Prompting of passwords in speaker verification systems

Håkan Melin, Johan Lindberg,
KTH, Dept. of Speech, Music and Hearing.

Abstract

The problem of how to prompt a client with a password in an automatic prompted speaker verification system is addressed. Text-prompting of four-digit sequences is compared to speech-prompting of the same sequences, and speech-prompting of four-digit sequences is compared to speech-prompting of five-digit sequences. Speech recordings are analyzed by comparing speaker verification performance and by inspecting the number and type of speaking errors that subjects made. It is found that text-prompting works better than speech-prompting for the four-digit sequences, given that speaker models were trained on text-prompted digits.

Introduction

Speaker verification systems can be classified as being text-dependent, text-independent or prompted. Systems of the prompted class work similarly to a text-dependent system but has the feature that the system prompts the client what to say each time the system is used.

There are two main reasons for wanting a speaker verification system to prompt the client with a new password phrase for each new test occasion: 1, the client does not have to remember a fixed password and 2, the system can not easily be defeated with the re-playing of recordings of the client's speech. In a telephony application the obvious way of prompting is by playing the password through the telephone with a prompting voice *(speech-prompting)*. An alternative approach would be to provide the client with a list of once-only passwords from which (s)he can read a password *(text-prompting)*. Which password to read could be the choice of the client himself or be indicated by the system. The latter approach might not be as convenient since the client must have the password list at hand, but has even greater security potential since an impostor who doesn't have the list has no way of knowing the correct password.

This paper addresses the problem of how to prompt the user with a password phrase by presenting two comparative experiments. In the first experiment (A), text-prompting a four-digit string is compared to speech-prompting the same string. The second experiment (B) then compares using four-digit and five-digit strings as the speech-prompted password. Each experiment is analyzed by looking at the number and type of speaking errors the subject made while saying the different passwords and by comparing the performance of an automatic speaker verification on passwords acquired in the different conditions.

The Speaker Verification System

The system used in the experiments has one left-to-right HMM for each digit (0-9). The HMM has two states per phoneme in the word (from two to four phonemes in Swedish words for digits) and two Gaussians per state. Speech parameterization is 12 LPCC coefficients plus an energy parameter, with appended delta and acceleration coefficients (total 39 parameters per frame). Cepstral mean subtraction is used to decrease inter-session variability. A world model with one HMM per digit and the same characteristics as the client model is used for log-likelihood normalization of the

score from the client model. A silence and garbage model (inter-word model) is shared by all client models and the world model.
When training the world and client models a word boundary segmentation of training sequences is needed. It is here assumed that an ideal segmentation component is available and this is simulated by using manual segmentations. At test-time the system makes its own segmentations given the prompted sequence as input, i.e. the system knows which sequence the client is supposed to say.
This system configuration is one of those that performed well in tests in the CAVE project reported in (Bimbot et. al., 1997). The system implementation used in the experiment is described in the same reference.

Database

Both experiments were conducted on the Gandalf database (Melin, 1996). Client models were built from 25 text-prompted five-digit sequences recorded in one session. Among these 25 sequences each digit occur at least twelve times and in all left and right contexts. The world model was built from such material from a separate set of *off-line* speakers, 15 male and 15 female speakers which are not used otherwise in the tests, neither as client nor impostor speakers. A silence/garbage model was trained on all between-word segments in the enrollment call of clients and off-line speakers. Silence model, world model and client models are the same in all experiments.
In experiment A, verification tests were made on pairs of text-prompted and speech-prompted versions of the same sequence. A pair was always recorded in the same telephone call and only pairs where both recordings contain precisely the requested four-digit sequence were used (recordings with repetitions of words or missing or additional words were sorted out through manual listening). Among the 1850 client test calls in Gandalf there are 455 such pairs which can be used for true-speaker tests, i.e. 455 true-speaker tests per prompt-type. For studying speaking errors on the other hand, all 1850 client test calls were used. For false-speaker tests, one pair from each of the client speakers plus one pair from each of 24 other speakers were used. This gives 109 false-speaker tests per prompt-type and target identity, and thus a total of 9374 false-speaker attempts per prompt-type.
In experiment B, verification tests were also made on pairs of sequences recorded in the same telephone call; one four-digit and one five-digit sequence in each pair. In Gandalf, five-digit speech-prompted sequences only exist for calls 19 and later (539 calls) and only a subset of 61 of the 86 client speakers have one or more recording of such a call. Data for false-speaker attempts were chosen similarly to experiment A. The number of speakers and tests used in each experiment is summarized in table 1.
The speech-prompts were synthesized with the KTH formant synthesizer (Carlson et al., 1991) for exact reproducibility of the prompting voice.

Table 1. The number of speakers and tests used in the verification test part of experiments A and B. The number of tests given are the number of tests per prompt-type in the respective experiment. The two numbers for the number of true-speaker tests in experiment B are with and without speaking errors.

Experiment	A	B
clients	86	61
additional speakers for false-speaker attempts	24	31
number of true-speaker tests	455	1070/952
number of false-speaker attempts per client	109	91.2
total number of false-speaker attempts	9374	5562

Results

Speaking and recording errors

Recording items used in the verification part of experiment A are those where the text in the recording is exactly that of the prompted text. This section will present some analyses of the remaining speech items divided into two groups: those where the password is complete and those where it is not. A password is here considered complete if the requested words are included in the recording and occur in the correct order. For the analysis of experiment A in this section, all test calls from the client speakers have been used.

Table 2 shows some observations from items where the password is not complete. It can be seen that in the text-prompted case some digits are spoken as numbers, e.g. *one two* spoken as *twelve*. Word substitutions for speech-prompted items are of two kinds: most of them are confusions between 1 and 6 and are due to a difficulty to hear which digits the speech synthesizer said. Note that the synthesized speech is being played through a telephone line. The other confusions are likely to come from "errors" in the short-term memory of the subjects. This is also the case for word-order errors.

The fraction of items with a difference between the contents of the recording and the prompted text, but still with a complete password, is 0.03 % for the text-prompted and 2.8 % for the speech-prompted items. 2.4 of the 2.8 % are due to the recording procedure, in which the speech-prompt was first played, followed by a short beep after which the recording started. The 2.4% are cases when the subject started speaking (before *or* after the beep), was somehow disturbed by the beep, and re-started saying the whole sequence. Most of those errors can be eliminated by changing the prompting procedure.

A detailed study of speech and recording errors for experiment B is not given here. Instead, verification results are given in the next section for the cases where those kinds of errors are included and excluded respectively. It can be noted that 1.2 % of the four-digit items and as much as 6.8 % of the five-digit items, recorded in the 539 calls where there are five-digit speech-prompted items, does not contain the complete password.

Table 2. Observations from 7400 text-prompted and 3700 speech-prompted four-digit items. Numbers are given as the fraction of the number of recorded items of the respective prompt-type. The numbers in each column sum up to the number given for "password incomplete". Bold-face numbers are referred to in the discussion.

	text-prompted	speech-prompted
password incomplete	0.67 %	3.8 %
digits spoken as number	**0.10 %**	
word substitution due to subject	**0.09 %**	**0.53 %**
word substitution due to synthesizer		2.25 %
wrong word-order		**0.62 %**
recording method	0.33 %	0.03 %
other	**0.15 %**	**0.38 %**

Speaker verification performance

Both experiments have been designed such that two sets of tests are compared. Table 3 shows gender-balanced sex-independent (GBSI) equal error rates (EER) (Bimbot and Chollet, 1995) for each of the two sets in the two experiments. When computing an EER the decision threshold is adjusted *a posteriori* to give equal false rejection and false acceptance rates within each test set. The threshold is adjusted individually for each client.

Table 3. Speaker verification GBSI-EER for each of the two sets in both experiments.

A	text-prompted	speech-prompted
	3.24 %	4.86 %

B	Number of digits	4	5
	including speaking errors	2.41 %	1.97 %
	without speaking errors	2.23 %	1.43 %

Discussion

In experiment A, the EER for text-prompted sequences was clearly lower than for speech-prompted sequences. One shall here keep in mind that the speaker models were trained on text-prompted speech. The result can be interpreted such that there is a difference in how subjects speak a phrase when it is given to him through text rather than speech. If the speaker models in the speech-prompted case were also trained on speech-prompted speech, the result would probably be another.

Of the sources of error included in table 2, some could be eliminated. In the speech-prompted case, the number of confused digits could be reduced by improving the prompting voice. In both cases, errors due to the recording method are mainly from the speech being cut off in the end, and those errors could be reduced by making the recording time window longer or by using robust speech detection. That leaves us with the bold-face numbers in table 2 which sum to 0.3 % in the text-prompted and 1.5 % in the speech-prompted case, a factor of five. This is a measure of how many times a prompted speaker verification system would have to give the client a new try just because the password was wrong.

In table 3.B five-digit sequences give lower EER than four-digit sequences, even when speaking and recording errors are included in the tests. Note that the large difference in EER for four-digit speech-prompted sequences in experiment A and B comes from the fact that A and B have very different test sets, and that of B include only calls from the favorite handset (Melin, 1996).

Conclusions

Two indications have been presented which show that text-prompting may work better than speech-prompting from a performance point of view. Firstly, the number of speech errors was larger for speech-prompted than for text-prompted four-digit sequences. Secondly, the verification error rate was also larger.

References

Melin H. 1996. Gandalf - A Swedish Telephone Speaker Verification Database. *Proceedings of the International Conference on Spoken Language Processing. ICSLP-96, Philadelphia, PA, USA*, 1954-1957.

Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J.-B. 1997. The CAVE Project: Caller Verification for Telephone Applications. *To be published in EUROSPEECH-97*.

Carlson R., Granström B., and Karlsson I. 1991. Experiments with voice modelling in speech synthesis. *Speech Communication* 10, 481-489.

Bimbot F. and Chollet G. 1995. Assessment of speaker verification systems. *In Spoken Language Resources and Assessment, EAGLES Handbook*.