ON MODEL QUALITY AND EVALUATION IN SPEAKER VERIFICATION

Johan Koolwaaij and Lou Boves,

A^2RT, Dept. of Language and Speech
University of Nijmegen
koolwaaij@let.kun.nl, boves@let.kun.nl

Hans Jongebloed and Els den Os

KPN Research, Leidschendam
h.a.jongebloed@research.kpn.com
e.a.denos@research.kpn.com

ABSTRACT

This paper addresses differences between operational speaker verification (SV) systems and laboratory experiments in terms of performance and methods for measuring performance. It is concluded that operational SV systems need an indication of the quality of newly enrolled speaker models, to decide whether to re-enrol or request more enrolment material. We have investigated the impact of ASR errors on model quality. While attempting to design measures for the quality of speaker models we have developed a novel method for assigning weights to the contribution of models in accordance with their discriminative ability.

1. INTRODUCTION

Speaker Verification in large scale telecommunication services is not yet a fully mature technology. Field tests carried out in the Language Engineering projects CAVE [1] and Picasso [2], have shown that it is difficult to reproduce the performance found in laboratory experiments in the field. This is even true for laboratory experiments with realistic databases like SESP [1].

There are many differences between laboratory tests on prerecorded databases and operational services. One very important difference, that interferes with performance measurement proper, is the classification of recorded utterances. When recording databases the concepts of true speaker (client) and impostor are relatively clear: subjects follow some recording protocol that collects a number of utterances, part of which is classified as 'client' or 'impostor' for the sake of experiments that are designed after the data collection is completed. In an operational service things are different. Some clients appear to tamper with the system, perhaps to convince themselves that impostors cannot break into their accounts. Some kind of testing behaviour has been observed in virtually all field tests with SV. As a consequence it is difficult to classify utterances as belonging to a client or to an impostor on the basis of the data that can be collected by the application in the field. Thus, in operational services it is virtually impossible to evaluate the accept decisions made by the SV system. The only way to know that an accept was false is through complaints of the true owner of an account. In the services that we have tested so far such complaints could not be expected. Therefore, one is effectively limited to an analysis of the reject decisions. In the Free Access to DA service in the Netherlands (called GGS using an originally Dutch acronym), reject proportions of 7.3% were observed (although about half of these rejects could possibly be explained after auditory analysis of the utterances) [6]. Moreover, it appeared that a very large proportion of all rejects occurred with a small proportion of the clients. Yet, the number of problem clients was too large to assume that they were all regular 'goats' [3] [4].

In all database experiments carried out so far the utterances used for enrolment and test were checked for their correct transcription. In defining experimental protocols on the SESP database only utterances containing a correct token of a 14-digit calling card number were included. Thus, in our previous experiments ASR was only used to find the optimal segmentation of the utterances, using *a priori* knowledge about the digit sequence that was spoken. In an operational application callers may make hesitations or mistakes. Therefore, the task of ASR is different, and recognition errors can no longer be excluded. These errors are especially annoying if they occur in enrolment utterances, because then they give rise to corrupted models.

In this paper we investigate the problems encountered in the GGS service in more detail. First, we scrutinise the data, to allow us to make more dependable statements on the 'true' performance. In addition, we want to develop objective measures of the quality of speaker models, that can help to monitor the performance of an operational SV system. In doing so, we address the issues introduced above (the impact of ASR errors, and the existence of problem speakers). The attempts to better understand 'model quality' have resulted in proposals for improved models and performance.

There is one extremely important problem with deployment of SV that cannot be covered in this paper, because the relevant data to conduct in depth analysis are missing. In [6] it was observed that only 84% of the subjects who started using the Free Access to DA service were able to complete SV enrolment. For future applications it essential that the causes of the failures are better understood.

2. ANALYSIS OF THE GGS RECORDINGS

The GGS corpus comprises recordings of 210 customers who enrolled in the Free Access to DA service [6]. Two enrolment calls were used to collect four tokens of the 10-digit telephone number to build speaker models. For the present study we selected 56 female and 76 male customers who produced enough tokens of their 10-digit telephone number to allow meaningful off-line experiments.

The GGS recordings did not come as a corpus in the sense that all utterances were transcribed and checked for speaker identity. Therefore, initial speaker models were trained using the Picassoft system [2], and all utterances were subsequently processed by the resulting SV system. All utterances that were rejected were then checked for speaker identity, by comparing them auditorily against the corresponding enrolment utterances. It appeared that the large majority of the rejects were not normal client utterances. Part of these utterances were spoken with a clearly abnormal voice, probably in an attempt of the client to check whether the system would accept disguised voices. Another part of the utterances clearly came from impostors: acquaintances of the customer attempting to break into the account. One account appeared to contain enrolment calls of two different individuals (one male and one female). Apparently, this is a case of a family that did not understand the instructions that came with the invitation to enrol for the service. All the above mentioned accounts (2 females and 5 males) were excluded from further experiments.
Although cleaning up the data removed only 7 of the 132 customers the detection cost function (definition in paragraph 3.4) dropped with 56%. However, we still observed that the SV errors are not uniformly distributed over the remaining 125 customers: a few speakers had exceptionally high false reject rates. It is in the service provider's interest to detect these problem speakers as early as possible. While [4] and [3] focused on the characterisation and assessment of speakers, this paper will focus on the question how to detect these problem speakers a priori, that is during enrolment of the speaker models.

3. EXPERIMENTAL SET-UP

In order to investigate the impact of potential ASR errors and to develop measures of model quality, we carried out off-line experiments on the GGS and the SESP database. In doing so, SESP was used as a reference (because many results on SESP are already available).
The SESP corpus has been described in previous papers (e.g. [1]). It comprises 20 female and 22 male speakers. Only 'correct' tokens of 14-digit calling card numbers were used. However, due to the recording protocol and the recording conditions many utterances contain high background noise levels.

3.1. Model topology and features

Since the vocabulary of the databases is small (the Dutch digits /nul(0)/, /een(1)/,/twee(2)/, ..., /negen(9)/) text-dependent modelling can be used. Separate client models are trained for all digits that occur in this client's calling card (SESP) or telephone (GGS) number; for the remaining digits the world model is substituted to ensure that each speaker has a complete digit model set, so that all possible digit utterances can be matched with the speaker's model set. The topology used is left-to-right HMM, with 4 states per phoneme, 2 mixtures per state and diagonal covariance matrix. Acoustic features are 12 liftered zero-mean cepstra (LPC based) together with the log energy and their delta's and delta-delta's. In addition to the client models there is a single set of sex independent world models, one for each of the ten digits. Finally, there is a silence model, that is shared by all clients and the world. The variances of the client model are trained using the client's training data, but a variance floor vector [5] is set, which prevents variances from becoming too small.

3.2. Training and testing data

The world models, silence model and variance floor vector are trained on a set of 288 utterances (2296 digits) from the Dutch Polyphone corpus (12 utterances per gender and per Dutch province). For the SESP experiments each speaker was trained on 3 (14-digit) calling card numbers recorded in a single session. For the GGS experiments each speaker was trained on 4 (10-digit) telephone numbers recorded in two sessions. Thus, the number of digit tokens used for enrolment is approximately equal for the two databases. However, SESP performance is expected to suffer from the fact that all utterances are recorded in a single session.

The test set contains 9734 attempts for the SESP experiments (5691 same sex, 4043 cross sex; 1817 client, 7917 impostor) and 21232 attempts for the GGS experiments (12501 same sex, 8731 cross sex; 3565 client, 17667 impostor). In this paper we only report on the same sex experiments, since the error rate on the cross sex experiments is typically a factor 4 lower than for same sex experiments. Each available utterance is matched with the true client speaker and with a randomly selected set of 5 impostor speakers. Duplicate tests are removed if they occur.

3.3. Scoring

The output of each hypothesis test is a log likelihood ratio (LLR), defined as the client log likelihood minus the world (or non-client) log likelihood. The LLR on a global (e.g. utterance) level is the time normalised integral over the LLRs on frame level: $\overline{LLR} = \int_S LLR(t)dt / \int_S dt$, with S the set of frames. The choice of the set S may have a significant impact on SV performance. Fig. 1 shows that for the SESP experiment (with training and testing data as described above) defining S as the set of frames with non-zero LLR gives the best performance. Frames with a zero LLR probably have exactly the same model for the client and the world hypothesis (e.g. the silence model or a client model which is a copy of the world model). Segmentation of the training and testing utterances is carried out using procedures described in paragraph 4.

3.4. Evaluation

Evaluation of the verification results is in terms of the detection cost function (DCF), which is a linear combination of the false reject (FR) rate and the false accept (FA) rate:

$$\begin{aligned} DCF &= C_{\mathrm{FR}} \times P(\mathrm{FR}|\mathrm{Client}) \times P(\mathrm{Client}) \\ &+ C_{\mathrm{FA}} \times P(\mathrm{FA}|\mathrm{Impostor}) \times P(\mathrm{Impostor}), \end{aligned}$$

with C the cost of a verification error: $C_{\mathrm{FR}} = C_{\mathrm{FA}} = 1$, and $P(\mathrm{Client})$, the prior probability for a client, arbitrarily set equal to 50%.

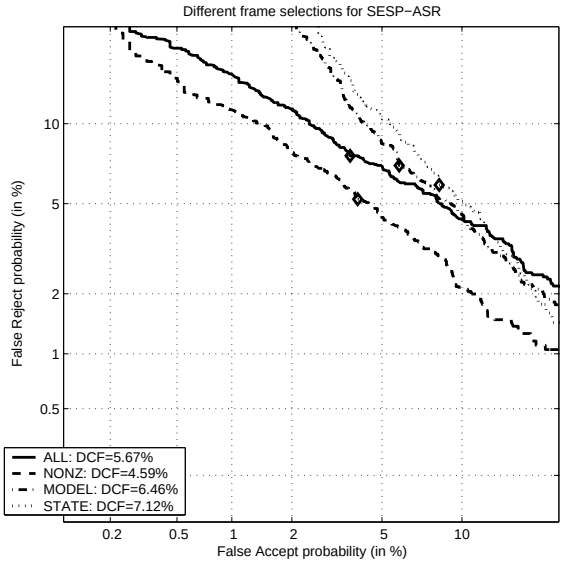

Figure 1: SV performance for different frame selection schemes: all frames (ALL), the frames with non-zero contribution to the LLR (NONZ), the nonz-frames assigned to the same model for both client and world (MODEL), the nonz-frames assigned to the same state for both client and world (STATE),

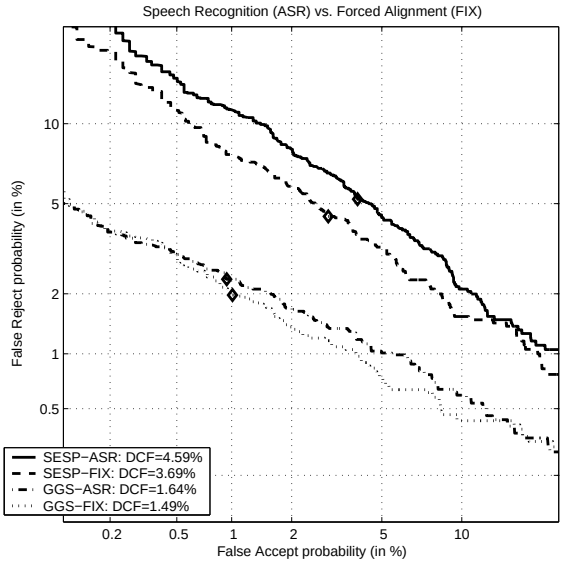

Figure 2: SV performance for using ASR versus performance using forced alignment

4. SEGMENTATION QUALITY

Training of text dependent speaker models needs a segmentation of the training utterances. Segmentation errors result in reduced model quality. Because ASR in telephone applications is usually speaker independent, it is to be expected that some speakers suffer more from the errors made by the ASR system than others. ASR errors are indeed concentrated in speech of a few speakers: for both the SESP and the GGS corpus 10% of the speakers account for about 45% of the ASR errors. To investigate how SV performance suffers from a worse model quality due to segmentation errors we compare two approaches:

ASR: Segmentation of the utterances is derived from a speech recogniser based on the SV world models.

FIX: Segmentation of the utterances is derived from a forced alignment of the world models with the transcription of the speech.

Fig. 2 shows the differences between SESP and GGS and between ASR and FIX. First, SESP shows much higher error rates than GGS. This is due to the single enrolment session, and because calls in SESP come from rather noisy environments and different handset types, while the calls in GGS come from mostly quiet home and office environments and callers almost always use the same handset. Second, the difference between ASR and FIX is rather large for the SESP experiment, and small for the GGS experiment, showing that errors in ASR hinders SV most on the noisy SESP, and less on the clean GGS. It was observed that most of the performance gain was obtained on the subset of speakers with relatively many ASR errors.

5. DISCRIMINATIVE ABILITY OF MODELS AND STATES

Model quality can only be described by indirect measures, that is measures derived from the model's *behaviour* and not calculated on the model directly. As a first approach we investigate one example of such an implicit measure. The distance Z between LLR scores on the client's training material and the LLR scores on a development set of impostor material for a given model m:

$$Z_m = \frac{\max\{0, \mu_C^m - \mu_I^m\}}{\sigma_I^m},$$

with μ_C^m and μ_I^m the mean LLR score of a model m on the client's training utterances and on a set of independent impostor utterances, respectively, and σ_I^m the standard deviation of the LLR scores on the set of impostor utterances. (This set of impostor utterances contains 60 scope numbers for SESP, and 70 telephone numbers for GGS.)

If the discriminative ability of a model is low we expect Z to be close to zero, and Z is large for a highly discriminative model. During training, the measure Z in combination with a threshold can be used to decide to re-enrol a model. However, Z can also be used to improve performance by weighting the contribution of the models according to their discriminative ability: if the test utterance t contains segments assigned to the models $[m_1, \ldots, m_n]$, the weight w_m for each model m can be written as

$$w_{m_i} = \frac{Z_{m_i}^p}{\Sigma_{j=1}^n Z_{m_j}^p}$$

and the LLR contribution of model m is weighted by w_m. The test utterance is segmented using the same approach as for the enrolment utterances (ASR or FIX). The exponent p controls the balance of the weighting scheme: $p = 0$

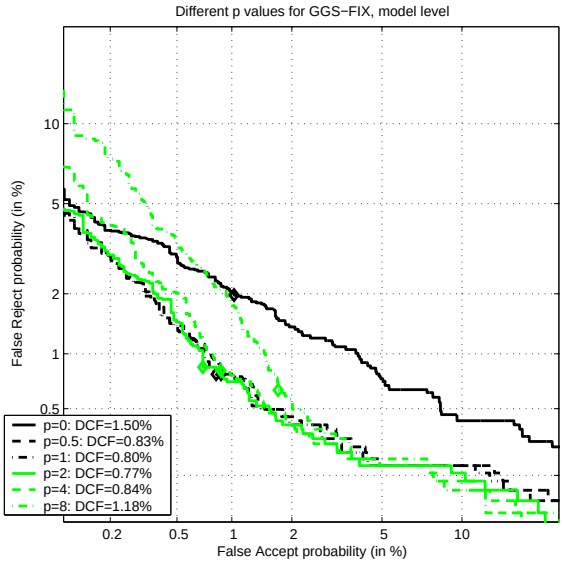

Figure 3: SV performance for GGS-FIX with the distance measure Z on model level for $p = (0, 0.5, 1, 2, 4, 8)$

gives equal weights to all models, if p increases the more discriminative models get a larger weight, until for $p = \infty$ only the most discriminative model has a weight equal to 1, while all other models have zero weight. In a similar way we can also compute the state quality, and apply weighting on state level.

Fig. 3 shows that introducing the model quality as weights for the model scores in the test utterance brings the DCF down from 1.50% for $p = 0$ to 0.77% for $p = 2$ in the GGS-FIX experiment. Also, for the SESP-FIX experiment the DCF reduces from 3.69% to 3.06% using model quality, and to 3.02% using state quality, as shown in Fig. 4. In general we can say that using the state quality measure gives a slightly better performance than the model quality measure, but is more unstable because the DCFs sky-rocket for higher p values. This is probably due to the fact that too much weight is given to a small part of the test utterance, and only a few states determine the final LLR on utterance level. Also parameter estimation problems on state level may play a role here. So preference goes to the more stable quality measure on model level.

6. CONCLUSIONS

In this paper we have investigated the difference in performance of an SV system between the laboratory and the field. In doing so it was shown that evaluating the performance of an operational SV system is not trivial. First, identity information of the caller may not always be available, because the clients may want to test how and if the SV system works. Second, users may have problems in adhering to the requirements of the system: they may make mistakes, hesitations, etc., not only during access, but also during enrolment. This requires powerful methods for detecting non-compliant utterances, which in their turn require very high performance ASR.

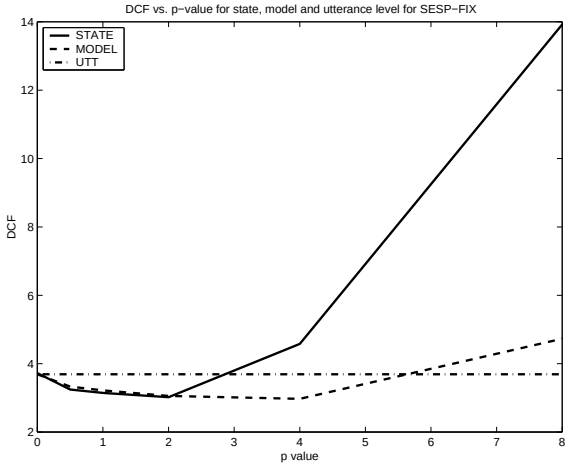

Figure 4: DCF as a function of p for SESP-FIX with the distance measure Z on state, model and utterance level

For real applications the proportion of 'goats' may be as important as the overall error rate. Even if only compliant utterances are taken into account, some clients still have relatively high false reject rates. It was shown that segmentation (ASR) performance may have a substantial effect on performance. In addition, a measure for the quality of speaker models is proposed, that can be used to decide whether newly enrolled models are adequate. The same technique, based on weighting the contribution of individual models according to their discriminative power, can also be used to improve performance per se. We have shown that incorporating the weighting improves performance substantially.

7. REFERENCES

[1] F. Bimbot et al. Speaker verification in the telephone network: research activities in the cave project. In *Proc. Eurospeech*, pages 971–974, Rhodes, 1997.

[2] F. Bimbot et al. An overview of the picasso project research activities in speaker verification for telephone applications. In *Proc. Eurospeech*, pages 1963–1966, Budapest, 1999.

[3] G. Doddington, W. Ligget, A.F. Martin, M.A. Przybocki, and D.A. Reynolds. Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proc. ICSLP*, pages 1351–1354, Sydney, 1998.

[4] J.W. Koolwaaij and L. Boves. A new procedure for classifying speakers in speaker verification systems. In *Proc. Eurospeech*, pages 2355–2358, Rhodes, 1997.

[5] H. Melin, J.W. Koolwaaij, J. Lindberg, and F. Bimbot. A comparative evaluation of variance flooring techniques on hmm-based speaker verification. In *Proc. ICSLP*, pages 2379–2382, Sydney, 1998.

[6] E. den Os, H. Jongebloed, A. Stijsiger, and L. Boves. Speaker verification as a user-friendly access for the visually impaired. In *Proc. Eurospeech*, pages 1263–1266, Budapest, 1999.