KTH Computer Science
and Communication

6. Corpora

HÅKAN MELIN

KTH/CSC/TMH
Stockholm, Sweden 2007

This is a self-contained re-print of Section 6.3 from Chapter 6 in:
Melin, H. (2006). Automatic speaker verification on site and by telephone: methods, applications and assessment. Doctoral Thesis, KTH, Stockholm, Sweden 2006. ISBN 978-91-7178-531-2.

Contents

Chapter 6

Corpora

6.3 The PER corpus

6.3.1 Introduction

The PER corpus is the result of data collection during 2003–2004 from actual use of a speaker verification system, namely that in the PER system described in Chapter 5. It consists of recordings of proper names and digit sequences spoken in Swedish and is suitable for experiments in text-dependent speaker verification.

The main design criteria for the corpus were to support an evaluation of the performance of ASV in the PER application, and to allow a fair comparison between speaker verification in on-site vs. telephone use.

The first criterion followed from our goal to evaluate the PER system during live use. New speech data were needed to this end because we had no suitable data before. The Gandalf corpus contained only telephone data, while PER used a wide-band microphone mounted in a reverberant room. It contained visually and aurally prompted digits, but no proper names, and the available text-dependent non-digit phrases were sentences shared by all subjects. Finally, Gandalf was recorded using a tape-recorder metaphor, where subjects were speaking to a machine without any feedback on how they were speaking, and we felt the difference to talking to a live ASV system could be important.

The second criteria came from a wish to relate results from on-site use of ASV to our previous research on its telephone use. Furthermore, we wanted to experiment with cross-condition enrollment and testing, where a client is supposed to enroll *once*, say by telephone, and then be ready to use his voice for authentication *anywhere*, be it a telephone or on-site application.

To allow a fair comparison between speaker verification in on-site vs. telephone use, the data collection was designed to include telephone data in parallel to on-site data. The telephone version of PER was thus created and a part of the test group was asked to make telephone calls in conjunction with their entry through the gate protected by the on-site version. Differences introduced in the telephone version

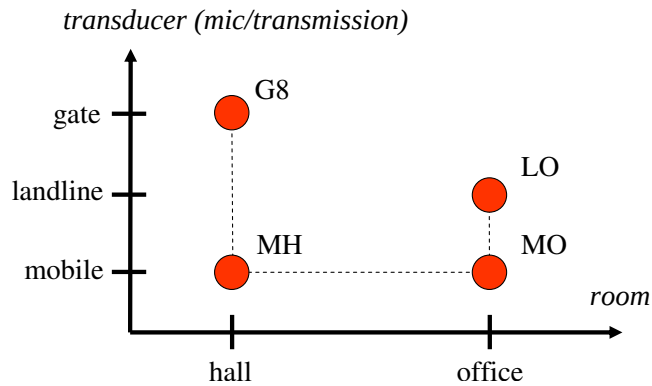

Figure 6.1: The four recording conditions in the PER corpus: gate/hall (G8), mobile/hall (MH), mobile/office (MO) and landline/office (LO).

with respect to the on-site version are motivated by the dissimilar prerequisites of the two cases as outlined in Section 5.3.

To allow for even better comparison between on-site data and the rather broad class of "telephone" data, and after noting that calling from a telephone rather than talking to the system on site potentially involves a change of room in addition to the recording transducer and channel, we decided to collect data in four separate *conditions*: through the microphone at the gate in the hall (stairwell), through a mobile telephone in the same hall, through the same mobile telephone from an office, and through a landline telephone from the same office. In this way, there is a change in one dimension at the time in a *room–transducer space* between the four conditions, and it should be possible to determine whether a difference in ASV error rate between any two conditions is mainly due to a change in the room (with associated background noise) or to the transducer (and associated transmission effects). The four conditions in the room–transducer space are labeled gate/hall, mobile/hall, mobile/office and landline/office in this thesis, and sometimes abbreviated G8[1], MH, MO and LO. They are illustrated in Figure 6.1.

While the PER system versions used for collecting data were not optimized for this particular application and the respective condition, a set of separate speakers were collected to serve as development data (background data) for creating optimized systems for later simulation experiments. The so called background speakers were recorded in each of the four conditions.

[1]G8 for the downsampled 8 kHz version of gate data

6.3.2 Data collection

This section describes the recording procedure and provides statistics on subjects and sessions in the corpus.

Given the two main design criteria for the corpus, two somewhat conflicting goals were pursued in data collection: to collect as much data in the primary gate/hall condition as possible, and to collect as many parallel data as possible from all four conditions. To resolve this conflict, client subjects were divided into two groups, where one of the groups provided data in the primary condition only, while the other group provided data in all four conditions.

Subjects were interacting with the on-site or telephone versions of the fully automated PER system described in Chapter 5, that uses automatic speech recognition and speaker verification to recognize the content of spoken utterances and to verify users' claimed identities. If an utterance (or pair of utterances in the telephone case) was not found to contain a valid claim the system prompted the user to try again. Each system session allowed up to three attempts, but there was no limit on how many consecutive sessions a user was allowed to initiate. If a valid claim was found, the on-site version of the system physically unlocked the gate, while the telephone version did not. Both versions welcomed the user verbally. If no valid claim was found after three attempts, the system informed the user of this verbally.

Audio data from the telephone version of PER were recorded through an ISDN-line and stored as one utterance per file in the format used in the Euro-ISDN network. Thus, the sampling rate is 8 kHz, samples are A-law coded and stored with 8 bits per sample, the same format used in the Gandalf corpus. Audio data from the on-site version were recorded at 16 kHz sampling rate with 16 bits per sample (linear amplitude scale) and stored as one utterance per file. The same data was also decimated to 8 kHz sampling rate as described in Section 5.2 and stored in the same format as telephone data.

6.3.2.1 Subjects

Subjects are divided into two disjoint groups: the *test group* and the *background speakers group*. Those in the test group have been assigned one or both of the functions *client* and *impostor*. As clients they are further divided into *group L* (limited) and *group E* (extended) with respect to how much effort they were willing to spend as subjects. The main difference between the tasks of client subjects in the two groups is that group E provides data in all four conditions and group L only in the gate/hall condition.

Out of 56 subjects who volunteered to the client group and attempted to enroll to the system, 54 (16 female and 38 male) succeeded to enroll[2]. They were all students or staff from the Department, with the age distribution shown in Figure 6.2 together with the age distribution of the Swedish population. Like in the Gandalf corpus, the age distribution in client subjects in the PER corpus has two pronounced peaks

[2]for results related to the enrollment process, refer to Section 10.3.1

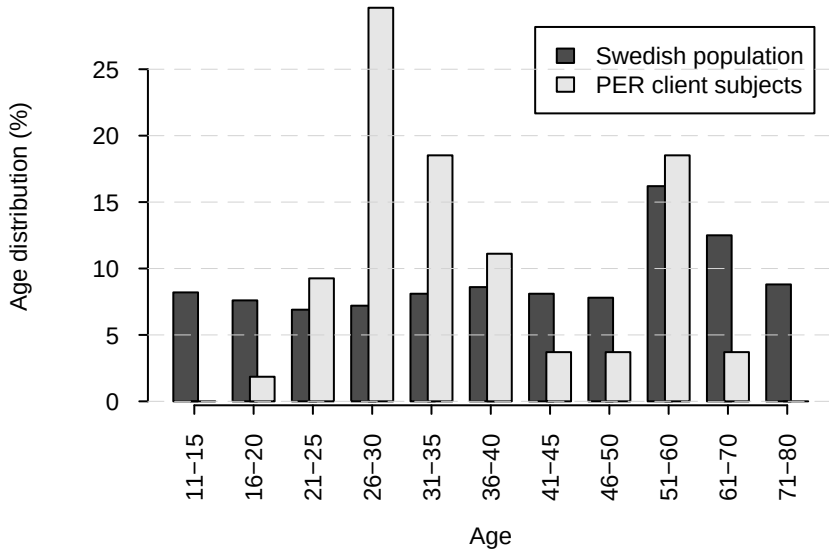

Figure 6.2: Age distribution, at the start of the recording period, among the 54 client subjects together with the distribution for the Swedish population between ages 11 to 80 (Statistics Sweden (SCB), 2004). Note that the three right-most age intervals span 10 years each while the others span 5 years per interval (for compatibility with the corresponding figure for the Gandalf database, Figure **??**, p. **??**).

at around 30 and 50 years of age. Among both male and female client subjects (who succeeded to enroll), half had previously been assigned to group L and half to group E.

Background speakers were recruited mainly from students and staff outside of the Department. 51 male and 28 female background speakers were recorded. While subjects in the test group used their own names, background speakers were assigned alias names. Alias names were chosen through the following procedure with the goal of including the most common Swedish names based on name statistics from Statistics Sweden (SCB) as of December 2002.

Starting from a list of the 100 most common family names in Sweden, 21 redundant names were removed. They were either homophones (e.g. Carlsson-Karlsson), phonetically similar (e.g. Jonsson-Jansson, Peterson-Petterson, Jonasson-Johansson), or substrings of other names (e.g. Ström-Strömberg). The order of the remaining 79 names was then randomized. The first 50 were then combined with a male first name and the remaining 29 with a female first name as described below. The 79 family names cover 31% of the Swedish population.

First names were processed similarly to family names, except they were used in

frequency order. Starting from the 100 most common male (female) first names, 10 (13) were removed because they were homophones of other names in the list, or phonetically similar. The frequency count of a removed name was added to the frequency count of the similar name kept in the list. Names were then re-ordered according to adjusted frequency counts and assigned to subjects in that order. The 51 male names used by a background speaker and seven corresponding similar names cover 51% of the Swedish male population, while the 28 female names plus two similar names cover 30% of the female population. More male first names than female first names were used since male subjects spoke only male first names, and vice versa, and more male background speakers than female ones were collected.

6.3.2.2 Recording conditions

As introduced above and illustrated in Figure 6.1, speech data were collected in four different conditions, referred to as gate/hall, landline/office, mobile/office and mobile/hall.

The primary condition was gate/hall. It was also the most naturally occurring condition of the four in that users had to pass through the gate to enter into the Department, and the PER system provided one of the three possible ways for employees to unlock the gate. The three telephone conditions were more artificial because the telephone version of the system did not give access to anything.

To allow comparison between the four conditions, parallel data were collected in *series* of sessions. A series consists of one session per condition recorded within a short time period with the same claimant speaker and the same claimed identity (target speaker) in all sessions. Subjects were asked to record sessions in a series as close as possible in time, preferably in immediate succession, and at the least to record them within the same day. They were also asked to vary the order of conditions between series.

In the gate/hall condition, every session by all claimants against any target is recorded through the same channel, i.e. with a single microphone, fixed amplifier gain, fixed recording level, etc. To establish a corresponding same-channel situation for telephone conditions (a single channel per target, but different channels for different targets), each client was asked always to use the same landline phone and the same mobile phone, and impostors were instructed to use the exact same telephone instruments as their target (to borrow phones from their target). These instructions were also followed in practice, with the exception that some subjects in the test group obtained a new mobile phone during the collection period and did not keep the old one. In these cases impostor attempts were made with the new telephone resulting in different channels between true-speaker and impostor tests, since most impostor sessions were recorded after the corresponding true-speaker sessions.

All telephone calls were made to a toll-free number to allow subjects to use their own mobile phone without being billed for their calls.

To balance out a potential bias from learning effects during enrollment in comparison between the gate/hall and landline/office conditions, half of the subjects within the test group and the background speakers group made their first enrollment session in the gate/hall condition and the second in landline/office, while the other half started with enrollment in landline/office. Enrollment in mobile conditions was always made after the other two enrollment sessions, however, allowing for a bias between mobile conditions on the one hand and gate/hall and landline/office on the other.

6.3.2.3 Client sessions

Clients in group L provided enrollment and test sessions in the gate/hall condition, plus an enrollment session in the landline/office condition. They were asked to provide at least 20 test sessions in the gate/hall condition during at least 15 different days. Clients in group E provided enrollment and test sessions in all four conditions. They were asked to provide 30 series of one session per each of the four conditions (for the definition of such series, see above) during at least 15 different days and then continue with at least 20 gate/hall sessions during different days.

During an enrollment session, the PER system collected one valid repetition of between eight and ten items per client and condition as described in Section 5.5. Each item consisted of the client's name and a string of five digits.

6.3.2.4 Impostor sessions

Since clients use their own name for verification, dedicated impostor sessions had to be collected. Impostor sessions against targets in client group L were collected in the gate/hall condition only, while impostor sessions against targets in client group E were collected as series of sessions in all four conditions. Impostor subjects were mostly the same people that also participated as clients. They knew most of their targets and were allowed to imitate the target's voice, though from listening through the recordings during annotation work, it turned out not many imitations were made in practice. Only same-sex impostor attempts were collected.

6.3.2.5 Background sessions

Background speakers made one complete enrollment session in each of the four conditions using office telephones and mobile phones mostly not used by subjects in the test group.

In each session they provided similar data as subjects in the test group (except they spoke their assigned alias name instead of their own name), plus five sentence items. The first sentence item was the same for all background subjects, "öppna dörren innan jag fryser ihjäl" (open the door before I freeze to death), while the remaining four were selected from a pool of 114 sentences such that one or two subjects of the same gender spoke the same sentence. Each subject spoke the same

sentences in all four conditions. Sentences were 5 to 14 words long (average 7.4 words) and between 21 and 49 phonemes long using a prototypical transcription (average 33 phonemes).

6.3.3 Annotation

Recorded data were manually annotated on session and file level, where a file is intended to contain a single utterance. Annotations were made with a graphical tool dedicated to this task. Wherever possible, the tool provided initial values for annotation fields that the annotator could confirm or change to the appropriate value. Initial values were taken from output saved by the PER system during data collection into a relational database and XML session log files (cf. Appendix G for a specification of log file contents).

6.3.3.1 Session level annotation

Sessions were annotated with *claimant identity*, *claimed identity* and *session status*. Figure 6.3 shows an example screen shot of the graphical tool used for this purpose.

The claimant identity was determined by comparing audio and video data recorded during the session to reference audio and video data for known identities. The annotation tool provided an *identity browser* where the annotator could traverse a list of known identities and inspect reference data for each of them, and a session browser where the annotator could listen to recorded files and view recorded images from the selected session. The annotator could create new identity entries as new subjects were encountered.

The claimed identity was determined by listening to one or more audio files for the spoken name. An instance of the identity browser was used for this field too, mainly to provide the annotator visual feedback on the currently selected identity.

The default value selected for both identity fields when the annotator loaded a new session was the identity corresponding to the name recognized by the PER system, *a priori* assuming the session was a true-speaker session. Since the annotator tool showed three images simultaneously (one from the selected session, one for the currently selected claimant identity and one for the currently selected claimed identity) the annotator could very quickly verify current selections by comparing the three images and listening to one or more audio files.

Session status is a categorization with the main categories “valid” and “invalid”. Valid sessions were further sub-categorized as “complete” or “incomplete”. A session is considered valid and complete if it contains at least one file (pair of files for telephone sessions) with a name and the requested number of digits (five for the gate/hall sessions and four for telephone sessions). Remaining sessions were classified as valid but incomplete if a user was trying to make a (serious) attempt but failed to record at least one complete attempt (e.g. user spoke very slowly and the last words were truncated in the recording, or in a telephone session, the spoken name was never recognized as the name of an enrolled client); or invalid if a person

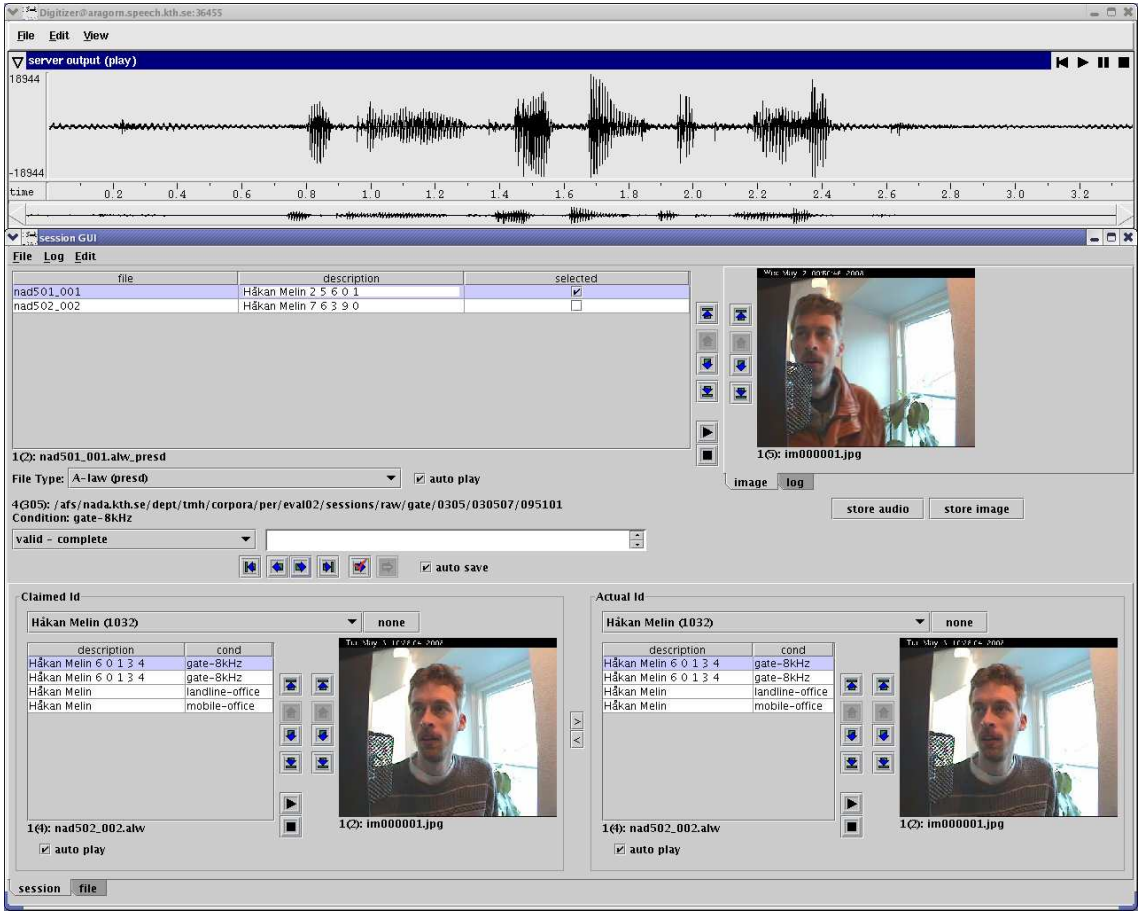

Figure 6.3: The session annotation tool showing a true-speaker session. The "Session GUI" window contains the session browser (upper half), the identity browser for selecting the claimed identity (lower left), and the identity browser for selecting the actual identity of the claimant (lower right). The upper window holds a WaveSurfer widget for listening to the audio file select in the session browser, or looking at some graphical representation of it.

was not judged by the annotator to make a serious attempt to use the system, if an unregistered identity was claimed, or if there was no recorded speech (a session was triggered by mistake).

6.3.3.2 File level annotation

Files were annotated with a *graphical transcription* and an optional *free-text comment*. If the speaker in a file was different from that selected for the session-level claimant identity (i.e. the speaker changed during the session), that file was also annotated with a *file-level claimant identity*.

Graphical transcriptions were based on SpeechDat conventions for transcription (**?**). Standard conventions used were markers for stationary noise (`[sta]`), intermittent noise (`[int]`), speaker noise (`[spk]`), mobile phone-specific noise (`%word`), truncated signals (`~word` or `word~`), and mispronounced or truncated words (`*word`). To these were added a weaker marker for pronunciation errors used specifically with names (`&name`), and variants of intermittent noise for the particular noise occurring when the bar gate was opened (`[igo]`) or closed (`[igc]`). The `&`-marker was used with impostor attempts where the impostor pronounced a target's name in a different way than the target himself, and the difference was distinct enough to be captured by a phonemic transcription with word accent markers, but not so much different as to merit a `*`-marker for an incorrect pronunciation. Words labeled with `*` or `&` in the graphical transcription were transcribed phonemically in the comment field. The comment field was also used to note cases of clearly altered voices in the speaker, such as a whispering or high-pitched voice.

6.3.4 Data sets

This section describes the enrollment and test sets used in this thesis (some additional data sets not used in the thesis are defined in Appendix C). A data set is defined by rules to select claimants, target speakers, sessions and files.

6.3.4.1 Notation

Data sets are denoted tix_c, where parameter t is E for enrollment sets, T for true-speaker test sets, I for impostor test sets and S for complete test sets (combined true-speaker and impostor test sets). Parameter c indicates the recording condition and takes values {G8,LO,MO,MH}. Parameter x indicates how files are selected from a given session. It is referred to as "accepted text status" in the set definitions below, and takes values {a,b} meaning

a: "accepted text status" means that both the target's name and the prompted digits were included in one of the hypotheses produced by the speech recognizer during collection;

b: "accepted text status" means that both the target's name and the displayed digits were spoken, as indicated by a manually verified transcription. To be more

specific, the following conditions must be met by the transcription of a file (pair of files in telephone conditions): the complete name is included, but no modifier-labeled (~, *, %) repetitions of any part of the name; and the prompted digits are included in the given order, without modifiers, and with no other words in between. Note that names with the &-modifier are allowed, but not names with the *-modifier. Noise markers are allowed anywhere in the transcription.

Parameter i is simply an index number.

The notation introduced here is more general than required to cover the data sets actually used in this thesis. Specifically, we have always used text acceptance rule a in enrollment sets, while b was used in all test sets, and we have only used one single-condition test set and one condition-parallel test set per condition, all with index number 2. We have chosen to keep this notation for consistency with unpublished results and potential future experiments using other data sets. Other data sets, not used in this thesis, are defined in Appendix C.

6.3.4.2 Client enrollment sets

Based on data collected during enrollment sessions, two enrollment sets per condition c were defined using text acceptance rule a:

- E1a_c: (half session) the first five items from the last recorded and complete enrollment session from each client speaker under condition c; the first repetition of each item with accepted text status (approximately 15 seconds of speech per speaker).
- E2a_c: (full session) all ten items from the last recorded and complete enrollment session from each client speaker under condition c; the first repetition of each item with accepted text status (approximately 30 seconds of speech per speaker).

Sets E2a_c use the exact same data as was used during on-line enrollment into the collection system, while sets E1a_c can be used to simulate enrollment with only half of the speech data actually collected.

The G8 and LO client enrollment sets include 38 male and 16 female clients, while the MO and MH sets include 19 male and 10 female clients.

Corresponding enrollment sets have been defined on background speaker data as presented in Appendix C. Background enrollment sets also include sets with pooled speakers for training multi-speaker background models.

6.3.4.3 Single-condition test sets

Separate true-speaker and impostor test sets T2b_c and I2b_c are first defined for each recording condition c. Those are then combined condition-wise into the complete test sets S2b_c. In this thesis only a single complete single-condition test set per condition is used, and only for the gate/hall and landline/office conditions.

Table 6.1: Test set sizes for the PER corpus. Number of subjects are specified as number of male / number of female subjects. All impostor attempts are same-sex attempts.

	Test set	S2b_G8	S2b_LO	S2b_Q:c
Test group	Targets	38 / 16	24[a] / 9[b]	19 / 8
	Impostors	76 / 22	37 / 16	35 / 16
	True-speaker tests	4643	1228	977
	Impostor tests	1121	422	393
Background speakers group	Speakers	51 / 28		

[a]Three of the 24 targets (`M1014`, `M1023`, `M1101`) have enrollment data but only a single true-speaker test each, and no impostor attempts. Two additional targets (`M1122`, `M1127`) have 7 and 20 true-speaker tests each, but no impostor tests.

[b]Target `F1124` has 20 true-speaker tests but no impostor tests.

Common to both true-speaker and impostor test sets is that they contain no more than one attempt from any given session, and only from login sessions annotated as valid and complete that contain at least one attempt whose file level transcription meet the conditions of the b-criterion for "accepted text status". The true-speaker test sets include one attempt per session from all such true-speaker login sessions (no limit on the number of sessions per day or per target speaker). The impostor test sets include one attempt per combination of impostor speaker and target where the impostor speaker has recorded at least one session where (s)he claimed the given target identity. If there is more than one such session, the first one is used. Only same-sex impostor tests are used.

In sessions where more than one attempt satisfies the b-criterion for "accepted text status", the first attempt is used. Note that this selection depends on the manual transcription of recorded files only, and is independent of the results of speech recognition and speaker verification in the automatic PER system that collected the data. Table 6.2 below shows examples of files that were *not* included in the S2b_G8 test set because they did not satisfy the file-level selection criteria defined by the b-selection rule.

Table 6.1 shows the number of speakers and tests included in the PER test sets, including condition-parallel test sets defined below, while Figure 6.4 shows how tests are distributed over targets in the G8 true-speaker and impostor test sets.

6.3.4.4 Condition-parallel test sets

To feature a comparison between conditions, a quadruple of condition-parallel test sets have been defined, each denoted S2b_Q:c, where Q is a short-hand notation for a list of the included conditions, Q={G8,LO,MO,MH}, and c is one of the four conditions. A condition-parallel test set is constructed such that there is always exactly one test from each of the listed conditions that correspond to each other

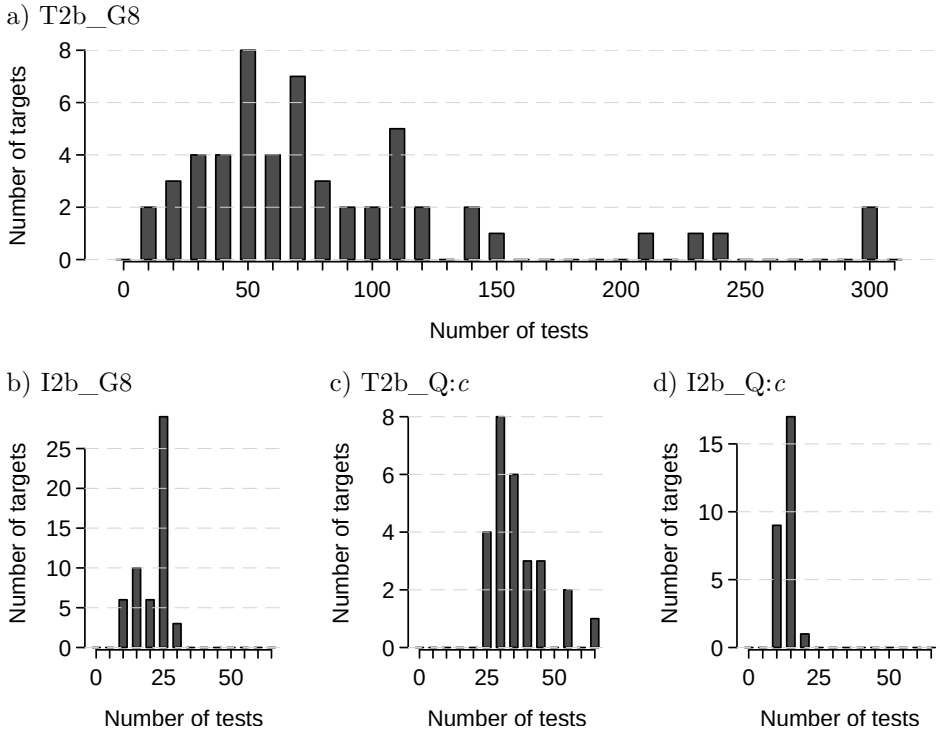

Figure 6.4: Histograms showing how many targets have how many tests in the gate/hall true-speaker and impostor test sets T2b_G8 and I2b_G8, and in the condition-parallel test sets T2b_Q:c and I2b_Q:c (per condition).

in the sense that they were recorded near each other in time. Such a group of one test per condition is called "a series", like it was during the data collection. The file selection criterion specified in the single-condition test set is applied to each condition individually. If there is no selectable file for one or more conditions, no corresponding series is constructed. Sessions in the various conditions to be grouped into a series should have been recorded as close as possible to each other in time. They must at least have been recorded during the same day.

6.3.4.5 Test set statistics

Handset use Impostor subjects in the test group were asked to make calls in telephone conditions from the same telephone instruments used by the target speakers during their enrollment, and this request was well responded to. A comparison between A-numbers in test calls included in the true-speaker-part of the condition parallel test sets (S2b_Q:c) and the corresponding enrollment calls shows that 6.0%

of calls in the landline/office condition and only 0.1% of calls from each of the mobile conditions were made from a different number. All but one of the different-number calls in landline/office were made by two subjects who changed their number and phone shortly after enrollment because they moved to other offices. `M1151` changed after the 8th of 31 calls and `F1160` after the first of 27 calls. After the change they consistently called from the same numbers, though the new numbers were different from the enrollment numbers. The remaining different-number calls (one per condition) were made by one female subject.

The corresponding proportions of impostor calls from a different number than the enrollment number are 23.9% in landline/office, 24.9% in mobile/office and 25.4% in mobile/hall. Most of these calls were made from a different number because either the target had left the Department before (four targets, 63% of different-number calls) or at the end of (two targets, 8% of different-number calls) the impostor data collection period, or because targets replaced their mobile phone during the same period (five targets, 15% of landline/office and 25% of mobile different-number calls). In all these cases, the enrollment mobile phone of the target was not available. Impostor subjects were then instructed to use non-enrollment phones in all telephone conditions. Moreover, in the landline/office condition, 9% of the different-number calls were made against target `F1160` from the same telephone that the target herself used in most of her true-speaker attempts. The remaining different-number calls (2-5 calls per condition) were made for other reasons, such as by mistake or by curiosity from impostor subject.

Note that a check for the same A-number in two calls doesn't guarantee that the same telephone instrument was used, and vice versa, but it is our belief there is a very good correspondence between telephone number and telephone instrument in our data.

Test file selection As an illustration of what definition b of "accepted text status" means in practice, Table 6.2 shows a categorization of transcription patterns for files that were skipped when selecting files for the S2b_G8 test set. The categories show what mistakes made by system or user caused files to be omitted from the test set. The majority of cases (62%) are omitted because one or more of the expected words are missing from the recorded file. This may be due to a speaker forgetting to say the digits (e.g. after the system responded to the previous attempt that it didn't perceive a name, then the subject often responded with only the name), speakers saying only their first name instead of the full name, or the system failed to record the entire (complete or incomplete) spoken utterance either because the speech detector pre-maturely signaled the end of the utterance or a programmed maximum recording time came to an end before the speaker finished the utterance. In many of these cases, missing words are due to a combination of a lacking capability in the system and unexpected behavior from the subject (e.g. a speech detector unaware of grammatical constrains in combination with very slow speech, a late start, or long pauses between words), and therefore the division

of cases between “user mistake, not repaired” and “system mistake” is somewhat arbitrary.

Table 6.2: A categorization of transcription patterns for files that were skipped when selecting files for the S2b_G8 test set according to "accepted text status" alternative b. The total number of cases is 142 from 131 different sessions. The total number of sessions included in the S2b_G8 test set is 5764. Transcription patterns are constructed by replacing the first and last names of the target speaker with `F` and `L`, digits with `D`, markers for extralinguistic sounds with `extral`, and other noise markers with `noise`. Patterns in the example column are delimited by a comma.

Category	Cases	Fraction	Example patterns
User mistake, not repaired			
missing digits part (or truncated signal)	26	18.3%	`F L`
extra out-of-vocabulary words	9	6.3%	`hallå F L D D D D D D`, `ska man trycka nånstans eller F~`
wrong digits	6	4.2%	`F L D D D D D`
wrong pronunciation of name or different name form	6	4.2%	`F *L D D D D D`, `Alexander L D D D D D` (`F=Alec`)
digits spoken in other language	3	2.1%	`F L zero four eight six five`
missing last name	3	2.1%	`F D D D D D`
speaker gives up	3	2.1%	`noise *F extral`
digits spoken as numbers	2	1.4%	`F L noise DD DDD`
mispronounced digit(s)	1	0.7%	`F L *D D D D D`
User mistake, with repair			
extra in-vocabulary words or fragments thereof	21	14.8%	`*L L D D D D D`, `F L D *D D D D D`, `F L D D D D D D`, `F D D D F noise L D D D D D`
System mistake (speech detection error or time-out)			
truncated signal or missing words	59	41.5%	`~F L D D D D D`, `noise F L D D D D~`, `noise F~`, `F L D D`
Other			
other	3	2.1%	

Table 6.3: 95% pre-trial confidence intervals for an observed false reject error rate on PER test sets given a "true" population error rate $p = 3\%$ or $p = 1\%$ and $N' = N^*/k = M\lfloor\bar{n}\rfloor^*/(1 + (\lfloor\bar{n}\rfloor^* - 1)\rho)$ independent tests for four choices of ρ. Intervals with lower limit 0.0 are one-sided confidence intervals, while others are two-sided confidence intervals.

test set	M[b]	N[c]	$\lfloor\bar{n}\rfloor^*$[d]	N^*[e]	95%[a] confidence interval (%) $\rho = 0$	$\rho = 0.1$	$\rho = 0.2$	$\rho = 1$
$p = 3\%$								
T2b_G8	54	4643	50	2700	2.3–3.7	1.5–4.6	1.2–5.2	0.0–7.4[f]
T2b_Q:c	27	977	36	972	1.9–4.1	0.9–5.6	0.0–5.8	0.0–7.4[g]
$p = 1\%$								
T2b_G8	54	4643	50	2700	0.6–1.4	0.2–2.0	0.0–2.0	0.0–3.7[h]
T2b_Q:c	27	977	36	972	0.4–1.7	0.0–2.3	0.0–2.5	0.0–3.7[i]

[a]Due to the discreteness of the binomial distribution, actual confidence levels for intervals in the table vary between 95.4% and 98.3% (cf. Section 2.5.2).
[b]number of target speakers
[c]total number of tests in the set
[d]average number of tests per target (floored) after truncating right tail in Figure 6.4a
[e]adjusted total number of tests in the set ($N^* = M\lfloor\bar{n}\rfloor^*$)
[f]confidence level 97.7%
[g]confidence level 95.4%, i.e. same limit as with T2b_G8 but with lower confidence
[h]confidence level 98.3%
[i]confidence level 97.0%, i.e. same limit as with T2b_G8 but with lower confidence

6.3.4.6 Statistical significance

Table 6.3 shows 95% pre-trial confidence intervals[3] for observed overall false reject error rates for pooled target speakers on the single-condition gate/hall test set and each of the condition-parallel test sets for assumed "true" population error rates 1% and 3%, respectively. Confidence intervals are based on the assumptions made in Section 2.5.2 and four cases of choosing a value for the intra-speaker correlation coefficient ρ in Eq. (2.20). Here we use

$$N' = \frac{N^*}{k} = \frac{M\lfloor\bar{n}\rfloor^*}{1 + (\lfloor\bar{n}\rfloor^* - 1)\rho} \tag{6.1}$$

where M is the number of targets in the test set and $\bar{n}$ is the average number of tests per target. $\lfloor\bar{n}\rfloor^*$ is the average number of tests per target rounded downwards ($\lfloor\cdot\rfloor$) and adjusted for the fact that a few targets have very many tests in T2b_G8 as shown by Figure 6.4a. We (somewhat subjectively) chose $\lfloor\bar{n}\rfloor^* = 50$ for this test set and used the unadjusted $\lfloor\bar{n}\rfloor$ for the other test sets. Note that with (6.1), N' tends to M/ρ as $\lfloor\bar{n}\rfloor^* \to \infty$.

[3]Confidence limits in the table were computed with the `qbinom` function in the R software (http://www.r-project.org/).

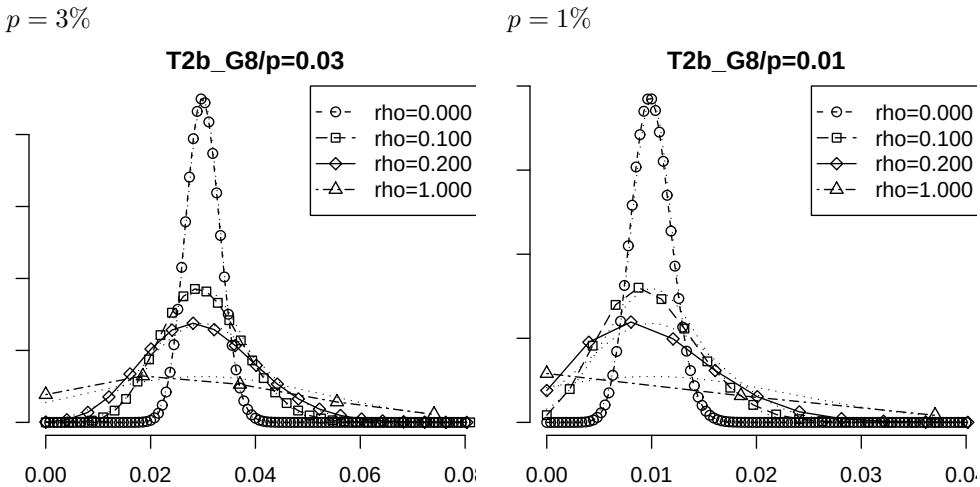

Figure 6.5: Binomial distributions used to compute confidence intervals for experiments on the T2b_G8 test set assuming "true" population false reject rate 3% or 1% and four choices of ρ. The normal approximation to each binomial is shown as a dotted line.

The first and the last case of choosing ρ in Table 6.3 are the two extremes in the assumption of partial dependence between observations discussed in Section 2.5.2.1, where we argued that the best choice of ρ should be somewhere in between the two extremes. Table 6.3 therefore also includes the cases $\rho = 0.1$ and $\rho = 0.2$. Unfortunately, we do not know at this point[4] which of these cases best describes reality. The table merely gives a perspective on the meaning of independence assumptions in terms of their effect on the length of confidence intervals. Figure 6.5 plots the binomial distributions from which confidence intervals in Table 6.3 for the T2b_G8 test set are computed, along with their normal approximations. Note that there are only a few points within the interesting range of observed error rate on the (discrete) binomial distributions with variances scaled according to $\rho > 0$, and resulting confidence limits are quantized by these points.

Table 6.4 shows example pre-trial confidence intervals for observed false reject rates for individual target speakers. It shows 95% confidence intervals for assumed "true" error rates 1% and 3%, respectively, and for a range of n (number of tests per target) spanning approximately (except for $n = \infty$) those values occurring for target speakers in the T2b_G8 test set as shown in Figure 6.4a. To calculate k of Eq. (2.17) we adopt (2.20) even though the latter was motivated by the beta-binomial distribution in the context of estimating an overall false reject rate from

[4]In Section 10.5.1 values for ρ are computed for post-trial experiments on the PER corpus, and Section 11.2.2 discusses the choice of ρ for pre-trial confidence intervals.

Table 6.4: 95% pre-trial confidence intervals for an observed false reject error rate for a single target speaker given a "true" error rate $p = 3\%$ or $p = 1\%$ for combinations of N and ρ_1. Intervals with lower limit 0.0 are one-sided confidence intervals, while others are two-sided confidence intervals.

	95%[a] confidence interval (%)			
$N = n$	$\rho_1 = 0$	$\rho_1 = 0.005$	$\rho_1 = 0.010$	$\rho_1 = 0.020$
$p = 3\%$				
∞		1.0–5.5 (96.5)	0.0–6.0 (96.9)	0.0–8.0 (98.3)
300	1.3–5.0 (96.0)	0.0–5.8 (97.1)	0.0–6.7 (97.5)	0.0–7.1 (96.3)
200	1.0–5.5 (96.5)	0.0–6.0 (96.9)	0.0–6.1 (95.2)	0.0–7.5 (96.9)
100	0.0–6.0 (96.9)	0.0–6.1 (95.2)	0.0–8.0 (98.3)	0.0–9.1 (98.3)
50	0.0–8.0 (98.3)	0.0–7.5 (96.9)	0.0–9.1 (98.3)	0.0–8.0 (96.2)
20	0.0–10.0 (97.9)	0.0–11.1 (98.4)	0.0–12.5 (98.9)	0.0–14.3 (99.2)
$p = 1\%$				
∞		0.0–2.5 (98.4)	0.0–3.0 (98.2)	0.0–4.0 (98.6)
300	0.0–2.0 (96.7)	0.0–2.5 (96.7)	0.0–2.7 (96.0)	0.0–4.8 (99.1)
200	0.0–2.5 (98.4)	0.0–3.0 (98.2)	0.0–3.0 (97.1)	0.0–5.0 (99.2)
100	0.0–3.0 (98.2)	0.0–3.0 (97.1)	0.0–4.0 (98.6)	0.0–3.0 (95.7)
50	0.0–4.0 (98.6)	0.0–5.0 (99.2)	0.0–3.0 (95.7)	0.0–4.0 (97.4)
20	0.0–5.0 (98.3)	0.0–5.6 (98.6)	0.0–6.3 (98.9)	0.0–7.2 (99.2)

[a]Due to the discreteness of the binomial distribution, actual confidence levels for intervals in the table vary as shown by parentheses in each cell (cf. Section 2.5.2).

multiple speakers with multiple tests. In (2.20) the coefficient ρ balances the correlation between speakers on the one hand and between trials from the same speaker on the other. It also widens confidence intervals to compensate for the distribution of individual false reject rates among targets in the test set. In the case of estimating a false reject rate for a single speaker, our ρ should correspond to correlation between single-speaker trials only, and possibly widen intervals because of inter-trial variation in the "true" underlying false reject rate, which we expect to be less than the corresponding inter-speaker variation. Thus we expect appropriate values for ρ in the single-speaker case to be smaller than in the multi-speaker case. To emphasize this difference we denote ρ in the single-speaker case as ρ_1. Thus, to calculate an equivalent number of attempts from the binomial distribution we use

$$N' = \frac{n}{1 + (n-1)\rho_1}. \tag{6.2}$$

Table 6.4 includes four choices of ρ_1, where $\rho_1 = 0$ corresponds to the case where all tests from a given target are independent and that the assumptions behind the error generation model motivating the use of the binomial distribution are assumed true (these assumptions were discussed in Section 2.5.2.1). Figure 6.6 shows plots of the binomial distribution (from which confidence intervals in Table 6.4 were

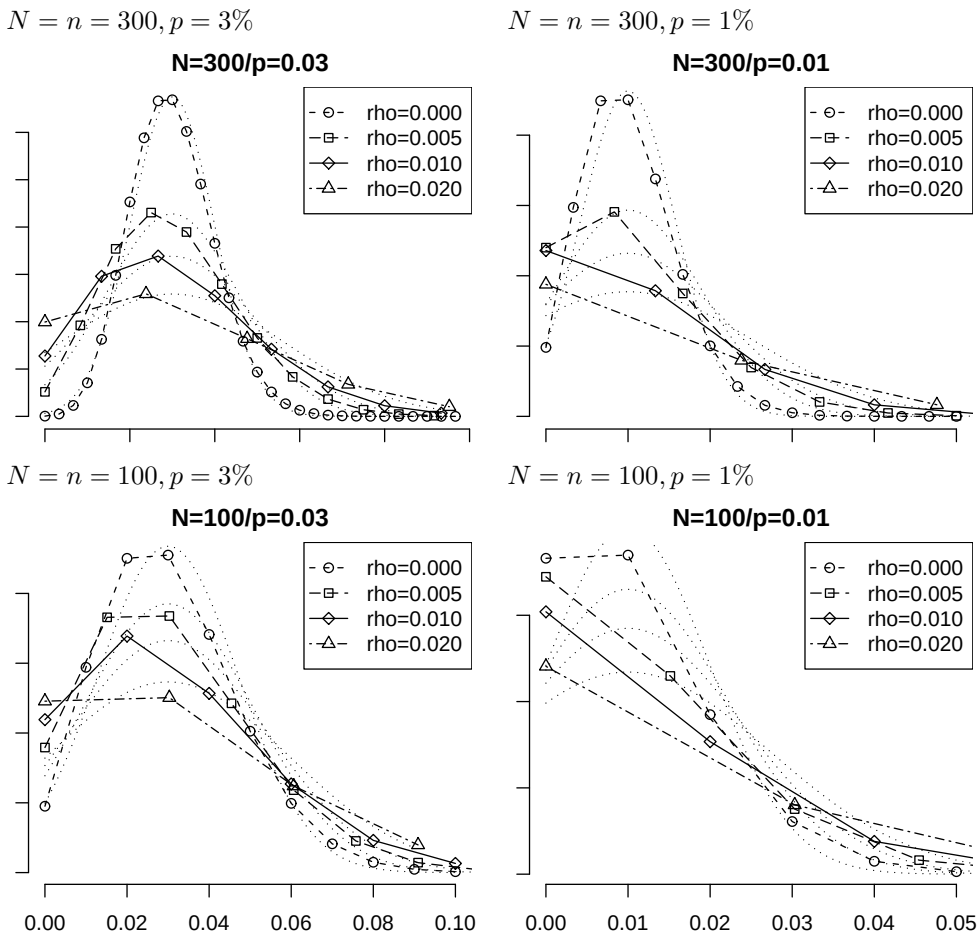

Figure 6.6: Binomial distributions used to compute confidence intervals for experiments on individual target speakers assuming "true" population false reject rate 3% or 1% and four choices of ρ_1. The normal approximation to each binomial is shown as a dotted line.

computed) and their normal approximations for $n = 300$ and $n = 100$. Also for this single-speaker case, we don't know what values for ρ_1 are appropriate. Values for the table were selected through our prior belief and after studying distribution plots like those in Figure 6.6. Note that under our model (6.2), and assuming $\rho_1 > 0$, the width of confidence intervals is bounded from below by the value of ρ_1 no matter how many trials are observed for a target speaker, since N' tends to $1/\rho_1$ as $n \to \infty$. These bounds are shown in the table for $n = \infty$.