GANDALF - A SWEDISH TELEPHONE SPEAKER VERIFICATION DATABASE

Håkan Melin

Dept. of Speech, Music and Hearing, KTH, Stockholm
email: melin@speech.kth.se

ABSTRACT

The Gandalf speech database has been designed for use in research on automatic speaker verification. 86 customer speakers have been recorded in up to 24 telephone calls per speaker during a period of up to 12 months, and an additional 100 impostor speakers are currently being recorded. In addition to speech files, Gandalf includes a relational database with a twofold function: it stores information on subjects and calls, and it is a tool for making quantitative and qualitative analyses of speaker verification test data.

The customer speaker part of the database is described, and some of the motivation behind the design is given. A small speaker verification experiment is then described that demonstrates how test results can be qualitatively analyzed using the relational database.

1. INTRODUCTION

Speaker verification (SV) is the task of accepting or rejecting the identity claim of a speaker based on recognition of the speaker's voice. In making the verification decision, two kinds of errors may occur: a false rejection of a genuine customer or a false acceptance of an impostor. The false rejection error rate for a customer will depend on the variability within the customer's voice (intra-speaker variability), while the false acceptance error rate will depend on the similarities between the customer's voice and the voices of the impostors attempting access to the customer's account (related to inter-speaker variability). In practice, the measurable speech signal will contain more variability than that related to the speaker himself. If the speech signal is transmitted through a telephone channel, for instance, the characteristics of different handsets will add to the measured intra-speaker variability.

One necessary resource for research on speaker verification is a speech database. The main difference between a database targeted for speaker verification and one targeted for speech recognition is the need of intra-speaker variability coverage. The verification database must include multiple recordings from each speaker. The recordings must be spread in time to capture both long-time changes and colds, sore throats, mood, and other sources of short-time variation.

This paper describes a Swedish speech database, Gandalf, targeted for research on speaker verification. Other speaker verification databases are described in [2], [3], and [4].

2. MOTIVATION

The database was designed for research on speaker verification in telephony applications. The three main design criteria were

1. to include both telephone line variation and intra-speaker variation,
2. to allow for a comparison between SV-systems with different text dependence, and
3. to enable an analysis of the significance of effects from different sources of variation in the speech signal.

The motivation for those three criteria will now be outlined.

Some SV databases have been designed to cover either long-term variations in the speaker or telephone handset variation. The reasons for including only one of the two are to isolate one source of variation in the experiments, and to limit the size of the database. Gandalf includes both types of variation. However, in order to enable a separation of variations due to speaker and handset, the following procedure was used: the subjects make half of the calls (every second call) from a "favorite handset", and the rest of the calls from different handsets in different environments.

Speaker verification systems are usually classified as being text-dependent, text-independent, or text-prompted. With a text-dependent system, the user will have to utter the same phrase in both enrollment calls and test calls. With a text-independent system, on the other hand, the utterance in a test call does not have to be the same as in the enrollment call. Finally, a text-prompted system will prompt the user for a random utterance to be repeated. In the text-prompted system, the text is not the same in enrollment and test calls, but the text is known to the system in each case.

Systems with different text dependence may differ in many aspects, such as user acceptance, system complexity, resistance against impostor attempts with tape recordings, and pure verification performance, where the last one is the main target for experiments with a speech database. To enable a comparison between these systems, the database should allow for testing of each system during the same conditions, which is only possible if all kinds of text are available in each call.

Most investigations on SV-systems have been quantitative. The general characteristics of a database are described and results are summarized in overall error rates. Such experiments will not answer questions about what happens if a user gets a cold, if he calls from a mobile phone, or if there is significant background

noise, etc. In order to take the analysis further, a more qualitative assessment of the SV-system is needed. Such an assessment is possible if more sources of variation are documented in the database, and the results from system tests are correlated with this information.

3. DESCRIPTION

3.1. Subjects

Subjects in the database were either recorded as customer speakers or impostor speakers. The customer speakers made multiple calls while impostor speakers made only two calls.

The customer subjects were mainly recruited from the employees at KTH and Telia Research AB, and from their friends and families. Subjects for the impostor part are currently being recruited and recorded, so no specific data on those subjects can yet be given. However, many of the impostor subjects are being recruited through the customer subjects with the request that people with some similarity be recruited, such as a close relative or someone with the same particular dialect.

The customer part contains 86 subjects. There are 48 male and 38 female subjects, with an age distribution as displayed in Figure 1. The dialect distribution is heavily biased towards the Stockholm region, with 52 of the subjects coming from Stockholm and 12 coming from the nearby city of Eskilstuna. The remaining 22 subjects speak various distinctive dialects.

3.2. Calls

Two types of calls were recorded: enrollment calls and test calls. The enrollment calls were longer in order to collect enough speech material for enrollment into an automatic SV-system. The duration of the entire call was about 7 minutes per enrollment call and 2½ minutes per test call. The calls were recorded with an automatic procedure through an ISDN-line.

The recording of the customer subjects were done in two parts. The first part included three enrollment calls and 14 test calls; one call per week during four months. The series of calls started with two enrollment calls, one from the favorite handset and one from another handset. Then the test calls were made with every second call from the favorite handset and the rest from other handsets. Finally, the third enrollment call was made, again from the favorite handset. Figure 2 shows statistics on the number of handsets used by subjects, and Table 1 shows the distribution of those handsets over handset type. As many as 82 of the subjects completed all the 17 calls (remaining subjects completed 4, 7, 14, and 16 calls each). The total number of calls in part one is 1435, corresponding to approximately 30 hours of speech.

The second part included seven calls from the favorite handset with a one-month interval between the calls. 67 of the subjects from part one volunteered to part two, and 57 of them have completed the 7 calls. With the second part, intra-speaker variation during up to 12 months has been included in the database. The second part contains approximately 420 calls with 8 hours of speech.

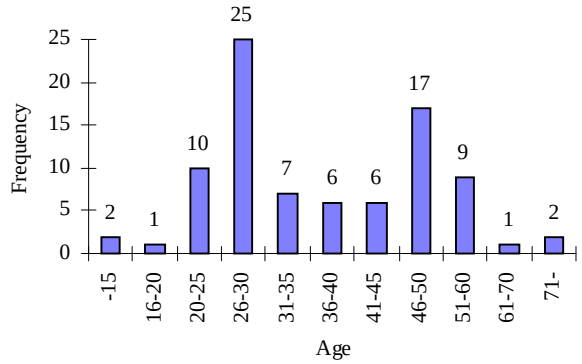

Figure 1: Age distribution among the 86 customer subjects.

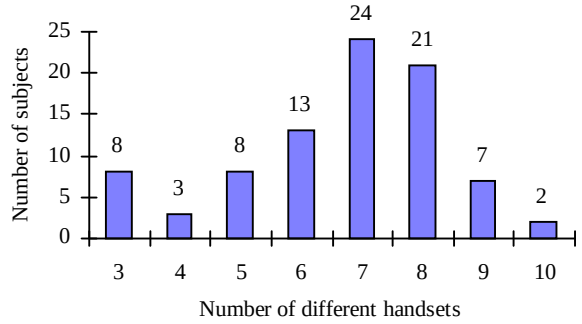

Figure 2: The histogram shows how many customer subjects called from a certain number of different handsets (including the favorite handset). The favorite handset has changed for some subjects.

Type	favorite handsets	non-favorite handsets	non-favorite calls
Stationary, button	84	259	394
Stationary, dial	1	43	65
Cordless	1	27	44
Mobile, GSM		58	81
Mobile, NMT		17	25
Pay phone, card		59	61
Pay phone, coin		12	12
Speaker phone		6	7
ISDN-phone		6	7

Table 1: The number of handsets of each type used by customer speakers, and the total number of calls made from a handset of the respective type in the non-favorite section.

The impostor subjects are recorded in two calls: one enrollment call and one test call. The two calls are made from two different handsets and no specific time is requested for the interval between the calls. Normally, an impostor speaker would have to be

recorded only once, and the recording would be used only to make impostor attempts on the identities of the customer speakers. However, recording one extra call per speaker involves little additional effort, gives the possibility of making limited customer tests with the impostor speakers, and will also give recordings from the same impostor with two different handsets. 100 impostor speakers will be recorded, resulting in 200 calls and approximately 7 hours of speech.

3.3. Text

The recorded phrases include short sentences and digit strings of various length. Some of the phrases are the same for each call while some are different. Some of the phrases are read from a script and some are given to the subjects by a voice prompt at recording time. In the last five calls the subjects were also asked to speak freely for 15 seconds. Table 2 shows the exact composition of phrases in each call. The scripted phrases are the same for all subjects (except the 7-digit identification number).

4. DOCUMENTATION

Apart from audio files, the Gandalf database also contains a relational database (RD). The RD application has two functions: it stores all the available information about subjects and calls, and it is a tool for analyzing results from SV tests.

4.1. Stored information

Information stored in the RD has been retrieved from the subjects through filled-out return forms and through post-processing of the recordings.

The subjects filled out one *voice form* plus one *call return form* for each call. The voice form had questions about fixed characteristics, such as gender, age, dialect, smoking habits, and education, while the call return form had questions about call-specific conditions, such as the handset, the room where the call was made, background noise, and colds and other conditions that potentially could affect the voice.

The post-processing has so far involved manually checking the recorded files for correctly spoken phrases. Files where there is some deviation from a normal pronunciation of the text have been sorted out for further annotation. Examples of such deviations are: repeated, extra, missing or mispronounced words.

4.2. Analysis tool

Procedures for computing SV results have been implemented within the RD. The results can easily be correlated with other stored information. In this way the RD can be used as an efficient analysis tool in making detailed analyses of SV-systems. Examples of questions which can be investigated are: What happens to the false rejection rate when a customer always calls from his favorite handset compared to when he uses different handsets? Is a system robust to a particular type of background noise? How much larger is the probability of a false rejection when the customer has a cold? How does the false rejection rate change with time from enrollment? Is the probability of a successful impostor attempt greater if the impostor is a close relative?

The statistical significance of the answers to questions like these will depend on the number of occurrences of an investigated phenomenon and the number of errors made by the system under test. Hence, different questions will have answers with different statistical significance. For instance, the number of calls where a subject has a severe cold is low (44 calls), and hence, conclusions on the influence of a cold will be relatively weak.

5. EXPERIMENT

A small text-independent speaker verification experiment was conducted to demonstrate the potential of the RD-based analysis tool. The SV method that was used was chosen for its simplicity and reproducibility. For this telephone speech task it gives high error rates, but the results still illustrate how a SV-system can be qualitatively evaluated as outlined in section 4.2.

5.1. Method

A second-order statistical measure (SOSM), the symmetric arithmetic-harmonic sphericity measure [1], was used in the test. This is a simple and computationally efficient measure of the similarity between two covariance matrices. The client model is a sample covariance matrix C_x computed from a 16 channel mel-spaced filter bank representation of the enrollment utterances. In a verification test, a test covariance matrix C_y is computed in the same way from the test utterances, and the SOSM produces a score s_c for the similarity between C_x and C_y.

The decision variable, s, is produced by normalizing s_c with the score s_w from a world model: $s = s_c/s_w$. A speaker-dependent *a posteriori* decision threshold was set to give a (mixed-gender) false acceptance rate of 10 % for each individual client.

5.2. Data

Only the customer part of Gandalf was used in the experiment. Client models were trained for each of the 86 customer speakers on the 10 varied sentences (30 s of speech) from call 1 (see Table 2). The two fixed sentences (6 s of speech) were used as test utterances. One within-speaker test was made for each of the available test calls (a total of 1605 tests). Call 99 (Table 2) for each customer speaker was then used in impostor tests against each of the other clients (a total of 7052 tests). The world model was built from a balanced set of 52 speakers in a separate database, the Swedish part of the Rafael database [5].

5.3. Results

The results on different sets of tests were computed with the analysis tool in the relational database. Some of those figures on false rejection (FR) rate are presented here as related to the two first questions put forth in section 4.2. The FR rate on a set of tests will be given as a percentage of the number of within-speaker tests in the set. This number of tests is given within parentheses

following the FR rate. The FR rate on the set of all calls was as high as 30% (on the set of 1605 tests).

Firstly, the FR rate on the set A={call from favorite handset} and the disjoint set B={call from non-favorite handset} were compared: 13% (995) versus 58% (610), indicating a very high sensitivity to a change in handset for the chosen SV method. Secondly, set A was divided into two disjoint sets A.1={subject reports no background noise} and A.2={background noise}: 9.4% (692) versus 21% (303). In particular, on a small subset of A.2 where the subjects reported continuous background music, the FR rate was as high as 53% (45).

The different types of handsets used in calls in set B was then studied. As expected, the FR rate for mobile phones was very high, 84% (25) for NMT and 85% (74) for GSM, but for public pay phones the case was even worse: 90% (70). It is worth noting that a big portion of these calls were made from a noisy environment. For stationary button phones the FR rate was 48% (339).

6. CONCLUSIONS

The Gandalf database has not yet been used in any large tests of speaker verification systems. Future tests will indicate the usefulness of the different aspects of its design.

It is believed that a qualitative test methodology which exploits additional information about subjects and recordings in a speech database will be useful for the understanding and improvement of speaker verification systems in the future. Further research will be carried out to develop this kind of methodology, especially on the statistical foundation of the analysis.

7. ACKNOWLEDGMENT

The author wishes to thank Antonio de Serpa-Leitão for much help with the recording and documentation process. The recording of the database and the work of the author have been supported by Telia Research AB.

8. REFERENCES

1. Bimbot, Magrin-Chagnolleau, Mathan (1995). Second-order statistical measures for text-dependent speaker identification. In: *Speech Communication 17*, 177-192.

2. Boves, Bogaart, Bos (1994). Design and recording of large data bases for use in speaker verification and identification. In: *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, April 5-7, 43-46.

3. Campbell Jr (1995). Testing with the YOHO CD-ROM voice verification corpus. In: *Proc. 1995 IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Detroit, Michigan, May 9-12, 341-344.

4. Godfrey, Graff, Martin (1994). Public databases for speaker recognition and verification. In: *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, April 5-7, 39-42.

5. Rosenbeck, Baungaard, Jacobsen, Barry (1994). Experience from the recording of Rafael.0: A 3000 speaker Scandinavian telephone speech database. In: *Proceedings, Nordic Signal Processing Symposium (NORSIG)*, Ålesund, Norway, June 2-4, 108-114.

			Call-ID							
Presentation	**Phrase**	**TD**	1-2	3-16	99	17-18	19-23	101	102	**Example**
Scripted	7 digit id. No.	x	1x2	2	1x5	1	1	1x5	1	0 8 8 0 3 2 5
	call No.		1	1	1	1	1	1	1	det tredje
	fixed sent.	x	2x5	2	2x5	2	2	2x5	2	Öppna dörren innan jag fryser ihjäl.
	varied sent.		10	4	10	4	4	10	4	Filmen med badbilderna har fastnat i tullen.
	1 digit		50							2
	3 digit seq.	x		6		6	6		6	6 5 8
	4 digit seq.	x		4	4x5	4	4	4x5	4	7 9 4 1
	5 digit seq.	x	25		25			25		1 4 4 7 2
Prompted	4 digit seq.			2		2	2		2	2 9 5 4
	5 digit seq.						2		2	3 4 0 8 9
	sent.			2		2	3		3	Barnen är lediga från skolan idag.
Free	free speech						1		1	
		Type	E	T	E	T	T	E	T	
		Part	Part 1			Part 2				
		Subject	Customer					Impostor		

Table 2: The number and types of phrases in different calls. *PxR* denotes *R* repetitions of *P* phrases. E and T in the Type-row denotes enrollment and test call respectively. An x in the TD-column means that the same text is used in every call, which allows for text-dependent tests.