KTH Computer Science
and Communication

10. PER experiments

HÅKAN MELIN

KTH/CSC/TMH
Stockholm, Sweden 2007

This is a self-contained re-print of Chapter 10 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.

© Håkan Melin, 2006, 2007

Contents

Chapter 10

PER experiments

10.1 Introduction

This chapter reports on findings from an evaluation of the on-site and telephone versions of the PER system described in Chapter 5. The evaluation was conducted with speech data collected through actual use of the two system versions. The data collection and data themselves were described in Section 6.3.

All our development of the speaker verification component of the PER system before the collection of evaluation data was made using general purpose telephone corpora Gandalf (Melin, 1996), Polycost (Hennebert et al., 2000), SpeechDat (Elenius, 2000) and Switchboard, since the Department's research was directed on telephone applications of speaker verification (Lindberg and Melin, 1997; Melin, 1998; Melin et al., 1998; Nordström et al., 1998; Melin and Lindberg, 1999; Bimbot et al., 1999, 2000; Neiberg, 2001). Hence, the system used to collect live evaluation data was not optimized for the particular application it was used in. However, in parallel to collecting evaluation data, separate, application-specific development data were collected allowing for off-line simulation experiments with an optimized system. In this chapter, results are presented both for the initial, general-purpose system and the optimized, application-specific system.

Besides the variants of our own research system, a commercial speaker verification system has also been tested on the collected corpus. Results from these tests serve as calibration of the data and the recognition tasks.

Results from practical use of ASV technology for person authentication in on-site applications have been reported in several publications. Test sites include Texas Instruments corporate headquarters in Dallas (Doddington, 1985), Siemens in Munich (Feix and DeGeorge, 1985) , LIMSI in Paris (Mariani, 1992), AT&T Bell Labs in cooperation with a large bank (Setlur and Jacobs, 1995), Fraunhofer Institute in Erlangen (Wagner and Dieckmann, 1995), University of Frankfurt (Schalk et al., 2001) and Panasonic Speech Technology Laboratory in Santa Barbara (Morin and Junqua, 2003). At AT&T Bell Labs the application was an automated teller ma-

chine (ATM), while at all other sites it was a voice-actuated lock that secured access to a physical room or building.

At Texas Instruments, a template based system was installed in the mid 1970s (Rosenberg, 1976; Doddington, 1985). It was aurally text prompted using strings of four words like "Proud Ben served hard", and used a sequential decision strategy where claimants were asked to speak new word sequences until a certain level of confidence was achieved. False reject and false accept rates (casual impostors) of below 1% are reported with on average 1.6 utterances required by the sequential decision strategy. Users were required to step into a booth to use the system.

At LIMSI, a text-dependent, template based system was first publicly demonstrated in 1985. It was installed in a voice-actuated door lock application at the lab in 1987 and was used by about 100 users (Mariani, 1992). A second generation system was installed in 1990 and a new generation, HMM-based system was developed in 1997 which has so far only been used for data collection (Lamel, 2005).

At Panasonic Speech Technology Laboratory in Santa Barbara a biometric terminal has been in service since April 2002 by the building's main entrance door (Morin and Junqua, 2003). It is a multi-modal access control system where any of the three modes speech, fingerprint or keypad (10-digit account number) can be used individually, or in combination for uncertainty recovery. The speech sub-system is template based and operates on user-selected pass-phrases in an open-microphone mode. Users can speak the pass-phrase at any time from within typically 0.3–3 meters from the terminal. The system has been in use by about 35 enrolled users and was reported to have about 8% FRR and 0.1% FAR (2.8% EER) for the speech mode only. Some of the initial rejections were recovered via another mode reducing the FRR to about 5%. Other results using data collected by this system have been reported in (Bonastre et al., 2003).

AT&T conducted a six month field trial with an ATM application where a text-prompted, HMM-based speaker verification system was used in addition to regular PIN codes typed on a keyboard (Setlur and Jacobs, 1995). Claimants were asked to repeat random 4-digit phrases into a handset connected to the ATM.

10.2 Development tests

This section describes what data was used for developing the PER system and how it was used.

Table 10.1 shows results from the development experiment to determine empirical values for weights ω_ξ used in combining scores from the HMM and GMM subsystems (Eq. 3.23).

The value of the decision threshold θ (Eq. 3.25) was also determined empirically as the same-sex EER threshold with the combined ASV system on the same development test set.

Table 10.1: Equal error rate ϵ_ξ, standard deviation σ_ξ of score distribution and combination weights ω_ξ for the HMM and GMM subsystems as determined from a development experiment on Gandalf data.

subsystem (ξ)	EER (ϵ_ξ)	stdev (σ_ξ)	weight (ω_ξ)
HMM (H)	7.51%	4.017	0.142
GMM (G)	6.11%	0.6747	0.858

10.2.1 Development data

Most experiments behind development decisions in the design of the HMM subsystem were done on various partitions of the Gandalf (Melin, 1996) and Polycost (Hennebert et al., 2000) corpora, e.g. (Melin and Lindberg, 1999) and (Nordström et al., 1998). With particular development for the PER application in mind, a PER-like development test configuration on Gandalf was created. It was used to optimize the configuration of the GMM subsystem and to determine the *a priori* score fusion weights and the decision threshold used during data collection.

The PER-like development test configuration uses one of two fixed sentences in place of names. Half of the target speakers were assigned one sentence and the other half the other sentence. Enrollment was performed with 10 repetitions of the sentence and 10 five-digit sequences taken from two recording sessions from different handsets (enrollment set d5+fs0x, cf. Table 6.8), while each test was performed with a single repetition of the same sentence and an aurally prompted string of four digits (test set 1fs+1r4-fs0x, cf. Table 6.9). All impostor tests used in development experiments were same-sex attempts. True-speaker test sessions were recorded from up to 10 different handsets per target, but at least half of the sessions came from one of the target's enrollment handsets. Impostor test sessions were generally *not* recorded from one of the target's enrollment handsets. Even though this development test configuration was designed to simulate the PER application as well as possible given the constraints of the already existing Gandalf corpus, it differs in several aspects from real PER data as summarized in Table 10.2 for the telephone version of PER. The on-site version of PER naturally adds the differences already identified between the two PER versions (Table 5.1).

Background models were trained on subsets of files from 960 speakers in the Swedish landline FDB5000 SpeechDat corpus (Elenius, 2000). Background models in the HMM subsystem were trained on a *digits* subset composed by five files per speaker that may contain pronunciations of isolated or connected digits (corpus and item identifiers with parentheses): a random 10-digit sequence (`B1`), a 10 or 7 digit prompt sheet number (`C1`), an 8-12 digit phone number (`C2`), a 16-digit credit card number (`C3`), and a 6-digit PIN-code (`C4`). Background models in the GMM subsystem were trained on a *mixed* subset composed by six files per speaker: a random 10-digit sequence (B1), three phonetically rich sentences (`S1-S3`), and two phonetically rich words (`W1`, `W2`). None of the 960 speakers occur in the Gandalf or PER corpora.

Table 10.2: Main differences between the PER-like development set on Gandalf and telephone subset of PER evaluation data.

Aspect	Gandalf development	PER evaluation
Elicitation	recording	use of ASV system
Enrollment data	two session, two different handsets	single session
Test data (per target)	multiple handsets; cross-handset impostors	single handset; same-handset impostors
Impostors	random pseudo-impostors	dedicated impostors
Vocabulary	sentence+digits	proper name+digits
Passphrase variation	1 sentence/20 targets	1 name/1 target

Acoustic models for speech recognition were trained on 4016 speakers (gender-balanced) in the referred SpeechDat corpus, including all files from each speaker with the exception of files transcribed with truncated signal, mispronunciations, unintelligible speech or phonetic letter pronunciations (Lindberg et al., 2000). The number of used speakers is less than 5000 because 500 speakers were withheld for testing, 37 more because they were included in the Gandalf corpus, and 10% of the remaining speakers were set aside for development testing. Hence, there is no speaker overlap between this data and the Gandalf data. There is also no speaker overlap between used SpeechDat data and PER data. The total duration of speech segments in this training data is approximately 120 hours.

Six of the subjects (`M1003`, `M1005`, `M1015`, `1032`, `F1025` and `F1031`) in the PER test group participating as clients (five in group E and one (`F1031`) in group L) and impostors are also included in the development set of the Gandalf corpus, together with three subjects (`M1002`, `M1166` and `F1009`) participating as impostors only in the PER collection (one with gate-only data, the other two with gate and telephone data). The unfortunate overlap between subjects in PER evaluation data with respect to Gandalf development data is thus 11% of the 54 clients and 9% of the 98 impostors in the PER gate-only test set and 19% of the 27 clients and 16% of the 51 impostors in the condition-parallel test sets. More details about subjects participating both in Gandalf and PER can be found in Section 6.4.

Table 10.3: Statistics on the average number of attempts per enrollment item and gross duration (*minutes*:*seconds*) of enrollment sessions, based on complete enrollment sessions included in enrollment sets E2a_*c*.

Condition, *c*	#Sessions	Attempts			Duration		
		Min	Avg	Max	Min	Avg	Max
gate/hall	54	1.1	1.9	4.5	1:42	3:41	8:44
landline/office	54	1.0	1.1	2.1	1:35	2:18	4:32
mobile/office	29	1.0	1.2	1.7	1:49	2:26	3:48
mobile/hall	29	1.0	1.4	3.9	2:02	3:08	8:28

10.3 Field test results

10.3.1 Enrollment

During the data collection period, 56 subjects started enrollment. 54 of them succeeded to complete the enrollment sessions they were asked to do (enrollment in two conditions for client group L and four conditions for client group E). Table 10.3 shows statistics on how many attempts per item they made and the total duration of the sessions. Durations are measured from session start to completed enrollment, including time for system prompts, system delays, etc. For all telephone enrollment sessions this includes the entry of a 7-digit enrollment code by voice for authorization, and for sessions from a mobile phone it also includes a sub-dialog to determine if the call was made from the office or the hall. Attempts statistics are based on the average number of attempts per enrollment item and session, e.g. 1.1 in the Min-column for the gate/hall condition means the session with the least number of attempts had 11 attempts total since there were ten items. Attempts are counted from the system point of view, disregarding whether users actually made an attempt to speak an enrollment item or not.

In the longer enrollment sessions, users typically experienced problems with a few of the enrollment items, which they had to repeat many times before the speech recognizer was able to recognize their utterance correctly, or they opted to skip the item. The skip-possibility was introduced as described in Section 5.5 as an attempt to limit user frustration in these cases and to allow the enrollment process to be completed despite such problems. Within the enrollment sessions that were eventually completed, eight subjects (15%) skipped one item and one subject (2%) skipped two items in the gate/hall condition, while a single subject (2%) skipped one item in the landline/office condition. In the two mobile conditions, no items were skipped.

39 of the 54 subjects who completed their requested enrollment sessions (72%) completed all their enrollment session at the first attempt, while the remaining 15 (28%) had one or more failed or aborted enrollment sessions before the complete ones. In failed sessions for nine of the latter, the actual enrollment procedure was never started because either subjects had not enabled enrollment through the

intranet or the enrollment window had expired (four cases); their name was incorrectly recognized (seven cases); or they did not have the enrollment code available (two cases). Six subjects terminated the enrollment procedure of one or more enrollment sessions pre-maturely. Four of the six terminated one session each (three in the gate/hall and one in the mobile/hall condition), probably after feeling disturbed by other people passing through the gate or otherwise making noise in the hall. One of the six, it appeared, had removed his last name through the web interface so the ASR grammar contained only his first name while he was still speaking his full name. After correcting this, his enrollment sessions were immediately successful. The last of the six had severe problems with getting the speech recognizer to recognize his utterances. He terminated three enrollment sessions in the gate/hall condition and one in the landline/office condition before succeeding with enrollment. The source to the system's problems with this subject appears to have been a combination of the subject being a non-native speaker of Swedish and him speaking very loudly to the system.

The remaining two of the 56 subjects (3.6%), one male and one female subject, failed to complete any enrollment session. The female subject, a non-native speaker of Swedish, tried to enroll in both the gate/hall condition and the landline/office condition, with similar results in both cases: The speech recognizer consistently failed on digit sequences including the digit 7, probably caused by her non-native pronunciation of this digit (a typical Swedish pronunciation would be [ɧʉ:] or [ʃʉ:]). The male subject terminated his first (and only) enrollment session in the gate/hall condition after being disturbed by noise from other people passing through the gate at the time. He suggested to try another time, but never did so.

10.4 Simulation results

Results in this section are from off-line simulations of speech recognition and speaker verification operations using the PER corpus (with recordings from actual use of the PER system; cf. Section 6.3). Results are presented in terms of DET curves and EER.

10.4.1 Baseline system

The original speech recognition and speaker verification components of the PER system (as described in Chapter 5) used to collect data, without the use of a speech detector, is designated as the baseline system.

Results for the baseline system in the gate/hall and landline/office conditions using E2a_*c* and S2b_*c* enrollment and test sets for the respective condition, are shown by the dashed DET curves in Figure 10.1. EERs are 6.4% for gate/hall and 4.0% for landline/office. Error rates are lower in the telephone case as was expected since both acoustic models (ASR) and background models (ASV) were developed on telephone data. However, test sets S2b_G8 and S2b_LO are not

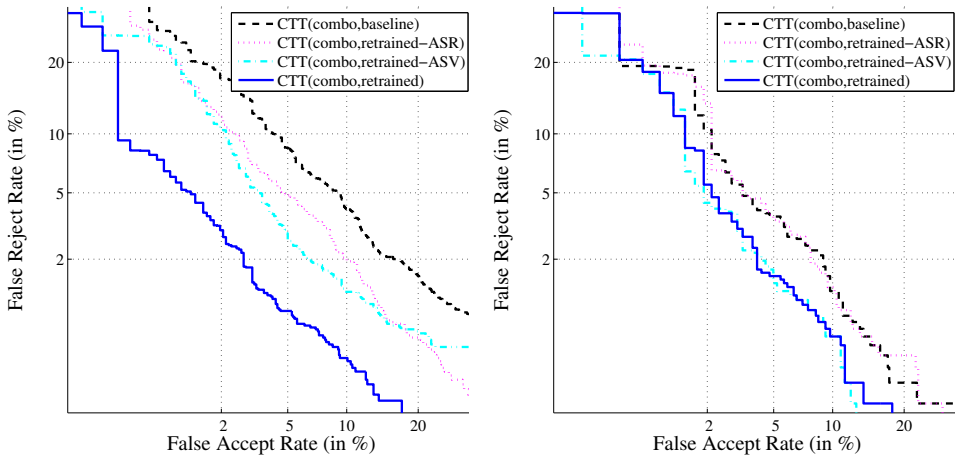


Figure 10.1: DET curves for baseline and retrained systems in the a) gate/hall (S2b_G8) and b) landline/office (S2b_LO) conditions. Baseline is with the original speech recognition and speaker verification components used during data collection, while in the retrained case both components have been adapted to condition-dependent data from background speakers. The remaining two plots show results where only one of the speech recognition or speaker verification components has been adapted.

directly comparable since they are based on different number of subjects, etc. A more fair comparison is shown in Figure 10.2 using the condition-parallel test sets S2b_Q:*c*. The comparison is more fair because, firstly, every test in a given condition has a corresponding test in all other conditions (Section 6.3.4.4), and secondly, a name and four digits is used per test in all four conditions, with a digit in a random position having been omitted from every test in the gate/hall data. Figure 10.2 confirms the lower error rates for landline/office than for gate/hall, however with a smaller difference than in Figure 10.1, even though one digit less per utterance is used in the gate/hall condition.

Figure 10.2 also indicates the operating points corresponding to the *a priori* decision threshold determined using the EER point on the Gandalf development test configuration. The resulting operating points are near the *a posteriori* EER point in the telephone conditions, while it is clearly far-off in the gate/hall condition.

10.4.2 Retrained system

Models of the original (baseline) PER system were adapted to PER-specific data from background speakers to create new, *retrained*, condition-dependent systems. These systems are expected to perform better in the PER application than the

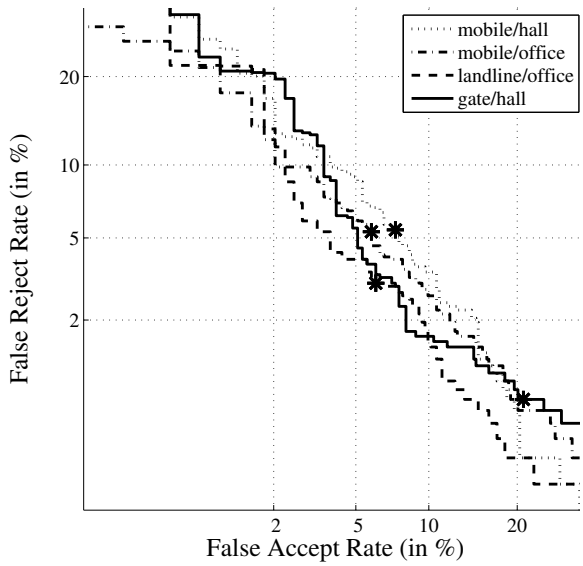


Figure 10.2: A comparison between conditions using the condition-parallel test sets (S2b_Q:c) and the baseline system. A name plus four digits is used in all conditions. EERs are 6.4% (MH), 5.8% (MO), 4.3% (LO) and 5.1% (G8). Asterisks (*) mark the operating points determined by the *a priori* threshold.

baseline system, but since background data was collected in parallel to evaluation data, the retrained systems have only been tested using off-line simulations on recorded data.

Acoustic models in the speech recognition component were not only trained on the new data, but their structure were also changed in two respects: models were made gender-dependent and the number of terms in the Gaussian mixture was reduced from eight to four. The new models were created with the following procedure. Gender-independent models with four terms per state were created with the same procedure as the original (eight-term) models. The four-term models were then cloned into male and female gender-dependent models and background speaker files were tagged as male or female. Mean vectors of the gender-dependent models were then adapted to the new data by a gender-independent Maximum Likelihood Linear Regression (MLLR) transform followed by a single MAP iteration using HTK (Young et al., 1999). The MLLR transform was made with a single transformation matrix for both male and female models, while the MAP adaptation was made with gender-dependent data.

Background models of the speaker verification component (both the HMM and GMM subsystems) were adapted to new data with three iterations of the EM-algorithm and the ML criterion, updating means, variances and mixture weights.

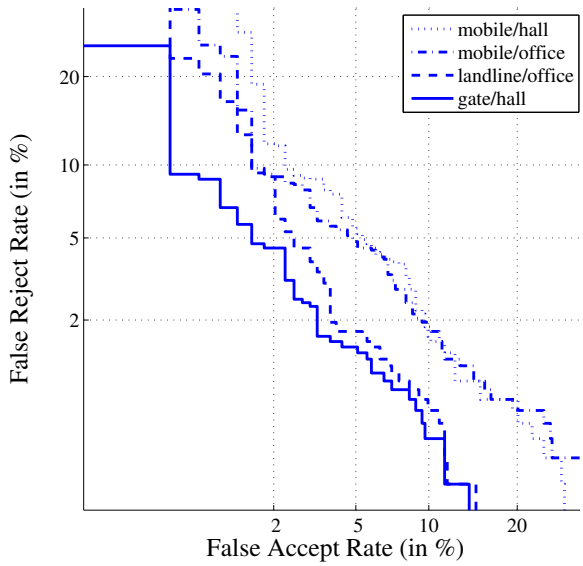


Figure 10.3: A comparison between conditions using the condition-parallel test sets (S2b_Q:c) and the retrained, condition-dependent systems. A name plus four digits is used in all conditions. EERs are 5.3% (MH), 4.8% (MO), 3.5% (LO) and 2.6% (G8).

Original models were used as the starting point for the first iteration. The HMM subsystem was trained on the digits subset of background speaker data, and the GMM subsystem on the name and digits subset.

The solid lines in the DET plots of Figure 10.1 show results for the retrained systems where both speech recognition and speaker verification components have been retrained, while dotted and dash-dotted lines indicate the contribution from retraining the individual components. The figure shows that adapting the speech recognition component improves performance considerable in the gate/hall condition while no effect can be seen in the landline/office condition, while adapting background models in the speaker verification component reduces error rates in both conditions. EER for the solid lines in Figure 10.1 is 2.4% for gate/hall and 3.1% for landline/office. This corresponds to a 63 % relative reduction in EER for gate/hall compared to baseline, and 23% for landline/office. Figure 10.3 shows DET curves for the condition-parallel test sets and the retrained systems.[1] Note that with the retrained systems, performance is better in the gate/hall condition than in the landline/office condition.

[1]See also 7.15 for an alternative comparison using parametric DET curves.

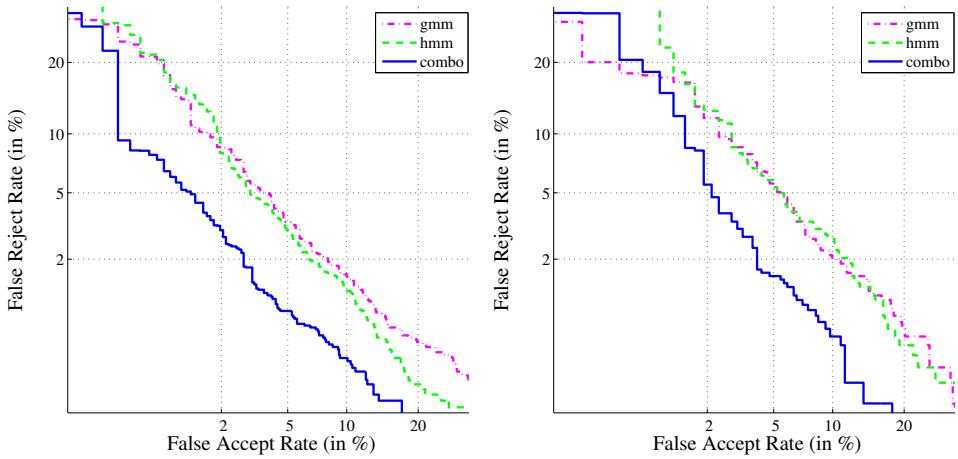


Figure 10.4: DET curves for the retrained system and its individual subsystems in a) gate/hall and b) landline/office conditions.

10.4.3 Fusion

Figure 10.4 shows DET curves for the individual HMM and GMM subsystems along with the combined system, all retrained on PER background speakers. Score combination weights are the *a priori* weights computed on Gandalf data. Clients are enrolled using the full enrollment session (E2a_*c*) and test sets are the single-condition test sets S2b_*c*. The GMM and HMM subsystems exhibit similar error rates in both the gate/hall and landline/office conditions, but note that the GMM subsystem uses more speech data than the HMM subsystem since it uses both the name and the digits. EER in the gate/hall condition is 4.2% and 4.0% respectively for the GMM and HMM subsystems and 2.4% for the combined system; 5.2% for both subsystems and 3.1% for the combined system in the landline/office condition.

10.4.4 Enrollment length

All of the above results were produced using target models trained on the full enrollment session represented by enrollment sets E2a_*c*. This includes 10 repetitions of name and digits for most targets, and 8 or 9 repetitions for a few targets where one or two enrollment utterances were skipped (cf. Section 10.3.1, p. 5). Figure 10.5 compares these results to the case with half of the enrollment data, exactly five repetitions per target, and the retrained systems. Note that background models were the same in both cases. They were trained on full enrollment sessions from each background speaker. EER is 2.4% and 5.3% in the gate/hall condition and 3.1% and 8.8% in the landline/office condition, i.e. the EER is more than doubled

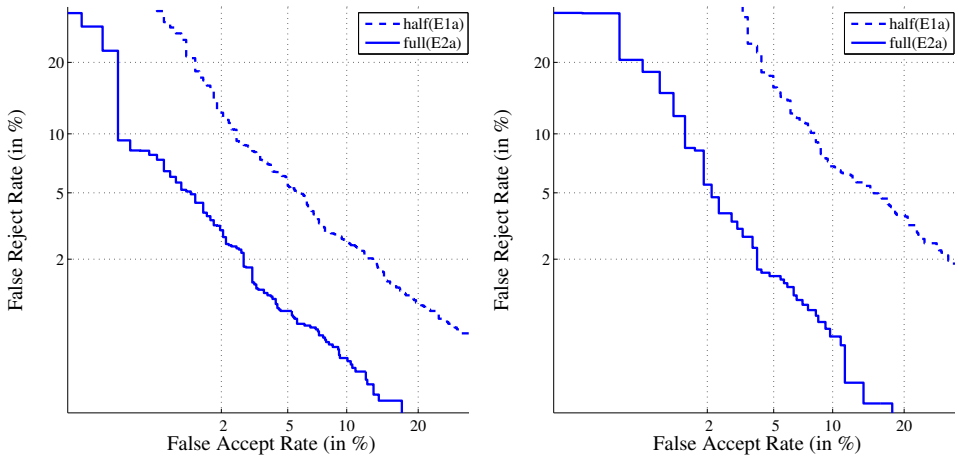


Figure 10.5: Client enrollment using the full enrollment session (E2a_c) and the first half of it (E1a_c) with the retrained system for the a) gate/hall and b) telephone/office conditions. Test sets are the single-condition sets S2b_c.

in the former condition with the reduction in enrollment data, and almost tripled in the latter condition.

10.4.5 Test utterance length

Test utterances collected in the gate/hall condition contain name plus five digits, while those in telephone conditions contain one digit less. Results for single-condition test sets are based on those test utterances directly, and thus the gate/hall condition has a slight advantage over telephone conditions, offered by the use of a display to prompt passphrases. To focus on speaker verification system performance in comparison between conditions, results on condition-parallel test sets in this chapter are produced with one digit removed from every test utterance in the gate/hall condition (with the exception of Figure 10.8 where all five digits were used with the commercial system).

Figure 10.6 displays the effect of the test utterance length directly. In the gate/hall condition, it compares DET curves for the retrained system with a name and two, three, four or five digits. To produce test utterances with less than five digits, digits in one or more random positions within each test utterance have been ignored in the feature vector stream, i.e. delta parameters in feature vectors were computed from the complete waveform to avoid discontinuities. The EER increases from 2.4% for the full test utterance to 2.9%, 3.2% and 4.0% when dropping one, two and three digits, respectively. These results for test utterances with less than five digits should be interpreted as approximate estimates of error rates for real

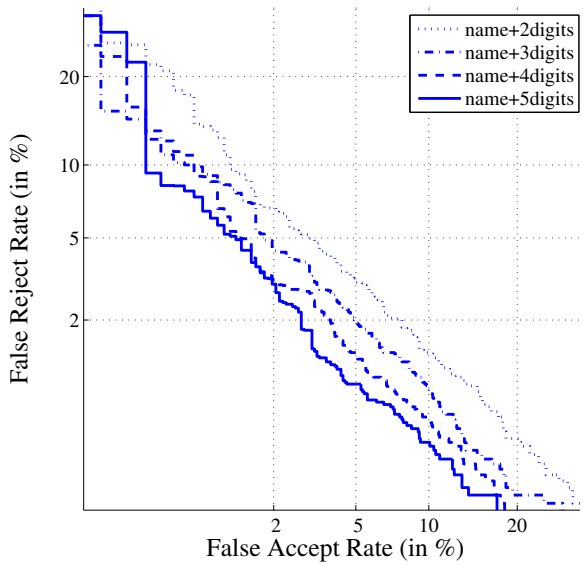


Figure 10.6: DET plots for the retrained system with the gate/hall single-condition test set S2b_G8 and including the name plus two, three, four or five digits in each test.

test utterances with the same number of digits, since synthetic short digit string utterances created by omitting digits from a longer utterance cannot be expected to be exactly equivalent to corresponding real utterances. Naturally, the prosody of the synthetic utterances will not be correct, but it may also be that digits in short strings are pronounced more clearly than longer strings. However, we believe the influence on presented results is small because the ASV system does not explicitly model sentence prosody or word context dependency.

10.4.6 Commercial system

Figure 10.7 shows DET curves for the commercial system for the single-condition test sets S2b_*c* and the gate/hall and landline/office conditions. Results are presented with the full and half session enrollment. EERs are 6.8% and 8.4% in the gate/hall condition and 6.0% and 7.6% in the landline/office condition (24% and 27% relative increase in EER for the two conditions with the reduction in enrollment data). Operating points marked with asterisks in the figure correspond to the EER-threshold determined from the Gandalf development experiment.

A comparison to Figure 10.5 shows that the commercial system performs better with less enrollment data relative to the retrained research systems.

Figure 10.8 compares the four conditions with the commercial system with full-

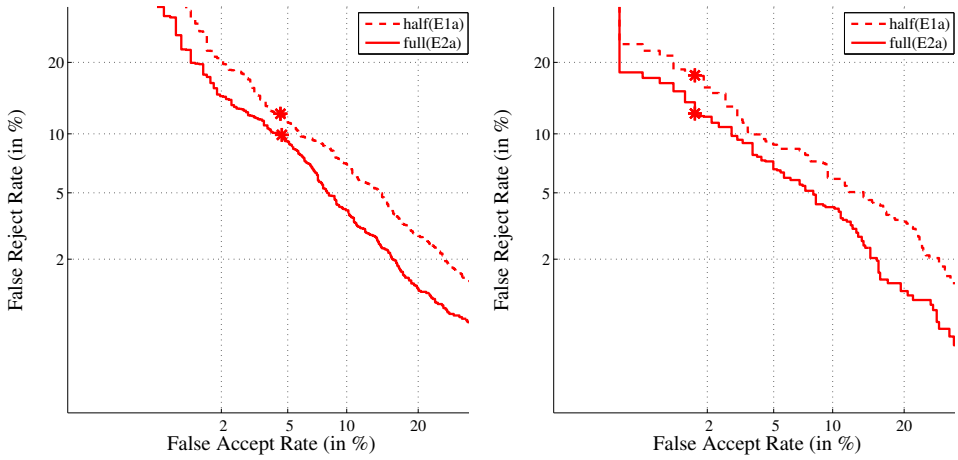


Figure 10.7: DET plots for the commercial system and client enrollment using the full enrollment session (E2a_*c*) and the first half of it (E1a_*c*) for the a) gate/hall and b) telephone/office conditions. Test sets are the single-condition sets S2b_*c*. Asterisks (*) mark the operating points determined by the *a priori* threshold. The speaker adaptation feature is turned off.

session enrollment and condition-parallel test sets. It also includes operating points determined from the EER point on Gandalf development data. As for the baseline research system (Figure 10.2), the operating point for the gate/hall condition is further to the lower right relative to those for the telephone conditions. However, all four points are shifted to the upper left compared to the same system.

10.4.6.1 Speaker adaptation

The commercial system has a speaker adaptation feature that allows a target model to be adapted to a test utterance if the verification score is greater than an adaptation threshold. Figure 10.9 shows DET curves for the commercial system on single-condition test sets (S2b_*c*) with tests run in a random order, the full enrollment session (E2a_*c*), and with the adaptation feature turned on. Since the adaptation threshold is specified relative to the decision threshold, an ideal decision threshold for the EER point was determined *a posteriori* for each condition from a previous run on the exact same test data with the adaptation feature turned off. This decision threshold was then used together with the default value on the adaptation threshold. EER with adaptation turned on is 3.2% in the gate/hall condition and 4.0% in the landline/office condition. This is a 53% relative reduction in EER for gate/hall and 27% for landline/office, compared to not using adaptation.

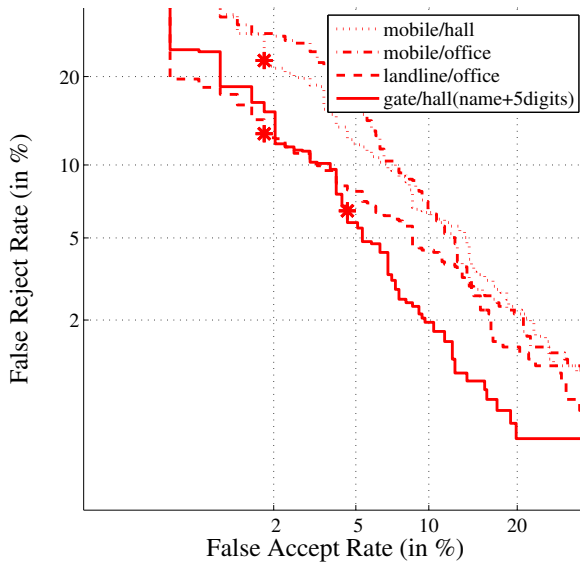


Figure 10.8: A comparison between conditions using the commercial system without speaker adaptation, full enrollment sessions (E2a_*c*), and the condition-parallel test sets (S2b_Q:*c*). A name plus four digits is used in telephone conditions and name plus five digits in the gate/hall condition. EERs are 8.4% (MH), 8.7% (MO), 6.4% (LO) and 5.3% (G8). Asterisks (*) mark the operating points determined by the *a priori* threshold.

With speaker adaptation the order of tests is relevant (e.g. Fredouille et al., 2000). In Figure 10.9 two cases were tested: *random* where all tests were run in a random order, and *optimistic* where all true-speaker tests were run before any impostor test. The latter case is an idealized situation for a speaker verification system, and was meant to estimate a lower bound on error rates with speaker adaptation. However, it turned out in Figure 10.9b that error rates are lower with the random order test than with the optimistic.

Table 10.4 shows how many of true-speaker and impostor tests resulted in a model adaptation (for each test the name and digits file were concatenated to form a single file per test).

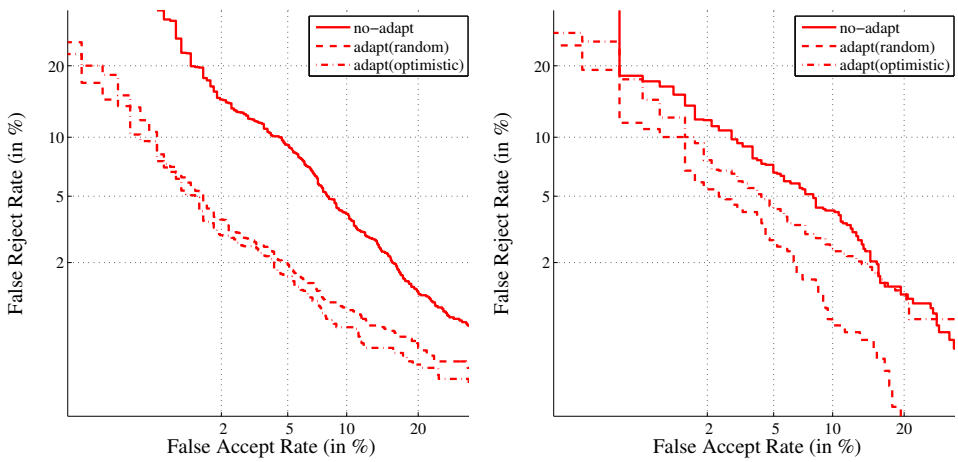
10.5 Discussion

10.5.1 Statistical significance

Table 10.5 summarizes EERs found in this chapter in the gate/hall and landline/office single-condition test sets. Table 10.6 show corresponding results for

Table 10.4: Proportion of true-speaker and impostor tests that resulted in model adaptation and the corresponding false reject (FRR) and false accept rates (FAR) in the experiments presented in Figure 10.9.

Cond.	Adapt	Test order	True-speaker tests	Impostor tests	FRR	FAR
G8	off	-	-	-	6.7%	6.8%
	on	random	97.4%	3.3%	0.75%	12.8%
	on	optimistic	98.1%	4.2%	0.39%	14.9%
LO	off	-	-	-	6.0%	6.2%
	on	random	96.7%	4.5%	0.57%	13.8%
	on	optimistic	96.2%	6.0%	1.55%	16.4%



a) b)

Figure 10.9: DET plots for the commercial system with and without its speaker adaptation feature turned on for the a) gate/hall and b) telephone/office conditions.

condition-parallel test sets. However, to allow the computation of post-trial confidence intervals (CI) based on Section 2.5.2, where we only considered false reject rates, we present EERs as if they were FRRs and compute CIs for the FRR. Basing a CI on an FRR this way is not statistically sound since EER is based on an *a posteriori* decision threshold determined after observing all the true-speaker test scores we are analyzing, plus a series of impostor test scores. The *a posteriori* threshold introduces a dependency between observations of decision errors. By treating observations of EER as observations of FRR we have basically assumed that previous experiments on development data resulted in a threshold that exactly meets an EER criterion on evaluation data (this was obviously not the case in most of our experiments). We claim the method still gives an idea of the uncertainty in our results.

The tables present 95% post-trial CIs for the "true" overall false reject rate given an observation of a fraction of errors $\hat{p} = x/N$ using two different methods. In both methods, intervals are computed from the binomial distribution as defined by 2.15. The methods differ in how the value for N in the binomial distribution 2.14 is determined:

- **Method 1**: N equals N' as determined by Eq. 2.18, i.e. such that the variance $\hat{p}(1-\hat{p})/N'$ of the fraction of errors predicted by the binomial equals the variance $s_{\hat{p}}^2$ in the estimate of $\hat{p}$ estimated from Eq. 2.22. In the computation of $\hat{s}_{\hat{p}}^2$, false reject rate p_i for target i is simply the fraction of errors observed for this target (defined as the non-parametric ML method in Eq. 7.1; $p_i = \mathrm{FRRd}(i)$). This is the "best practice" approach suggested by Mansfield and Wayman (2002), but we use the binomial directly to compute intervals, instead of its normal approximation.

- **Method 2**: N is fixed for a given test set and equals N' computed according to Eq. 6.1 with $\rho = 0.2$.

Since the variance estimation step in Method 1 can be viewed as a way to determine ρ, the resulting values of ρ are included in the tables. After estimating the variance $\hat{s}_{\hat{p}}^2$ with 2.22, ρ was computed relative to the adjusted total number of tests N^* (defined in Section 6.3.4.6) using 2.20 and 2.19 with N substituted by N^* and n substituted by $\lfloor \bar{n} \rfloor^*$, i.e. by solving for ρ in

$$1 + (\lfloor \bar{n} \rfloor^* - 1)\rho = \frac{N^* \hat{s}_{\hat{p}}^2}{\hat{p}(1-\hat{p})}. \tag{10.1}$$

Figure 10.10 shows the binomial distributions behind the confidence intervals for the retrained research (combo) system and single-condition test sets. Appendix F provides corresponding plots for condition-parallel test sets for the retrained combo system (Table F.3) and for the baseline research system and the commercial system (Table F.4 and F.5).

Table 10.5: Summary of observed FRR (%) with 95% confidence intervals computed from the binomial distribution with Methods 1 and 2 to select N. In all cases, the threshold equals the *a posteriori* EER (EERd) threshold. Systems above the dashed line within each method section of the table have been retrained on PER-specific, condition-dependent background data.

Test set	T2b_G8			T2b_LO		
ASV system	FRR	interval	ρ^a	FRR	interval	ρ^b
<Method 1>						
Combo, retrained						
- full enrollment	2.4	(1.2–3.6)	0.066	3.1	(0.0–6.1)	0.285
- half enrollment	5.3	(3.5–7.2)	0.076	8.8	(3.8–14.3)	0.267
GMM, retrained						
- full enrollment	4.2	(2.4–6.2)	0.101	5.2	(1.8–10.2)	0.259
HMM, retrained						
- full enrollment	4.0	(2.4–5.8)	0.093	5.2	(2.1–8.7)	0.141
Combo, baseline						
- full enrollment	6.4	(3.8–9.3)	0.156	4.0	(0.8–7.8)	0.240
Commercial system						
- full enrollment	6.8	(4.0–9.9)	0.181	6.0	(1.1–11.2)	0.320
- half enrollment	8.4	(5.1–11.9)	0.214	7.6	(1.6–14.3)	0.463
<Method 2>		$\rho = 0.2$ ($k = 10.8$)			$\rho = 0.2$ ($k = 8.8$)	
Combo, retrained						
- full enrollment	2.4	(0.8–4.4)	0.200	3.1	(0.7–6.6)	0.200
- half enrollment	5.3	(2.8–8.4)	0.200	8.8	(4.4–14.0)	0.200
GMM, retrained						
- full enrollment	4.2	(2.0–6.8)	0.200	5.2	(1.5–9.5)	0.200
HMM, retrained						
- full enrollment	4.0	(1.6–6.4)	0.200	5.2	(1.5–9.5)	0.200
Combo, baseline						
- full enrollment	6.4	(3.6–9.7)	0.200	4.0	(0.7–7.4)	0.200
Commercial system						
- full enrollment	6.8	(4.0–10.0)	0.200	6.0	(2.2–10.3)	0.200
- half enrollment	8.4	(5.2–12.1)	0.200	7.6	(3.7–12.5)	0.200

[a] $N^* = 2700$ in calculations of ρ (cf. Table 6.12)
[b] $N^* = 1200$ in calculations of ρ (30 targets with $\lfloor \bar{n} \rfloor = 40$ true-speaker tests/target)

Table 10.6: Summary of observed FRR (%) with 95% confidence intervals computed from the binomial distribution with Methods 1 and 2 to select N. In all cases, the threshold equals the *a posteriori* EER (EERd) threshold. Systems above the dashed line within each method section of the table have been retrained on PER-specific, condition-dependent background data.

Test set	T2b_Q:G8[a]			T2b_Q:LO			T2b_Q:MO			T2b_Q:MH		
ASV system	FRR	interval	ρ	FRR	interval	ρ	FRR	interval	ρ	FRR	interval	ρ
<Method 1>												
Combo, retrained												
- full enrollment	2.6	(0.0–5.5)	0.277	3.5	(0.0–7.4)	0.314	4.8	(1.0–9.4)	0.261	5.3	(2.7–8.0)	0.067
Combo, baseline												
- full enrollment	5.1	(2.0–8.2)	0.113	4.3	(1.1–8.9)	0.280	5.8	(2.3–9.1)	0.130	6.4	(3.6–9.9)	0.096
Commercial system												
- full enrollment	5.3[b]	(2.5–8.5)	0.110	6.4	(1.2–12.0)	0.306	8.7	(4.2–14.4)	0.207	8.4	(3.0–14.1)	0.252
<Method 2>	$\rho = 0.2$ ($k = 8$)			$\rho = 0.2$ ($k = 8$)			$\rho = 0.2$ ($k = 8$)			$\rho = 0.2$ ($k = 8$)		
Combo, retrained												
- full enrollment	2.6	(0.0–5.0)	0.200	3.5	(0.8–7.5)	0.200	4.8	(1.6–9.1)	0.200	5.3	(1.6–10.0)	0.200
Combo, baseline												
- full enrollment	5.1	(1.6–9.1)	0.200	4.3	(0.8–8.3)	0.200	5.8	(1.6–9.9)	0.200	6.4	(2.5–10.8)	0.200
Commercial system												
- full enrollment	5.3[c]	(1.6–9.9)	0.200	6.4	(2.5–10.8)	0.200	8.7	(4.1–14.1)	0.200	8.4	(4.1–13.3)	0.200

[a] using name and four digits, where not otherwise specified
[b] using name and five digits
[c] using name and five digits

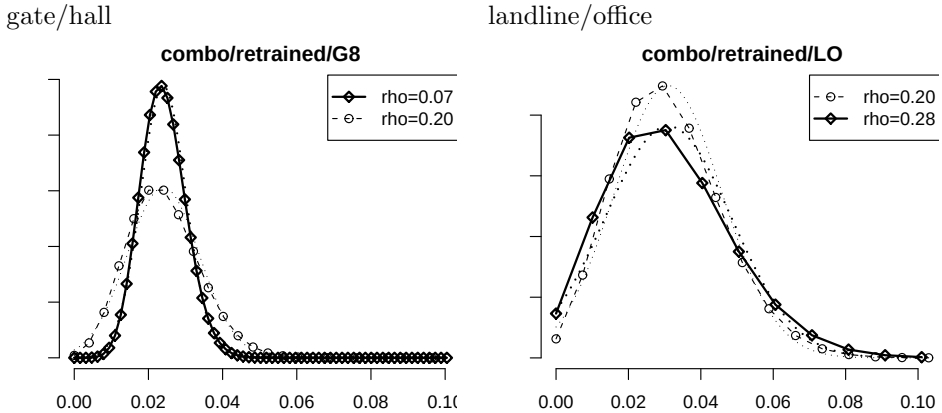


Figure 10.10: Binomial distributions used to compute confidence intervals for the retrained research system, test sets T2b_G8 and T2b_LO and full enrollment (E2a_*c*). ρ (rho) for solid lines with diamonds are computed with Method 1, while the distributions for $\rho = 0.20$ (dashed lines with circles) correspond to Method 2 with an *a posteriori* choice of ρ. The normal approximation to each binomial is shown as a dotted line.

For Method 2, we have chosen[2] a constant intra-speaker correlation coefficient $\rho = 0.2$ corresponding to values for k between 8 and 11 for the different test sets. Our prior belief about k was that a constant $k = 2$ would be a good value. Compared to values for ρ (and k) resulting from Method 1 based on an estimated variance, we chose to show intervals for a constant $\rho = 0.2$ instead, being the average over all ρ values found with Method 1 within Tables 10.5 and 10.6, respectively. Thus, the CIs shown for Method 2 in the table are based on an *a posteriori* choice of ρ.

The potential usefulness of Method 2 lies in predicting pre-trial CIs rather than estimating post-trial CIs. We include results from Method 2 here for comparison. The motivation for a constant ρ in Method 2 is that the "intra-speaker correlation" should depend mainly on the speakers and not so much on the particular test set or ASV system under test.

The evaluation strategy of Bolle et al. (2004) (applied to fingerprint data) should be applied also to ASV data to evaluate post-trial CI estimation methods. With this strategy, a corpus is randomly divided into two disjunct halves. CIs are estimated with each method on one half and compared to the "true" error rate estimated on the other half. The procedure is then repeated a number of times to estimate the coverage, i.e. the probability that a CI covers the true error rate[3].

[2] this choice is discussed in Section 11.2

[3] ideally, a 95% confidence interval should have a coverage of 95%.

Given confidence intervals from Method 1 in Tables 10.5 and 10.6, we can get an idea of which experimental differences observed in this chapter are statistically supported, and which are not, by comparing confidence intervals. Well separated, non-overlapping confidence intervals indicate strong support for the difference, while intervals that overlap to a great extent indicate no support for a difference. Note that to make formal conclusions about differences being statistically significant or not, the results normally require more rigorous analysis. In particular, our confidence intervals are derived to say something about measurements on one ASV system compared to some underlying "true" value. Comparing two systems on the same speech data, or comparing the performance of a single system on different types of data, requires other types of statistical tests, for example McNemar's test (e.g. Siegel, 1956).

Informally comparing[4] confidence intervals in the tables, we find for example:

- A positive effect from retraining the (combined) research system on PER-specific, condition-dependent tuning data is well supported in the gate/hall condition by results on the single-condition test set T2b_G8, while it is not supported in the telephone conditions. On the condition-parallel gate/hall test set (T2b_Q:G8) this difference is weakly supported.

- Performance degradation from halving the amount of enrollment data with the retrained (combined) research system is supported in the gate/hall and landline/office conditions.

- An improvement from combining the (retrained) HMM system with the GMM system (including the additional use of proper names for verification) is weakly supported in the gate/hall condition, and not support in the landline/office condition.

- Difference in performance of the retrained (combined) research system between the four conditions are not supported, except for the difference between the gate/hall and mobile/hall condition which is weakly supported.

10.5.1.1 McNemar tests

McNemar's test for the significance of changes (e.g. Siegel, 1956) is a non-parametric test that can be applied to pair-wise related measures on a nominal scale (labeled data). To apply this test in speaker verification with good theoretical justification, FRR and FAR should be treated jointly somehow (Bengio and Mariéthoz, 2004). For simplicity, however, we will take the same approach as above and compare false reject error rates only, at a global *a posteriori* EER threshold. The problem is

[4]We used the following definitions: Call two cases under comparison case A and case B, and the estimated false reject rates and confidence intervals from the two cases $\hat{p}_A$, $\hat{p}_B$, $\hat{\mathrm{CI}}_A$ and $\hat{\mathrm{CI}}_B$. A difference is *well supported* when $\hat{\mathrm{CI}}_A$ and $\hat{\mathrm{CI}}_B$ are non-overlapping; *supported* when $\hat{p}_A \notin \hat{\mathrm{CI}}_B$ and $\hat{p}_B \notin \hat{\mathrm{CI}}_A$; *weakly supported* when $\hat{p}_A \notin \hat{\mathrm{CI}}_B$ but $\hat{p}_B \in \hat{\mathrm{CI}}_A$; and *not supported* if $\hat{p}_A \in \hat{\mathrm{CI}}_B$ and $\hat{p}_B \in \hat{\mathrm{CI}}_A$.

again that FRR observations are then dependent through the threshold and also depend on impostor tests.

To compare two cases, say A and B, with the McNemar test, we compare individual FRR for each target speaker and determine if the FRR is higher or lower in case B than in case A, assuming each target has the same number of tests in both cases. Denote as $p_{\mathrm{A}i}$ and $p_{\mathrm{B}i}$ the FRR for target i in the two cases. Denote as M_{AB} the number of targets for which $p_{\mathrm{A}i} < p_{\mathrm{B}i}$ (better result in case A than in case B)[5], and as M_{BA} the number of targets for which $p_{\mathrm{B}i} < p_{\mathrm{A}i}$. Designate as the null hypothesis H_0 that there is no difference between cases A and B. Under H_0, expected values of both M_{AB} and M_{BA} would then equal $(M_{\mathrm{AB}} + M_{\mathrm{BA}})/2$. The McNemar test tests if observed values M_{AB} and M_{BA} are sufficiently different from their expected values. It proceeds by computing the test statistic

$$T_{\chi^2} = \frac{(|M_{\mathrm{AB}} - M_{\mathrm{BA}}| - 1)^2}{M_{\mathrm{AB}} + M_{\mathrm{BA}}} \tag{10.2}$$

and the probability of the value T_{χ^2}, or a more extreme value, under the χ^2-distribution with one degree of freedom ($df = 1$). If this probability is less than $(1-\alpha)/2$ (two-sided test), H_0 is rejected in favor of the alternative hypothesis, that there is a difference between cases A and B. We use the same level of significance $\alpha = 0.05$ as with confidence intervals above.

Table 10.7 shows the results of applying McNemar's test to some of the comparisons made in this chapter. Note that the results from McNemar are consistent with findings from our comparisons of confidence intervals above. All differences that were found to be at least *weakly supported* in the comparison of confidence intervals were found statistically significant with the McNemar test, while differences that McNemar tests did not find statistically significant were found *not supported* by confidence interval comparison.

Note that the McNemar test does not take into account the magnitude of differences in individual FRR between the two cases, only the sign. Since our measures are ordinal (FRR differences can be ranked with respect to their magnitude), the Wilcoxon matched-pair signed-ranks test (e.g. Siegel, 1956) could also be used, which does take the magnitude of differences into account. This is a more powerful statistical test. However, since our approach of comparing FRR at case-dependent *a posteriori* thresholds introduces dependencies between speakers, and thus the assumptions behind both tests are not quite true, we decided to use the more "blunt" McNemar test instead.

10.5.2 Length of enrollment and test data

It is clear from Figure 10.5 and confidence intervals in Table 10.5 that the (retrained) research system benefits from the rather large number of repetitions of name and

[5] M_{AB} is equivalently the number of targets for which fewer false reject errors are observed in case A than in case B, given our assumption about an equal number of tests per case for each target.

Table 10.7: Results of McNemar's test of differences at 5% level of significance.

case A	case B	common	test set	p^a	diff^b
Effect 1: *retraining on PER condition-specific background data*					
baseline	retrained	combo system,	T2b_G8	<0.001	x
		full enrollment	T2b_LO	1.0	-
			T2b_Q:G8	0.016	x
			T2b_Q:LO	1.0	-
			T2b_Q:MO	0.24	-
			T2b_Q:MH	0.45	-
Effect 2: *reducing enrollment data by a factor two*					
full	half	combo system, retrained	T2b_G8	<0.001	x
			T2b_LO	0.002	x
		commercial system	T2b_G8	0.015	x
			T2b_LO	0.34	-
Effect 3: *combining HMM subsystem with GMM subsystem*					
HMM	combo	retrained systems,	T2b_G8	0.004	x
		full enrollment	T2b_LO	0.043^c	-
Effect 4: *changing PER condition*					
G8	LO	combo system, retrained	T2b_Q:c	1.0	-
G8	MO			0.30	-
G8	MH			0.006	x
LO	MO			0.15	-
LO	MH			0.015	x
MO	MH			0.33	-

[a] probability that test statistic x has observed value T_{χ^2} or greater ($P_{\chi^2}(x >= T_{\chi^2})$)
[b] 'x' indicates a statistically significant difference detected by a two-sided test at $\alpha = 0.05$
[c] difference would have been significant with a one-sided test

digits in the full enrollment session, since cutting it to half more than doubled the EER. The same is not true for the commercial system, for which the EER increased by only about 25%. This difference between the two systems can be partly explained by target model size: gross model size is about five times larger for the research system than for the commercial system after compressing each model set using Lempel-Ziv coding to partly compensate for an ineffective storage format used with the research system. The research system was dimensioned to operate with rather large amounts of enrollment data.

Test utterances in this study are an order of magnitude shorter than the total length of enrollment and consist of a name and a string of digits. It was argued in Section 5.3 that it is difficult to collect digit strings with more than four digits from users in a telephone application through aural prompts, but collecting longer digit strings through visual prompts as in the gate case should be feasible. In this study we collected only five digits per utterance in the gate/hall condition, and simulated the use of two, three and four digits per utterance with the results in Section 10.4.5 (p. 11) and Figure 10.6. Figure 10.11 shows a prediction of EER for test utterances with longer digit strings, based on an exponential fit to the EERs of Figure 10.6 extended with the EER for the corresponding experiments with a name only, and a name plus a single digit. It suggests that the EER with a name plus six digits would be 1.8%, a 37% relative reduction compared to 2.8% for a name plus four digits. The prediction of 1.1% EER for a name plus eight digits is uncertain because, firstly, it is not evident that the exponential prediction model is valid for longer digit strings; and, secondly, it is also not evident that users would accept such long strings, and it is likely that they would generate significantly more disfluencies, such as substitutions, hesitations, repairs, etc. Such disfluencies are likely to generate errors in the speech recognition process and the resulting segmentation used by the speaker verification system.

A complementary approach for collecting more test data efficiently is to rely on a sequential decision strategy such as a heuristic method (Furui, 1981; Naik and Doddington, 1986) or one based on Wald's sequential probability ratio test principle (Lund and Lee, 1996; Surendran, 2001).

10.5.3 Effects of fusion

Fusion results in Section 10.4.3 (p. 10) show a large error rate reduction from each of the individual systems to their combination. This may be surprising since both subsystems are based on similar features, classifiers and normalization techniques, and their output score should therefore be correlated and not be very good candidates for score fusion. However, one major difference is their use of data: the GMM subsystem uses both names and digits, while the HMM subsystem ignores the name and uses the digits only. To understand/explain the underlying factors, we tested the separate and combined systems on the individual and combined parts of the gate/hall test utterances. In all cases were clients enrolled using the full name plus

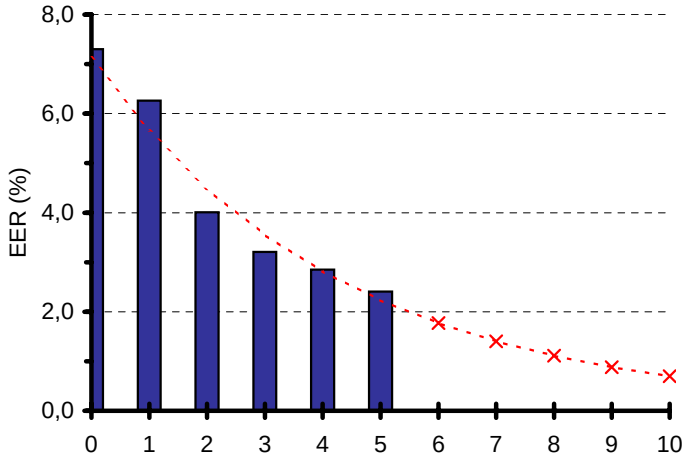


Figure 10.11: The EER values for the retrained system with the gate/hall single-condition test set S2b_G8 and including the name plus zero through five digits in each test (bars) and a prediction (dashed line) of EER values for a different number of digits using an exponential model (cf. Figure 10.6).

digits enrollment set (E2a_*c*). EERs are shown in Table 10.8, where cases C, D and F match the DET curves included in Figure 10.4a.

Using the cases in the table, the formation of the final result (case F) can be illustrated with the two alternative paths of information fusion shown in Figure 10.12. Each path consists of conceptual information fusion along two axes: score fusion of separate systems and vocabulary fusion of the name and the digit parts of the test utterance. The lower path (via case E) consists of one fusion step along each axis: a system fusion on the digits part of the test utterance, followed by fusion of the two parts of the utterance. Hence, each step combines independent sources of information. The upper path (via case C) more closely reflects the actual structure of the ASV system, but it contains the fusion (C,D)→F with simultaneous fusion along both axes, since it combines system scores based on different parts of the test utterance, and hence combines information sources that are not independent. We propose that the lower path better explains the formation of the system output (from a conceptual point of view).

10.5.4 On-site vs. telephone use of ASV

Is there a greater potential for well-performing ASV in an on-site application than in a telephone application? This is a very general question, and of course we don't have a foundation to answer it in a general sense, but we do have some clues for our particular application instances.

Table 10.8: EER for the individual subsystems and their combination applied to digits-only or name-only subsets of the S2b_G8 test set, or the complete test set using both name and digits. The enrollment set is the full (name and digits) E2a_G8. The combined system in case E uses the same score combination weights as the system in case F.

Case	System	Vocabulary	EER
A	gmm	name	7.3%
B	gmm	digits	6.9%
C	gmm	name, digits	4.2%
D	hmm	digits	4.0%
E	combo	digits	3.4%
F	combo	name, digits	2.4%

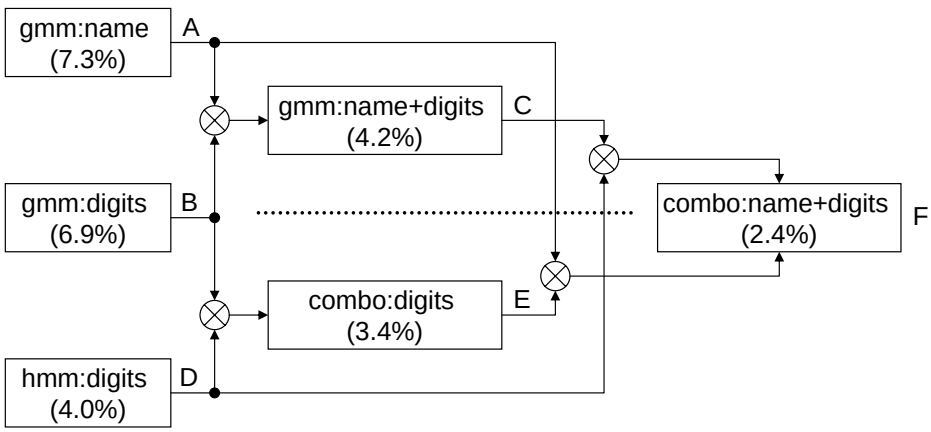


Figure 10.12: Two alternative conceptual information fusion paths that explain how the output of the ASV system is formed. Numbers within parentheses are the measured EERs for the gate/hall condition with the system and test utterance content represented by each box.

We believe our comparison between ASV in the on-site and telephone version of PER is fair. First, the design differences introduced between the on-site and telephone versions of PER are well founded. For example, verification based on a client's proper name and a digit sequence collected in a single utterance, where the digits are visually prompted, works well in the on-site application, while the name and digits must be separated in the telephone case. Aural prompts are the only alternative with most telephones (since they don't have a display). Second, our data collection procedure and design of the condition-parallel test sets based on series of sessions recorded in chronological proximity in the four conditions, allow for similar prerequisites in the four conditions. If a subject suffered a head cold during a gate session, the same was true in telephone sessions within the same series. If there was a noisy background, it was probably there both during a gate session and a corresponding mobile telephone session in the hall. Random between-session variation in for example background noise will naturally have occurred, but such variation can only be excluded by stereo recordings, and stereo recordings in our case would have meant a more artificial context for the recordings. Systematical differences between conditions may also have occurred, however. For example, in many series the gate/hall session was made before telephone sessions because it was recorded when the subject arrived to work in the morning. Since a number of steps had to be climbed to reach the gate, (true-speaker) subjects might have been more out-of-breath at the gate than after arriving in the office and sitting down to make the landline/office call. There is also the possible difference in motivation in subjects, since the gate version of PER could actually open the door, while telephone calls were made for recording purposes only.

Results from the condition-parallel test sets indicate that, provided acoustic models in the speech recognition and speaker verification components are tuned using proper development data, ASV error rate may be lower in the on-site condition than in all three telephone conditions, though a statistically significant advantage was measured only relative to the mobile/hall condition (Table 10.7). With acoustic models trained on a general-purpose telephone corpus, little or no difference was seen between the conditions. The performance difference introduced by tuning on proper development data was large for the on-site application and non-significant for the telephone application. This highlights an important difference between using a variety of ubiquitous telephone handsets vs. using a single microphone in a particular room in an on-site application: the need for dedicated tuning data is larger for the on-site application than for telephone applications. It should be easier to create an ASV system that will perform consistently at a near optimum error rate between instances of telephone applications without tuning it for every particular application, while tuning data will be important for any on-site application. To achieve the best possible performance, however, tuning data from the application will usually be needed in either case.

Our point estimates of EER in the gate/hall and landline/office conditions were 2.6% vs. 3.5% on the condition-parallel test sets using the same number of digits in the test utterance. While this corresponds to a 25% relative reduction for the on-

site application, the statistical uncertainty in the estimates is large and we can not infer a difference between the two conditions based on these measurements alone. But that was with the same number of digits. In our on-site version of PER we used five digits that caused no apparent trouble for subjects, and we believe six digits would have worked well too. Considering Figure 10.11, we would expect a 37% relative reduction in EER for six digits compared to four, suggesting the 2.6% EER for the on-site application could be reduced to 1.6%[6]. We are then up to a 54% reduction for gate/hall relative to the landline/office condition. The corresponding reductions are 67% relative to mobile/office and 70% relative to mobile/hall.

We can further speculate into factors we have not tested. In our experiments with on-site data, we used a downsampled 8 kHz version of the original wide-band audio recordings made at 16 kHz. Given good development data for wide-band speech and a proper modification of the system's speech feature representation to operate with 16 kHz data, it should be possible to achieve a further reduction in error rate in the on-site system. Furthermore, we saw in Section 6.3.4.5 that the proportion of different-number calls (test calls from a different telephone number than the target's enrollment call) was higher in the impostor part of the telephone condition test sets (around 25%) than in the true-speaker part, suggesting that in a fully same-channel test set error rate in telephone conditions could have been higher than what we saw in our data.

To conclude the discussion on the potential for well-performing ASV in an on-site application vs. in a telephone application, our data suggests that, given the availability of application-specific tuning data, ASV error rate may be less than half in an on-site application than in a corresponding telephone application.

10.6 References

Bengio, S. and Mariéthoz, J. (2004). A statistical significance test for person authentication. In *Proc. Odyssey04 - The Speaker and Language Recognition Workshop*, pages 237–244, Toledo, Spain. [20]

Bimbot, F., Blomberg, M., Boves, L., Chollet, G., Jaboulet, C., Jacob, B., Kharroubi, J., Koolwaaij, J., Lindberg, J., Mariethoz, J., Mokbel, C., and Mokbel, H. (1999). An overview of the PICASSO project research activities in speaker verification for telephone applications. In *Proc. 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1963–1966, Budapest, Hungary. [1]

Bimbot, F., Blomberg, M., Boves, L., Genoud, D., Hutter, H.-P., Jaboulet, C., Koolwaaij, J., Lindberg, J., and Pierrot, J.-B. (2000). An overview of the CAVE

[6]Note that Figure 10.11 and the previous discussion on test utterance lengths are based on the single-condition test set S2b_G8, while in this paragraph we look at the condition-parallel test set S2b_Q:G8.

project research activities in speaker verification. *Speech Communication*, 31(2-3):155–180. [1]

Bolle, R., Ratha, N., and Pankanti, S. (2004). An evaluation of error confidence interval estimation methods. In *Proc. 17th International Conference on Pattern Recognition (ICPR)*, volume 3, pages 103–106, Cambridge, UK. [19]

Bonastre, J.-F., Bimbot, F., Boë, L.-J., Campbell, J., Reynolds, D., and Magrin-Chagnolleau, I. (2003). Person authentication by voice: A need for caution. In *Proc. 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 33–36, Geneva, Switzerland. [2]

Doddington, G. (1985). Speaker recognition - identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664. [1, 2]

Elenius, K. (2000). Experiences from collecting two Swedish telephone speech databases. *International Journal of Speech Technology*, 3:119–127. [1, 3]

Feix, W. and DeGeorge, M. (1985). A speaker verification system for access-control. In *Proc. 1985 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 399–402, Tampa, Florida. [1]

Fredouille, C., Mariéthoz, J., Jaboulet, C., Hennebert, J., Bonastre, J.-F., Mokbel, C., and Bimbot, F. (2000). Behaviour of a bayesian adaptation method for incremental enrollment in speaker verification. In *Proc. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1197–1200, Istanbul, Turkey. [14]

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2):254–272. [23]

Hennebert, J., Melin, H., Petrovska, D., and Genoud, D. (2000). POLYCOST: A telephone-speech database for speaker recognition. *Speech Communication*, 31(2-3):265–270. [1, 3]

Lamel, L. (2005). Personal communication. [2]

Lindberg, B., Johansen, F., Warakagoda, N., Lehtinen, G., Kacic, Z., Zgank, A., Elenius, K., and Salvi, G. (2000). A noise robust multilingual reference recogniser based on SpeechDat(II). In *Proc. 2000 International Conference on Spoken Language Processing (ICSLP)*, pages 370–373, Beijing, China. [4]

Lindberg, J. and Melin, H. (1997). Text-prompted versus sound-prompted passwords in speaker verification systems. In *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 851–854, Rhodes, Greece. [1]

Lund, M. and Lee, C. (1996). A robust sequential test for text-independent speaker verification. *The Journal of The Acoustical Society of America*, 99(1):609–621. [23]

Mansfield, A. and Wayman, J. (2002). Best practices in testing and reporting performance of biometric devices, version 2.01. Technical report, UK Biometric Working Group (BWG). [16]

Mariani, J. (1992). Spoken language processing in the framework of human-machine communication at LIMSI. In *Proc. 5th DARPA Speech and Natural Language Workshop*, pages 55–60, Harriman NY, USA. [1, 2]

Melin, H. (1996). Gandalf - a Swedish telephone speaker verification database. In *Proc. 1996 International Conference on Spoken Language Processing (ICSLP)*, pages 1954–1957, Philadelphia PA, USA. [1, 3]

Melin, H. (1998). On word boundary detection in digit-based speaker verification. In *Proc. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C)*, pages 46–49, Avignon, France. [1]

Melin, H., Koolwaaij, J., Lindberg, J., and Bimbot, F. (1998). A comparative evaluation of variance flooring techniques in HMM-based speaker verification. In *Proc. 1998 International Conference on Spoken Language Processing (ICSLP)*, pages 1903–1906, Sydney, Australia. [1]

Melin, H. and Lindberg, J. (1999). Variance flooring, scaling and tying for text-dependent speaker verification. In *Proc. 6th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 1975–1978, Budapest, Hungary. [1, 3]

Morin, P. and Junqua, J.-C. (2003). A voice-centric multimodal user authentication system for fast and convenient physical access control. In *Proc. Workshop on Multimodal User Authentication*, pages 19–24, Santa Barbara CA, USA. [1, 2]

Naik, J. and Doddington, G. (1986). High performance speaker verification using principal spectral components. In *Proc. 1986 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 881–884, Tokyo, Japan. [23]

Neiberg, D. (2001). Text independent speaker verification using adapted gaussian mixture models. Master's thesis, KTH/TMH, Stockholm, Sweden. [1]

Nordström, T., Melin, H., and Lindberg, J. (1998). A comparative study of speaker verification systems using the Polycost database. In *Proc. 1998 International Conference on Spoken Language Processing (ICSLP)*, pages 1359–1362, Sydney, Australia. [1, 3]

Rosenberg, A. (1976). Automatic speaker verification: A review. *Proceedings of the IEEE*, 64(4):475–487. [2]

Schalk, H., Reininger, H., and Euler, S. (2001). A system for text dependent speaker verification - field trial evaluation and simulation results. In *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 783–786, Aalborg, Denmark. [1]

Setlur, A. and Jacobs, T. (1995). Results of a speaker verification service trial using HMM models. In *Proc. 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 639–642, Madrid, Spain. [1, 2]

Siegel, S., editor (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, New York. [20, 21]

Surendran, A. (2001). Sequential decisions for faster and more flexible verification. In *Proc. 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 763–766, Aalborg, Denmark. [23]

Wagner, T. and Dieckmann, U. (1995). Sensor-fusion for robust identification of persons: A field test. In *Proc. IEEE International Conference on Image Processing*, volume 3, pages 516–519, Washington D.C., USA. [1]

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1999). *The HTK Book (for HTK version 2.2)*. Cambridge University, Cambridge, UK. [8]