotheses, each specified by a word sequence and start and end times of each word segment. The application selects one hypothesis per utterance to be used as the segmentation of the utterance by the speaker verification system. The hypothesis is selected based on the knowledge of what the claimant is supposed to say, as the hypothesis with the highest score whose text matches the expected text. If there is no hypothesis with the expected text, the dialog system rejects the utterance and prompts the claimant for a new one. The system always knows what digits to expect from the claimant, since during both enrollment and test, digit sequences are prompted to the user.

5.7 Speaker verification

The speaker verification component of the PER system is a score-level combination of an HMM-based subsystem and a GMM-based subsystem. The entire speaker verification system was described in detail in Chapter 3.

5.8 References

Armerén, E. (1999). Site access controlled by speaker verification. Master's thesis, KTH/TMH, Stockholm, Sweden. [1]

Elenius, K. (2000). Experiences from collecting two Swedish telephone speech databases. *International Journal of Speech Technology*, 3:119–127. [7]

Lindberg, B., Johansen, F., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recog-

niser based on SpeechDat(II). In *Proc. 2000 International Conference on Spoken Language Processing (ICSLP)*, pages 370–373, Beijing, China. [7]

Nordqvist, P. and Leijon, A. (2004). An efficient robust sound classification algorithm for hearing aids. *The Journal of The Acoustical Society of America*, 115(6):3033–3041. [3]

Ström, N. (1996). Continuous speech recognition in the WAXHOLM dialogue system. *TMH-QPSR*, 37(4):67–96. [7]