STEPS TOWARD THE INTEGRATION OF SPEAKER RECOGNITION IN REAL-WORLD TELECOM APPLICATIONS

Axel GLAESER[(a)] *and Frédéric BIMBOT*[(b)(c)]

(a) Ascom AG, Applicable Research & Technology Unit, 5506 Mägenwil, Switzerland
e-mail : Axel.Glaeser@ascom.ch
URL : http://www.ascom.ch/systec/ART

(b) ENST – Dept SIGNAL, CNRS – URA 820, 46 Rue Barrault, 75634 PARIS cedex 13, France
e-mail : bimbot@sig.enst.fr

ABSTRACT

The current market situation is characterized by a significant interest in speaker recognition functionalities in telecommunication systems (e.g. phone banking). This paper presents a field-test assessment of a speaker recognition algorithm, in a realistic context. Such field tests are particularly useful because the requirements for those real-world systems can be significantly different from those focused on by the research laboratories. Therefore, the results presented in this paper are divided into two groups. The quantitative ones describe the performance achieved in terms of Equal Error Rates (EER) as a function of the field-test conditions and different limitations on the enrollment and test duration. On the other hand, we discuss some innovative qualitative outcomes which are mainly based on non-technical but subjective impressions reported by the participants.

1. INTRODUCTION AND MOTIVATIONS

In the past few years, phone banking and phone shopping have become more and more popular. These new needs have mainly been initiated by the service providers (banks, etc…) which are looking for cost-efficient and secure solutions to enable their clients a flexible (24 hours a day, 7 days a week) and comfortable service access. In this context, the assessment of a speaker recognition system must take into account subjective factors that are directly linked to user acceptance criteria.

Beside this market-oriented view of the scenery, the technological part is characterized by different established approaches. The underlying algorithms classically used in speaker verification are mainly based on Long-Term Statistics, Hidden Markov Models and/or Neural Networks.

In order to merge the technological and marketing aspects into a real-world system, a collaboration was initiated in 1997 between Ascom, which supplies systems in the field of telecommunications and service automation, and ENST, a research unit which has been working for several years in speech processing, and in particular, speaker recognition.

The speaker recognition technology used for this cooperation is based on full-covariance Single-Gaussian Models (or Second-Order Statistical Models) of the speaker acoustic features [1], together with a symmetrized likelihood ratio measure [2]. This approach offers an interesting compromise between performance and computational effort.

Based on this technology, Ascom performed two field tests within the framework of text-independent speaker recognition. The two different scenarios under investigation are a telephone-based application and a test under poor conditions in terms of S/N-ratio.

In section 2, we present briefly the technology used in the field test. Then, in section 3, we describe the field test conditions and in section 4 the task, for both scenarios. In section 5, we review some subjective factors linked to the question of user-acceptance. In section 6, we present the results obtained, with corresponding quantitative performance. Finally, in section 7, we expose the outline of the optimal configuration that Ascom shall retain for the next version of its speaker recognition system.

2. SPEAKER RECOGNITION TECHNOLOGY

The speaker recognition technology used for this field test is based on full-covariance Single-Gaussian Models (or Second-Order Statistical Models) of the speaker acoustic features [1], together with a symmetrized likelihood score [2]. In this context, each speaker is modeled as the covariance matrix (*X*) of acoustic features computed from the training utterance. The same process is applied to the test utterance, which yields an other covariance matrix. The test utterance is also modeled as a covariance matrix (*Y*).

The matching score between *X* and *Y* is then computed as a function of the arithmetic, geometric and harmonic means (*a*, *g* and *h*) of the eigenvalues of matrix YX^{-1}. However, the

(c) Frédéric BIMBOT is now with IRISA (CNRS & INRIA), Campus Universitaire de Beaulieu, 35042 RENNES cedex, FRANCE. (e-mail : bimbot@irisa.fr)

computation of the score does not require the explicit extraction of the eigenvalues, as *a*, *g* and *h* can be obtained directly from the trace of YX^{-1}, the trace of XY^{-1} and from the determinants of *X* and *Y*.

In the experiments described in this paper, the acoustic features are 24 filter-bank coefficients in Mel scale, with long-term average substraction over the entire utterance (exactly as described in [2]). Though other approaches (such as Gaussian Mixture Models) have been shown to be more efficient in terms of performance, Single-Gaussian Models have an interesting property of economy in terms of computational requirements.

3. FIELD TEST CONDITIONS

We investigated two different scenarios for assessing the text-independent speaker recognition algorithm in terms of objective factors like the error rate and computational requirements as well as the more subjective factor of user acceptance. The main characteristics of these two scenarios are described in Table I, below.

	Scenario A	**Scenario B**
application	telephone	exhibition with poor S/N Ratio
bandwidth	300 Hz – 3400 Hz	0 – 4000 Hz
sampling freq.	8 kHz	8 kHz
quantization	8 bit	8 bit
number of participants	150	170
duration	5 months	2 days

Table 1 : The main characteristics of the two field test scenarios.

The enrollment phase in scenario A is characterized by an unsupervised procedure. The speaker is guided by a telephone dialog system without any help facilities. In contrast to this method, the enrollment in scenario B is done in a supervised manner.

Moreover, scenario A includes calls from different locations and with different telephone equipment, e.g. national and international calls, as well as different speech qualities caused by codecs for DECT or GSM. On the other hand, the speech quality in scenario B was mainly degraded by time-varying noise sources like messages over loudspeakers and low-altitude flights of helicopters and aircrafts.

The recordings of both scenarios include different languages (german in various dialects, english and french) as well as a large distribution of speaker ages (6 – 70).

Under these circumstances, the results reported in this paper reflect a good variety of real-world factors.

4. TASK

In the system's test phase defined by Ascom, the client has to claim his identity in form of a Personal Identification Number (PIN) by using the touch-tone function of his telephone (scenario A) or the PC-keyboard (scenario B). Afterwards, this claim is validated by the identification of his voice signature. This is done by scoring the test utterance against the models of all *N* members of the database. A ranking list is drawn up starting with the best matching model. When the client's model is within the first *n* ranks, he is accepted, otherwise rejected. Moreover, it is assumed that an impostor possesses a valid PIN and is member of the database. We simulate the impostor attacks by providing each test utterance against all identities in the database. In these respects, the task can be understood as "closed-set" speaker verification with exhaustive impostor attempts. Note also that, under this configuration, the False Acceptance Rate is strongly related to the number of participants ($f_a \sim (n\text{-}1)/N$).

For scenario A, we have 1'105 valid versus 123'760 impostor trials for the 113 participants. For scenario B, the figures are: 210 friendly trials and 35'490 attacks from 170 participants.

5. QUALITATIVE RESULTS

The outcome of the field trials can be distinguished into qualitative and quantitative results. In this section, we discuss the qualitative aspects and we focus on their concrete influence.

The qualitative results described in the following items are mainly based either on general observations drawn from the results or on subjective statements from the participants. They are mostly given as indications, and a more systematic study would surely be necessary to confirm these trends.

- The performance of the speaker verification system used in our experiments seems independent of the speakers age and the language, even if the latter is changed between enrollment and test.
- The false acceptance for family members of the same sex is higher than the average. This effect corresponds to the general experience, like confusing father/son or sisters on short telephone calls.
- As expected, the use of DECT- and GSM-codecs leads to higher error rates. This problem can be tackled using the following approach, which is currently under development: after the detection of the used codec, the stored speakers voice signature can be filtered in a predetermined way, so that both, enrollment and test files are homogeneous.
- People do accept relative long enrollments (30 seconds), when they are sufficiently informed about its use and when they are prepared for this task. Moreover, only 5% of the enrollment files in scenario A were absolutely unusable and had to be removed from further tests. This observation can be explained with the excessive demand on some participants to speak freely

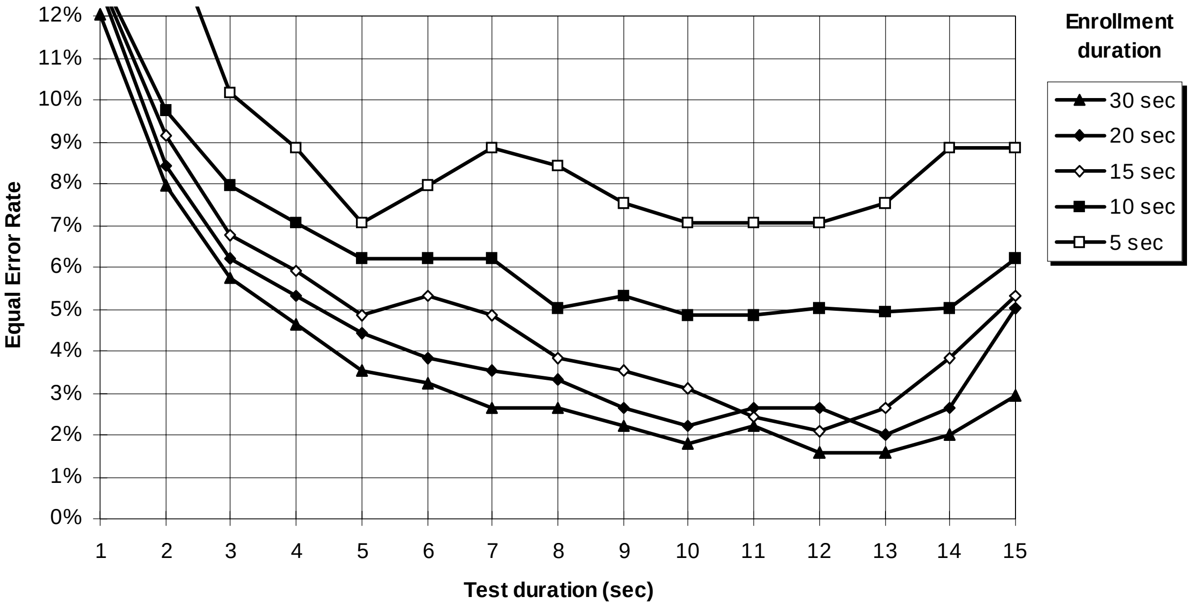

Figure 1: Equal Error Rates for scenario B (170 speakers) with variable durations for enrollment and test.

for 30 seconds during the automatic enrollment procedure. Those hardly corrupted files can be detected using a heuristic approach: after the enrollment, we tested the database using the second half of the same enrollment files. The likelihood values provided by the algorithm were analyzed and led to a reliable determination of malfunctioning recordings. This kind of cross-validation of the enrollment material appears to be a significant factor of overall performance improvement. Nevertheless, many of the remaining enrollment files of scenario A are of worse quality due to longer pauses and discontinuous speech, especially in the final 10 - 15 seconds of the enrollment.

- In contrast to the observation during the enrollment procedure, the participants were discontented about the recommended duration of the test phase which was chosen to 10 seconds. This is understandable, because it does not correspond to our human experience in normal speech communications. Therefore, we focused our quantitative analysis on optimizing this aspect.

6. QUANTITATIVE RESULTS

A conventional way of measuring performance in the field of speaker verification is to estimate the Equal Error Rate. This performance measure corresponds to the system operating conditions where the False Acceptance Rate is equal to the False Rejection Rate. It must be noted that this figure gives a rather optimistic idea of the system performance, as the decision threshold is set a posteriori on the test data, in order to reach this particular point.

The False Acceptance Rate (FAR) was calculated for different values of the threshold *n*, mentioned in section 3. For each of these values, the corresponding False Rejection Rate (FRR) was estimated after pooling all genuine trials together. The EER was chosen as the average of the FAR and FRR for the value of *n* that minimizes their difference. Note that, given the particular nature of the task treated here (“closed-set” speaker verification with exhaustive impostor attempts), the EERs obtained can not be compared with those corresponding to the more conventional task of speaker verification, with external impostors.

During off-line tests, we analyzed the influence of different durations of the enrollment and test files on the EER. Figure 1 shows that the enrollment duration seems more important for the system performance than the test duration. The latter leads to an EER which remains constant after 6 to 7 seconds in the case of scenario B. Compared with that, the same effect takes place in scenario A after 8 to 9 seconds of recognition.

The evolution of the EER in the course of time was another essential question for our study. Only the database acquired with scenario A is appropriate to answer this question due to the recordings which were spread over 5 months. Figure 2 tends to show that, for the algorithm that we used, the temporal drift in the speaker characteristics does not influence the performance, as the same levels of performance are obtained with recordings of October 1997 and March 1998.

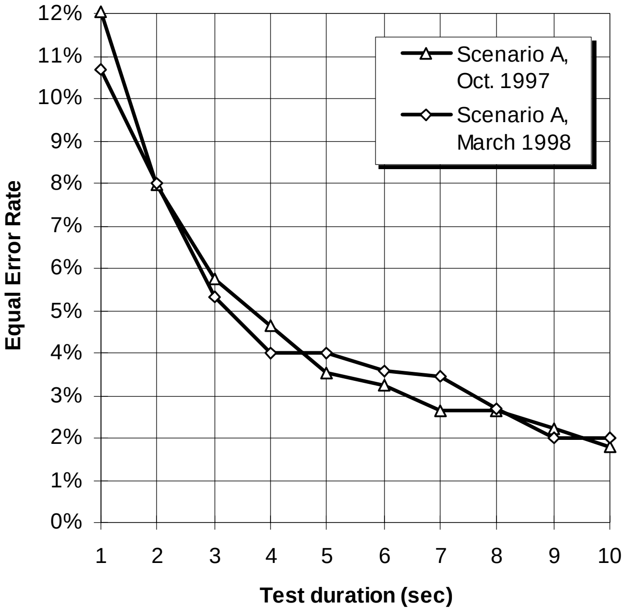

Figure 2: Equal Error Rates for scenario A obtained with recordings of October 1997 and March 1998.

Beside the performance measured in terms of EER, the computational performance and requirements of the system should be mentioned:

- The computation time per comparison is only 6 ms on a 233 MHz PC.
- 3 kB of memory is required for storing the speaker-specific parameters.

7. FUTURE IMPROVEMENTS

The impact of the described qualitative and quantitative results on the outline and assumptions for our planned, improved speaker recognition system can be summarized with the following items:

- The influence of the enrollment on the whole system performance is a significant factor. Therefore, the enrollment duration should be as long as possible. The participants will accept this, if they are informed in an appropriate way and if the process is made relatively natural (through some kind of pseudo-dialog, for instance). Moreover, the sufficient quality of an enrollment can be checked either in a supervised context or by some heuristic off-line methods. We will incorporate a reliable VAD (Voice Activity Detection) in order to get rid of dominant parts of voiceless background or channel noise.
- The duration of the test file can be limited to 6-8 seconds which is viewed as an adequate trade-off between user-acceptance and recognition performance.
- The system performance is constant for at least several months. So, no adaptation to speaker-specific features which are based on the first enrollment are required. However, it may prove efficient to incorporate progressively new (test) speech material in order to improve the voice signature by covering more variability in the reference model.

8. CONCLUSIONS

This paper reports on an initial study on the use of Single-Gaussian Models for speaker verification in a real life situation.

It seems that an important factor for the overall system performance is the duration and quality of the enrollment file. Its quality can only be guaranteed by a supervised recording or some off-line analysis and assessments, whereas the user acceptance for its duration can be achieved by appropriate user information and ergonomy.

Beside these critical aspects, we have noticed that the performance of the speaker recognition algorithm is mainly time-invariant which was observed for recordings over 5 months.

Thanks to this initial field test, we were able to identify a few key factors for the success of a speaker recognition application, which is a first step towards closing the gap between fundamental algorithms and real-world products.

9. REFERENCES

1. Gish, H., Krasner, M., Russell, W., Wolf, J. “Methods and experiments for text-independent speaker recognition over telephone channels”, *Proceedings ASSP, pp. 865-868, Tokyo, 1986.*
2. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L. “Second-order statistical measures for text-independent speaker identification”, *Speech Communication 17, pp. 177-192, 1995.*