# Speaker Verification Based on the German VeriDat Database

*Ulrich Türk, Florian Schiel*

Bavarian Archive for Speech Signals (BAS)
c/o IPSK, Ludwig-Maximilians-Universität München
Schellingstr. 3/II
80799 München, Germany
`{tuerk,schiel}@bas.uni-muenchen.de`

## Abstract

This paper introduces the new German speaker verification (SV) database VeriDat as well as the system design, the baseline performance and the results of several experiments of our experimental speaker verification (SV) frame work. The main focus is how typical problems using real-world telephone speech can be avoided automatically by rejecting inputs to the enrollment or test material. Possible splittings of the data sets according to network type and acoustical environment are tested in cheating experiments.

## 1. Introduction

This paper reports on recent experiments we carried out using the German SpeechDat style speaker verification (SV) database VeriDat. The focus of this presentation will be on the problems that arise when using "real world" telephone data in a SV application, namely an "under-representative" world model, varying acoustical environments and telephone channels, linguistic and meta-linguistic "noise", and "goats" that show an unusual high false rejection rate.

We tested several hypothesis on how the knowledge about certain properties of the enrollment or the test data may be exploited to yield a better performance. Furthermore, we analyzed some of the data sets in more detail to find out why certain speakers act as "goats" and others do not and how that behavior may be detected a priori.

The following section will give an overview of the features of the VeriDat corpus (first time reported here) and discuss some of the problems we found. Section 3 describes our SV system used as the framework for the experiments described in section 4. Section 5 presents and discusses the results of our experiments and the implications for our future work.

## 2. The VeriDat database

### 2.1. Overview

The VeriDat database was created by T-Nova, Deutsche Telekom Innovationsgesellschaft mbH, and the Bavarian Archive for Speech Signals (BAS) as a database for German speaker verification in fixed and mobile telephone networks. It is an extension of the standardized specification for speaker verification databases as published in the SpeechDat project ([1]). VeriDat contains an additional set of 19 recording items including number tripletts and spontaneous speech aside from the 21 items defined in the SpeechDat specification.

The main idea of VeriDat was to create a resource suitable for all kinds of speaker verification systems that covers the whole range of German dialects and uses different recording environments (quiet and noisy background) as well as different networks (fixed network and cellular phones).

The database comprises 7GB of speech material and is distributed on two DVD-Rs containing all signal and label files, support files and documentation. The speech material is stored according to the European ISDN standard as 8bit, 8kHz, A-law encoded and uncompressed data. The label files are formatted according to SAM ([2]) and contain recording and speaker information, the prompted text, the transcription and a small set of noise markers.

Each of the 150 speakers was recorded in 20 sessions with a minimum break of three days between the sessions. The recordings took mainly place in autumn and winter 1999/2000. Additional sessions were recorded to replace corrupt material found in the full-cover validation resulting in a nearly 100% error free database. The SpeechDat compatible part of the database has been validated successfully by the Speech Processing EXpertise center (SPEX) in Nijmegen, Netherlands[1].

### 2.2. Speaker population

The recorded speakers closely represent the German population with respect to the distribution of German accents (13 dialects plus one bin for foreign accent). The gender distribution within each accent group and within the five age groups is perfectly balanced. Some of the participating speakers are related to each other and their relationship (brothers, twins, etc.) is documented.

### 2.3. Recording conditions

The recordings can be divided by two sets of criteria: the recording environment (labeled "Quiet" resp. "Noisy" in the following text) and the network called from (labeled "Fixed" resp. "GSM"). The recording protocol defined for all speakers which session had to be recorded in which environment and from which network. The ratio of quiet to noisy sessions is independent of the partitioning in networks and vice versa (see table 1).

There had been no restrictions on the handsets used. Telephones using DECT technique but connected to ordinary fixed networks were treated as regular fixed network connections. The judgment about the degree of noise or quietness was left to the speakers though they were instructed by simple rules and sample recordings. This leads to a great variation of noise especially in the "Noisy" part ranging from static noise to loud

---

[1] www.spex.nl

|       | Fixed | GSM |    |
|-------|-------|-----|----|
| Quiet | 7     | 7   | 14 |
| Noisy | 3     | 3   | 6  |
|       | 10    | 10  | 20 |

Table 1: *Partitioning of recording sessions per speaker with regard to environment and network.*

cross talk (e.g. a small boy grouching about his father doing a telephone call).

### 2.4. Number triplets

For the experiments presented in this paper we selected only the triplets of two-digit numbers which are part of the VeriDat extension to the SpeechDat standard. Each session contains seven recordings of number triplets taken from a set of 140 triplets. This set is derived from the YOHO database ([3]) which uses 136 triplets and is extended by four additional triplets at the end. The different session structure between YOHO and VeriDat causes problems when selecting single sessions from the VeriDat database:

- YOHO uses 4 enrollment sessions with 24 triplets each and 10 test sessions with 4 triplets each. In each session the distribution of the different sub-words of the numbers ("sechzig", "drei", ...) are designed to be close to equal.

- VeriDat does not have predefined sessions for enrollment and test. All sessions contain seven triplets which are derived from the order given by the YOHO database.

These facts make it difficult to select a single VeriDat session for enrollment when using (like in our case) individual sub-word models. Some of the sessions show highly varying frequencies of the individual sub-words. The solution is to make a selection of seven triplets from two sessions with the same acoustic conditions (common type of environment and network).

## 3. System design

The previously described triplet recordings are used to build a a text-prompted speaker verification system based on sub-word-HMMs. For the feature extraction and HMM training/testing we use the HTK tools ([4]).

### 3.1. Experimental protocol

In order to set up a methodically sound experiment protocol for testing the open-set performance of the SV system, we divide the 150 speakers of the database into four sets:

- the client set (30 speakers). Individual models are build for these speakers.

- the world set (30 speakers). Used for building the world model(s) (might also be used for other score normalization methods e.g. cohort normalization).

- the impostor set (60 speakers). Used for testing the client models with non-genuine claims.

- the development set (30 speakers). Used for calculating various parameters (e.g. world model quality, a priori thresholds for the verification, ...).

The speakers of the four sets are chosen by a random selection scheme. In order to achieve a representative distribution of the speakers in each set, we test the generated four sets with regard

to an equal gender, accent and age group distribution and repeat the random selection until the constraints on the distributions are met.[2] If not stated otherwise we use all the recordings of a client speaker for testing which were not used for enrollment.

### 3.2. Preprocessing

Common to all experiments in this paper is the parameterization of the speech data. Speech features are calculated from a 14th-order LPC analysis with a Hamming window of 25ms length. The frame period is 10ms. A 1st-order pre-emphasis with a coefficient of -0.97 is used. 12 cepstral coefficients are derived from the LPC coefficients and joined with the unnormalized energy to a 39 features vector including delta coefficients and acceleration coefficients. In spite of long parts of non-speech in the recordings at the beginning and at the end we did not use a silence detection. Due to different noise levels the setting of a unique level threshold would not give predictable results. Instead we use three "silence" models to discriminate between speech and non-speech segments.

In order to get some insight in the recording conditions we measure the signal-to-noise ratio (S/N-ratio) of the speech material using the development set. Table 2 shows the results. The recordings with quiet background yield an increased S/N ratio by 4.6dB compared with the recordings having a noisy background. A comparison of the recordings using the two different networks shows no significant difference in the S/N-ratio.

| Quiet  | Noisy  | Overall |
|--------|--------|---------|
| 17.2dB | 13.6dB | 15.5dB  |

Table 2: *S/N-ratio of triplet recordings: calculated separately for the two environments and overall value.*

### 3.3. Modeling

Both the world model and the client models have the same HMM structure: they use a simple left-right topology without jump-over transitions. Each sub-word, e.g. "zwanzig" (twenty) or "ein" (one) from the triplet "ein-und-zwanzig" (twenty-one), is represented by a HMM model. The number of emitting states is determined by the number of phonemes in the canonical pronunciation of the sub-words. Two "silence" models are used for the leading and the trailing non-speech segment and an optional single state model is used for the pauses between the triplets.

The probability distributions are modeled by single Gaussians per state and a diagonal covariance matrix is used.

The world model is trained in un-supervised mode using a flat start scheme; the complete material of the world set is used, if not stated otherwise.

Practical constraints demand that a speaker verification system uses a minimum of enrollment material. In preliminary tests we found that building a client model from scratch based on 4 recording sessions (28 triplets) yields a poor model quality and thus a poor performance of the SV system. Therefore we chose to build the client models by performing one embedded Baum-Welch re-estimation step starting from the corresponding world model. This applies to each of our experiments.

Noises and transient sounds in the enrollment may have an impact on the model quality. In the VeriDat transcriptions additional labels for speaker noises (lip smacks, etc), for transient

---

[2]For a complete listing of the speaker sets see http://www.bas.uni-muenchen.de/Bas/SV.

and stationary noises and for signal truncations were used. We refer later on to "clean" recordings when using a sub-selection of the speech material without any of these noise labeled recordings. Note that even "clean" recordings may contain linguistic or meta-linguistic variation such as hesitations, mispronunciations, repetitions etc.

### 3.4. Score computation and performance measure

The computation of the final score starts with a forced alignment of a given utterance to a model sequence based on the expected triplet. Note that in our case the temporal segmentation of the utterance is predefined by the world model [5]. The sum over all log-likelihood scores of the speech segments is calculated, thus discarding the silence and pause models. The resulting score is normalized by the score of the world model which is calculated in the same way. The equal error rate (EER) for each speaker is computed with equal costs for a false acceptance (FA) and a false rejection (FR). All reported mean EERs are gender-balanced according to [6]. This implies using speaker-dependent thresholds for the decision making.

## 4. Experiments

Table 3 summarizes the parameters of all experiments. Aside from the base line performance on the German VeriDat corpus we wanted to test the general hypothesis that additional information about the context and/or acoustical environment may be exploited for a better performance. To test this hypothesis two series of cheating experiments were designed as described in the following.

### 4.1. Base Line Performance (Base)

To yield a base line performance we first trained on the corpus data with no regards to the meta information within the transcription. The world model was trained on 4200 recordings covering all recording conditions "Fixed/GSM", "Quiet/Noisy". Each client model was trained on an enrollment material consisting of 4 sessions covering all recording conditions in 28 recordings.

### 4.2. Data subsets (FilterFQ, FQ, F, Q)

To test the influence of the "difficult" recording conditions we repeated the base line experiment with data of some subsets of the recordings. The subset 'FilterFQ' uses only data from the recording conditions "Fixed" and "Quiet" and filters any recordings that are marked with a noise tag in the annotation. The sets 'FQ', 'F' and 'Q' contain unfiltered data sets from the resp. recording conditions. Note that the amount of enrollment data is kept equal except for the data set 'FilterFQ'.

### 4.3. Cheating Experiment: "Fixed" vs. "GSM"

Assuming that the type of telephone network is known to the system, we wanted to test in a cheating experiment, whether this information can be successfully exploited by either modeling two distinct world models or client models or both for the two recording conditions "Fixed" and "GSM". To ensure equal amounts of training and enrollment data the baseline experiment was modified: the world model is trained alternatively on a set of 5+5 or on 10 sessions of each of the two recording conditions; the same is done in the client model with 1+1 vs. 2 sessions. This results in four possible combinations of models (experiments 'Fixed/GSM').

| Exp. | World Model(s) | Client Model(s) |
|---|---|---|
| Base | one model: all cond. 4200 recs. | one model: all cond., 4 sess. 28 recs. |
| FilterFQ | one model: filtered recs. FixedQuiet 1358 recs. | one model: filtered recs. FixedQuiet, 4 sess. $\leq$ 28 recs. |
| FQ | one model: FixedQuiet 1470 recs. | one model: FixedQuiet, 4 sess. 28 recs. |
| F | one model: Fixed 2100 recs. | one model: Fixed, 4 sess. 28 recs. |
| Q | one model: Quiet 2940recs. | one model: Quiet, 4 sess. 28 recs. |
| Fixed/ GSM 1/1 | one model: Fixed, 5 sess. GSM, 5 sess. 70 recs. | one model: Fixed, 1 sess. GSM, 1 sess. 14 recs. |
| F/G 2/2 | two models: Fixed, 10 sess. 70 recs. GSM, 10 sess. 70 recs. | two models: Fixed, 2 sess. 14 recs. GSM, 2 sess. 14 recs. |
| F/G 1/2 | one model | two models |
| F/G 2/1 | two models | one model |
| Quiet/ Noisy 1/1 | one model: Quiet, 3 sess. Noisy, 3 sess. 42 recs. | one model: Quiet, 1 sess. Noisy, 1 sess. 14 recs. |
| Q/N 2/2 | two models: Quiet, 6 sess. 42 recs. Noisy, 6 sess. 42 recs. | two models: Quiet, 2 sess. 14 recs. Noisy, 2 sess. 14 recs. |
| Q/N 1/2 | one model | two models |
| Q/N 2/1 | two models | one model |

Table 3: *Test parameters for base line and cheating experiments (key word explanantions see text)*

### 4.4. Cheating Experiment: "Quiet" vs. "Noisy"

In the second series of cheating experiments we wanted to test the hypothesis whether the knowledge about the presence of strong background noise might help, if modeled in two distinct models. The method is equal to the cheating experiments 'Fixed/GSM'; the number of sessions for the world model are 3+3 vs. 6 and 1+1 vs. 2 for the client model respectively (experiments 'Quiet/Noisy').

## 5. Discussion of Results

Table 4 shows the EER for the experiments described in the previous section averaged for all client test and impostor data.

The base line performance of roughly 11% EER reduces dramatically to 2.8%, if only the data subset 'FQ' is used for the world and client model training. Using the filtered recordings ('FilterFQ') degrades the performance. The subsets using only "Fixed" or only "Quiet" data show EERs in between the

| Experiment | EER mean % |
|---|---|
| Base | 11.2 |
| FilterFQ | 3.4 |
| FQ | 2.8 |
| F | 7.2 |
| Q | 7.8 |
| Fixed/GSM 1/1 | 12.5 |
| Cheat Fixed/GSM 2/2 | 12.5 |
| Cheat Fixed/GSM 1/2 | 15.2 |
| Cheat Fixed/GSM 2/1 | 13.0 |
| Quiet/Noisy 1/1 | 12.6 |
| Cheat Quiet/Noisy 2/2 | 14.5 |
| Cheat Quiet/Noisy 1/2 | 15.2 |
| Cheat Quiet/Noisy 2/1 | 12.2 |

Table 4: *Results for the baseline and the cheating experiments.*

| Experiment | EER mean % |
|---|---|
| Base | 9.2 |
| FilterFQ | 1.6 |
| FQ | 1.9 |
| F | 5.5 |
| Q | 6.3 |
| Fixed/GSM 1/1 | 11.0 |
| Cheat Fixed/GSM 2/2 | 10.8 |
| Cheat Fixed/GSM 1/2 | 13.5 |
| Cheat Fixed/GSM 2/1 | 11.5 |
| Quiet/Noisy 1/1 | 10.9 |
| Cheat Quiet/Noisy 2/2 | 12.8 |
| Cheat Quiet/Noisy 1/2 | 13.5 |
| Cheat Quiet/Noisy 2/1 | 10.7 |

Table 5: *Results for the baseline and the cheating experiments. Worst three outlier speakers removed.*

most restrained case ('FQ') and the baseline system including all data.

Splitting the world model for 'Quiet/Noisy' seems to give a slight advantage over the standard modeling. This goes conform with the measured S/N-ratio (see table 2): Partitioning the data according to the environment gives a observable difference in the mean S/N-ratio, while partitioning the data by the network does not.

Some of the clients exhibit an extremely low verification performance caused by a broader distribution of log-likelihood (LLH) scores on their own testing data. This so-called "goat-like" behavior was further explored for three speakers who showed the worst EERs in most experiments. Two of these speakers misspeak once resp. two times in the enrollment; the third speaker did use two acoustically noticeable different GSM handsets for the "GSM" sessions (spk 0131). One of these outlier speakers is a 14 year old boy speaking rather soft and un-consistently in his speech rate. In addition, a lot of meta-linguistic noise occurs before and after his utterances.

Table 5 shows the mean EERs without the data of the three worst outlier speakers in each experiment. Basically the results are shifted by roughly 2%. However the small performance gain achieved by the experiment 'Cheat Quiet/Noisy 2/1' nearly vanishes in this case. Filtering the recordings (FilterFQ) gives noteable performance gain for the reliable part of the client population; the EER drops down to 1.6%. In contrast, the three outlier speakers suffer from this filtering, because compared to the average speakers a substantial part of their enrollment data is removed. But the kind of filtering applied here would require a supervised recording of the world speakers and – even worse – a supervised enrollment process, which is probably not feasible. However, an analysis of the filtered recordings might lead to an automated rejection rule for corrupt input data; this remains to future work.

Detecting a goat-like behavior like the three outlier speakers would therefore be helpful in a real world SV system. One possible way to perform such a detection using only the client's enrollment and the world model could be based on the variance of the LLH with regard to the world model. Figure 1 shows histograms of the world LLH of two clients: one is an outlier speaker (0131) without misspeaking in the enrollment, the other one is a speaker with an average EER. The outlier speaker shows a broader distribution of the scores than the average speaker. Our future work will include finding reliable criteria to predict such outlier behavior.
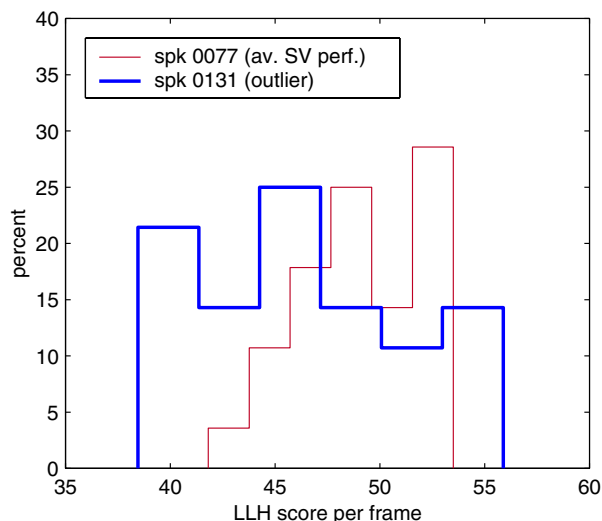


Figure 1: *Histogram of LLH scores per frame of the world model using client's enrollment data*

# 6. References

[1] SpeechDat: Public Specifications, *www.speechdat.org*

[2] ESPRIT "SAM" Project No 2589 : Speech Input and Output Assessment Methodologies and Standardization, *www.icp.inpg.fr/Relator/standsam.html*

[3] Campbell, Joseph P., "Testing with the YOHO CD-ROM voice verification corpus", ICASSP 1995, pp. 341-344, 1995.

[4] Steve Young et al (1995): The HTK Book. Cambridge University, htk.eng.cam.ac.uk

[5] J. Mariethoz, D. Genoud, F. Bimbot, C. Mokbel, "Client / World Model Synchronous Alignment for Speaker Verification", Eurospeech 1999, pp. 1979-1982

[6] F. Bimbot, G. Chollet, "Assessment of speaker verification systems" in "Handbook of Standards and Resources for Spoken Language Systems", Mouton de Gruyter, 1997