EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

TEXT-PROMPTED VERSUS SOUND-PROMPTED PASSWORDS IN SPEAKER VERIFICATION SYSTEMS

Johan Lindberg and Håkan Melin
Department of Speech, Music and Hearing
KTH
S-100 44 Stockholm, Sweden.
Tel. +46 8 790 9269, Fax: +46 8 790 7854, E-mail: {lindberg,melin}@speech.kth.se

ABSTRACT

The problem of how to prompt a client for 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 digits is compared to speech-prompting of five digits. Speech recordings are analyzed by comparing speaker verification performance and by inspecting the number and type of speaking errors that subjects made. From the experiment it is clear that text-prompting gives the subjects an easier task and fewer speaking errors are produced in that context. When enrolling clients with text-prompted speech and performing verification with an HMM-based system, the average EER was larger for speech-prompted items compared to text-prompted items, but changes in individual EERs varies across the test population.

10.21437/Eurospeech.1997-287

1. 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 have the feature that the system prompts the client what to say each time the system is used [1].

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)*. Either the client or the system could decide which of the passwords to use. The text-prompting 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 does not 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 sequence is compared to speech-prompting the same sequence. The second experiment (B) compares using four-digit and five-digit sequences 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 system on passwords acquired in the different conditions.

2. SPEAKER VERIFICATION SYSTEM

An HMM-based system [2] was used in the experiments. Client models have one left-to-right HMM for each digit (0-9). Each HMM has two states per phoneme (there are between two and four phonemes in Swedish digit words) and two Gaussians per state. Speech is parameterized using 12 LPCC coefficients plus an energy coefficient, with appended delta and acceleration coefficients (totally 39 elements per frame). Cepstral mean subtraction is used to decrease inter-session variability. A world model with the same characteristics as the client models is used for log-likelihood normalization of the score from a client model. An inter-word model (silence and garbage) is shared by all client models and the world model.

When training the world and client models a word boundary segmentation of the training sequences is needed. It is here assumed that an ideal segmentation component is available and this is simulated by using manual segmentations. During the test session the system automatically makes its own segmentations given the prompted sequence as input, i.e., the system knows the sequence the client is supposed to say.

The system configuration is one of those that performed well in tests in the CAVE project reported on in [2]. The system implementation used in the experiment is described in the same reference.

The experiments were conducted on the Gandalf database [3], i.e., data were not collected during actual usage of the speaker verification system. In the database recording, speech-prompting was implemented by playing the prompt followed by a 100 ms beep sound. The recording started after the beep. Speech prompts were synthesized with the KTH TTS-system [4] to ensure exact reproducibility of the prompting voice. Text-prompting was implemented by printing digit strings on a form that the subject was reading. The individual digits were separated by a space to indicate that they should be read as digits and not as numbers. During a recording session, the four text-prompted items were always recorded before the speech-prompted items.

Experiment	A	B
clients	69	61
average number of true-speaker tests per client	5.9	15.5
total number of true-speaker attempts	405	947
additional speakers for false-speaker attempts	37	43
number of false-speaker attempts per client	105	103
total number of false-speaker attempts	7245	6283

Table 1. The number of speakers and tests used in the verification test part of experiments A and B. The numbers of tests per prompt type in the different experiments, are given excluding items with some speaking or recording error.

3. EXPERIMENT

Client models were built from 25 text-prompted five-digit sequences recorded in one session. In these 25 sequences each digit occurs at least twelve times and in all left and right contexts. The world model was built from similar material from a separate set of so called *off-line speakers*, 15 male and 15 female speakers who are not used otherwise in the tests, neither as client nor impostor speakers. The inter-word model was trained on all non-word segments in the enrolment call of clients and off-line speakers. The silence, world, 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 contained 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 1820 client test calls in Gandalf there are 455 such pairs, which can be used for true-speaker tests. Among those, 405 were chosen, so that to each client there is at least 4 true-speaker tests per prompt type. This selection gives 69 clients with on average 5.9 true-speaker tests per client. For false-speaker tests, one pair from each of the client speakers plus one pair from each of 37 other speakers were used.

The main goal of experiment B is to compare four-digit versus five-digit speech-prompted sequences. Since the test material for that comparison must be chosen differently (five-digit speech-prompted sequences are only recorded in the 17th and later test calls in Gandalf), the comparability between results from A and B run a risk of being lost. Therefore, text-prompted four-digit sequences were also included in experiment B. Hence, verification tests were made on triples of items recorded during the same telephone call, where each triple contains one text-prompted four-digit sequence plus one four-digit and one five-digit speech-prompted sequence. 61 client speakers have 8 or more such triples. The average number of triples per client is 17.5 including speaking errors and 15.5 excluding them. Data for false-speaker attempts were chosen analogously to experiment A; one triple per speaker was chosen with no items in the triple having speaking or recording errors. The number of speakers and tests used in each experiment is summarized in Table 1.

prompting by:	*text*			*speech*		
range of test calls:	*1-4*	*5-16*	*17-26*	*1-4*	*5-16*	*17-26*
speaking or recording error	1.5	0.86	0.74	15	5.9	2.4
passwords complete	0.20	0.38	0.28	10	1.4	1.2
recording method				8.81	1.02	0.47
other	**0.20**	**0.38**	**0.28**	**1.19**	**0.41**	**0.74**
password incomplete	1.33	0.48	0.46	5.1	4.5	1.2
digits spoken as number	**0.10**	**0.07**	**0.09**			
word substitution due to subject	**0.00**	**0.03**	**0.14**	**0.82**	**0.68**	**0.19**
word substitution due to synthesizer				3.07	2.73	0.46
wrong word-order				**0.41**	**0.75**	**0.56**
recording method	1.13	0.20	0.09	0.00	0.07	0.00
omitted word	**0.00**	**0.03**	**0.05**	**0.82**	**0.27**	**0.00**
other	**0.10**	**0.14**	**0.09**			
sum of bold-face factors	**0.20**	**0.27**	**0.37**	**2.05**	**1.70**	**0.75**

Table 2. Observations on four-digit items with some speaking or recording error. Numbers are given as the percentage of the number of recorded items of a prompt type. Rows with indented left column show a factorization into different kinds of errors. Bold-face numbers indicate errors that are considered systematically related to the prompt type, while other errors are more related to the particular implementation used when recording the Gandalf database.

For studying speaking and recording errors in text-prompted versus speech-prompted items, all available calls from the 61 subjects used in experiment B were used. Subjects with at least 20 recorded test calls were selected so that potential learning effect can be observed. The total number of calls for this speaking errors study is 1511, with four text-prompted and two speech-prompted items in each call.

4. RESULTS

4.1. Speaking and recording errors

Recorded items used in the verification part of experiment A, are those where the text content of the recording is exactly that of the prompted text. This section will present some observations on 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.

The division into three groups of calls (1-4, 5-16, 17-26) in Table 2 is somewhat arbitrary, but allows the observation of potential short and long term changes in error rate while subjects get more used to the prompting procedures. The last group was chosen to match calls used in experiment B where five-digit speech-prompted sequences are also available.

As can be seen in Table 2, the recording procedure with speech-prompting caused trouble initially. Subjects frequently started speaking but were somehow disturbed by the beep, and re-started saying the whole sequence. Most of those errors could perhaps be eliminated by removing the beep from the prompting procedure.

The lower part of Table 2 shows observations from items where the password is not complete. "Digits spoken as numbers" refers to cases like *one two* spoken as *twelve*, which naturally occur only in text-prompting.

A very large portion of the word substitution errors turned out to be confusions between digits 1 and 6. Since those errors are likely to come from misinterpretations of the prompting voice, they are separated from other word substitution errors in Table 2. Digits 1 and 6 are confused especially in the context *after the digit 6*, e.g., 6-1 ([sɛksɛtʰ]) was often perceived as 6-6 ([sɛksɛks]). Note that the synthesized speech was played through a telephone line and hence the high-frequency components of /s/ were attenuated. From the 17th test call the sequences with the pair 6-1 were no longer included in the pool of possible prompts and therefore the error rate for word substitution due to synthesizer decreased considerably.

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, however, that the proportion of five-digit items where the password is incomplete is as high as 6.8 %, to be compared to the 1.2 % for four-digit speech-prompted items in calls 17-26 in Table 2.

4.2. 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) [2,5] for each of the 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.

Figure 1 shows the distribution of individual EERs for the two sets in experiment A, where the EERs in the speech-prompted case tend to be more smeared. Figure 2 shows the distribution of changes in individual EERs when going from text-prompting to speech-prompting. For 35 % of the clients, the EER does not change, while for 26 % it is lower and for 39 % higher in the speech-prompted case. This indicates that while the increase in

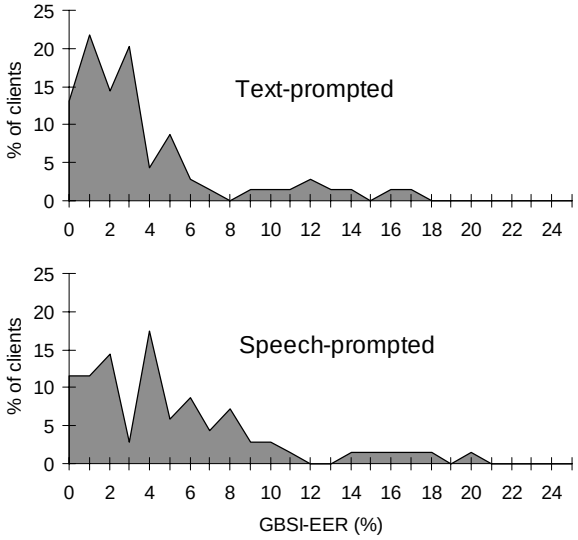

Figure 1. Distribution of individual speaker verification error rates (GBSI-EER) in the test sets in experiment A.

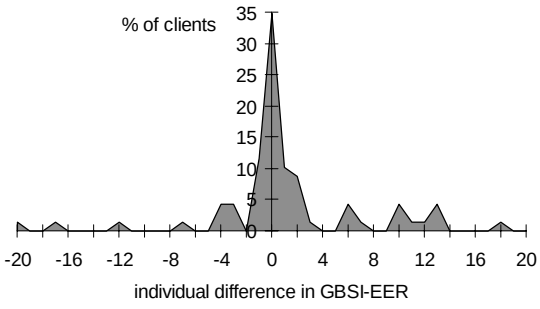

Figure 2. Distribution of differences in individual GBSI-EER when comparing the two sets in experiment A. A positive difference *x* indicates that the individual EER for speech-prompted items was *x* units higher than for text-prompted items.

Experiment	Number of digits / prompt type	4 / text-prompted	4 / speech-prompted	5 / speech-prompted
A	without speaking errors	3.24 %	4.86 %	
B	without speaking errors	1.99 %	2.38 %	1.55 %
	including speaking errors	2.14 %	2.57 %	2.08 %

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

average EER in Table 3 is substantial, the changes are not consistent over the client population.

Finally, note that the large differences in EER for four-digit text and speech-prompted sequences in experiments A and B (Table 3) come from the fact that A and B have very different test sets. B includes only calls from the same handset, the so called favorite handset [3], while A includes calls from many different handsets. The error rates in experiment A are therefore generally higher.

5. DISCUSSION

In experiment A, the average EER for text-prompted sequences was clearly lower than for speech-prompted sequences, though the spread among individual subjects is large. One should keep in mind, though, 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 different.

In experiment B, the addition of a fifth digit in the speech-prompted case lowers the EER, even when the large number of speaking and recording errors are included in the tests. The same improvement can be expected for text-prompted five-digit sequences, but without an increase in the number of speaking errors.

The system's apparent insensitivity to speaking errors may come from the use of relative scores and the fact that the system makes a forced alignment to the prompted text. When the spoken text is not the same as the prompted text, the scores of both client and world models are very low and more or less random. In this case the relative score is not a good decision variable. This effect was also observed in [1]. In the current system, this problem could be solved by performing an explicit check of the text contents in the response.

The portion of speaking errors is a measure of how often a speaker verification system would have to give the client a new try just because the password was wrong. The only systematic sources of speaking errors related to text-prompting seems to be reading disfluencies (included in the "other" group in Table 2) and digits pronounced as numbers. Observed error rates for both are very small relative to the EER of the verification system. For speech-prompting, observed speaking error rates are higher and of the same order of magnitude as the EER. In this case, speaking errors seem to come from two sources: either the subject does not hear the prompt correctly, or the short-term memory fails him and he repeats the wrong sequence, either with the wrong word order or with word substitutions.

6. CONCLUSION

From the study of speaking errors produced by subjects in response to text prompts and speech prompts respectively, it is clear that speech-prompting leaves the subject with a more difficult task and more speaking errors are therefore produced.

Experiments on text-prompted and speech-prompted passwords indicate that there is a difference in speech produced in response to the different prompt types, which affect the performance of the HMM-based speaker verification system.

7. ACKNOWLEDGEMENT

Johan Lindberg is supported by CAVE; EU Telematics programme, grant LE-1930. The work of Håkan Melin is supported by Telia Research AB.

8. REFERENCES

[1] Higgins A., Bahler L., Porter J., "Speaker Verification Using Randomized Phrase Prompting", *Digital Signal Processing* 1, pp. 89-106, 1991.

[2] Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J., and Pierrot J.-B., "Speaker Verification in the Telephone Network: Research activities in the CAVE Project," *Proc. EUROSPEECH'97,* Rhodes, Greece, 1997.

[3] Melin H., "Gandalf - A Swedish Telephone Speaker Verification Database," *Proc. ICSLP-96,* pp. 1954-1957, Philadelphia, USA, 1996.

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

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