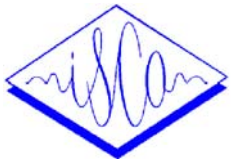


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DYNAMIC TEST DURATIONS FOR TEXT-INDEPENDENT SPEAKER VERIFICATION SYSTEMS

Axel GLAESER

Ascom AG, Applicable Research & Technology Unit, 5506 Mägenwil, Switzerland
Email: Axel.Glaeser@ascom.ch
URL: http://www.ascom.ch/systec/art

ABSTRACT

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 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. During field tests we found out that the main point of criticism was the test duration. While the enrollment duration was mainly accepted as a single, initial effort, the test duration of 10 seconds was not accepted as feature of a user-friendly system. Therefore, we developed a new approach based on Dynamic Test Durations (DTD) by taking the quality of speech and special characteristics of the classification process into account. This approach results in a decrease of the test down to 2,4 seconds (average).

10.21437/Eurospeech.1999-280

1. INTRODUCTION

The current market situation is characterized by a significant interest in speaker recognition functionalities in telecommunication systems (e.g. phone banking). The requirements for those real-world systems can be significantly different from those focused on by the research laboratories. The common way for assessing speaker recognition systems is a technology-driven approach, which mainly analyzes the objective performance in form of error rates, computational complexity and memory effort. Ascom, which supplies systems in the field of telecommunications and service automation, tries to take also more customer-oriented arguments into consideration. These are subjective impressions of non-technical users who assess such a speaker recognition system under a completely different view. Therefore, we performed two field tests within the framework of text-independent speaker recognition (verification) in order to focus on the relation between objective recognition capabilities and subjective assessments from the field test participants. The two different scenarios under investigation are a telephone-based application and a test under poor conditions in terms of S/N-ratio. In this paper, we report about our investigations concerning the main point of criticism: the duration of the test phase (10 seconds). Based on a heuristic approach, we developed an enhanced version of the recognition process during the test phase, which is characterized by Dynamic Test Durations (DTD). It offers a remarkable trade-off between time (test duration) and system performance (error rate). The basic assumptions of this approach are:

- The computational processing time of the algorithm is significantly shorter than the test duration.
- The underlying recognition system performs a speaker verification: the speaker claims his identity in form of a PIN. This statement is then checked with his voice signature.

Our approach using DTD works in principle independently from the underlying speaker identification algorithm. Due to the fact that we have no practical proof of this assumption, in sections 2, we present briefly the technology used in the field test. Then, in section 3, we review the framework, the field test was embedded. Section 4 describes our new approach for reducing the test durations including some examples and our results.

2. SPEAKER RECOGNITION TECHNOLOGY

The speaker recognition technology is provided by ENST, a research unit, which has been working for several years in speech processing, and in particular, speaker recognition. The 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 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.

Though other approaches 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 TESTS

3.1 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 1, 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 aircraft.

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.

3.2 Field test task

In the system's test phase, 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. Note that, under this configuration, the False Acceptance Rate (FAR) is strongly related to the number of participants ($f_a \sim (n-1)/N$). 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. This configuration focuses exceptionally good on the evaluation of the FAR.

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.

3.3 Performance measurement

A conventional way of measuring performance in the field of speaker verification is to estimate the Equal Error Rate (EER). This performance measurement corresponds to the system operating conditions where the False Acceptance Rate (FAR) is equal to the False Rejection Rate (FRR). 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 FAR was calculated for different values of the threshold n, mentioned in section 3.2. For each of these values, the corresponding 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. But in the framework of our investigations, we have of course comparable results whose principle statements can be transferred to other performance measurements.

3.4 Field test results

In [3], you will find an overview about all of our qualitative and quantitative results. In this paper we will concentrate on the duration-dependent (test and enrollment) outcomes of the field trials. They are summarized in the following items which are mainly based either on general observations drawn from the results or on subjective statements from the participants.

- People do accept relative long enrollments (30 seconds), when they are sufficiently informed about its use and when they are prepared for this task.
- 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 (scenario A) or 15 seconds (scenario B). This is understandable, because it does not correspond to our human experience. In conventional telecommunication systems, human beings recognize the identity of a known speaker in a dynamic way. The duration for this natural recognition process is influenced by background noise and individual, probably time-dependent characteristics of the speaker.

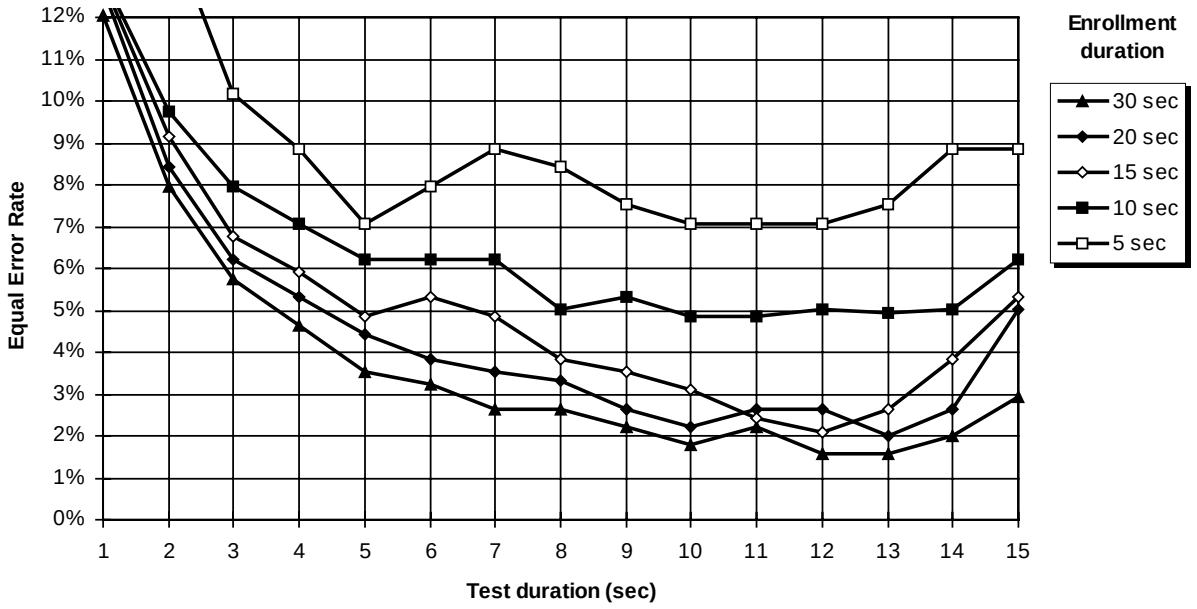


Figure 1: Equal Error Rates for scenario B (170 speakers) with variable duration for enrollment and test.

Figure 1 reflects the dependencies between different durations for enrollment and test for the static case. It has to be mentioned, that the small decrease in performance for test durations longer than 12 seconds can be explained by a specific behavior of many field test participants: they stopped speaking before they reached the end of the recording. The remaining seconds were filled with noise which significantly degraded the system's performance.

Nevertheless, form the participant's point of view, the test durations have the main impact on their subjective system assessment.

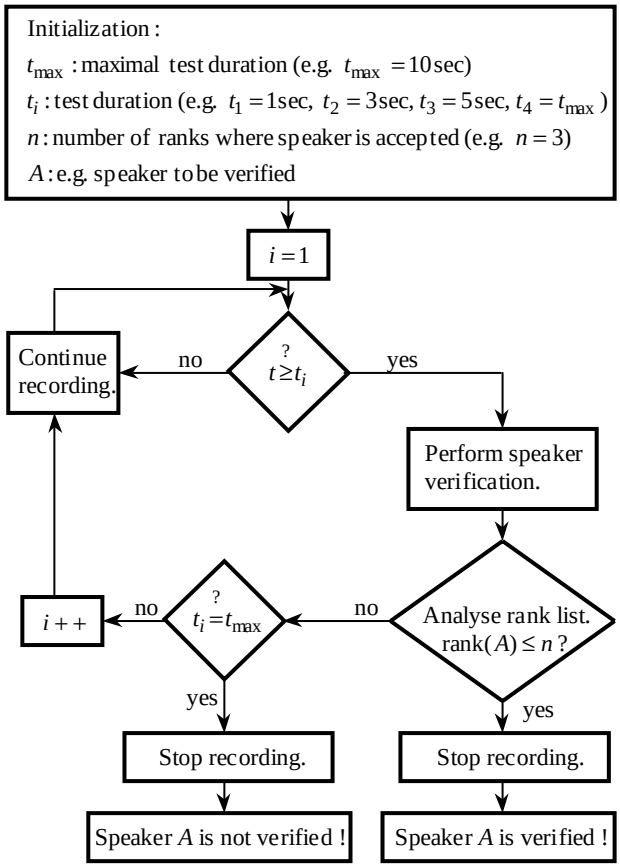


Figure 2: Flow diagram for Dynamic Test Duration

4. DYNAMIC TEST DURATIONS

The determination of appropriate test durations in conventional speaker verification systems is normally based on a single, fixed criterion. Either a duration close to the convergence-time t_{conv} of the underlying algorithm is chosen for achieving best system performance or the test duration is directly related to a predefined degree of quality (trade-off between EER versus test duration).

4.1 Enhanced recognition approach

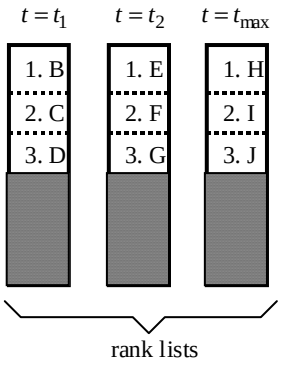
Our enhanced approach overcomes the drawback of a fixed, predefined test duration by performing a dynamic number of tests. This proceeding enables the system to accept a speaker after a very short duration t_1, if his voice signature identifies himself after the first test. If not, the tests are continued, taking longer test duration into account until either the speaker is accepted or the maximum test duration is reached, followed by a speaker rejection. Figure 2 illustrates the proceeding in more detail.

On the one hand, it is advantageous that the FRR is not influenced by this approach while on the other hand, the principle increase of the FAR is also a system-immanent feature. A more systematic study is surely necessary to estimate its relevance on the system. This study can either be carried out by some mathematics using severe assumptions to form an appropriated statistical model. We preferred the more direct, heuristic approach based on several offline tests.

Examples 1 – 3 demonstrate the influence of the DTDs on the FAR compared with conventional solutions. This procedure is performed for three scenarios with different underlying assumptions and statistical models:

Example 1:

- n = 3 ranks are accepted
- Speaker *A* is not recognized
- FAR (DTD) = 9/N (speakers B, C, …, J are impostors with PIN of A)
- FAR (conventional) = 3/N (only speakers H, I, J are impostors with PIN of A)



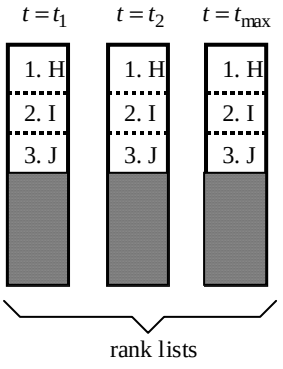
Example 1: Worst case for DTD-approach due to statistically independent recognition results.

Example 1 reflects the worst case for the DTD because no correlation between the three test results can be detected. Therefore, the degradation of the system's performance is proportional to the number of tests. But for obvious reasons example 1 does not model the DTD behavior correctly. This is because the verification task can be viewed more accurate as a classification process which assigns the current voice signature to a set of nearest neighbors in the feature space. According to figure 1, this procedure will converge over time towards the optimal solution. Taking this reflection into account, example 2 becomes more appropriate, where a strong correlation of the different tests is assumed.

While examples 1 and 2 are dealing with the case that the correct speaker can not be verified, example 3 formulates the most positive side of DTD. The correct verification is carried out after the first test, while the conventional solution requires the maximal time for the same result.

Example 2:

- n = 3 ranks are accepted
- Speaker *A* is not recognized
- FAR (DTD) = 3/N (speakers H, I, J are impostors with PIN of A)
- FAR (conventional) = 3/N (speakers H, I, J are impostors with PIN of A)



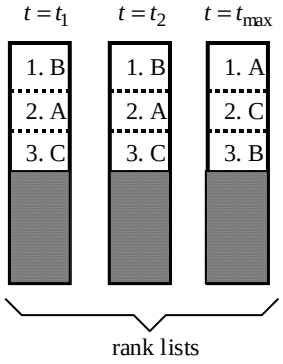
Example 2: No performance degradation for DTD-approach due to statistically correlated recognition results.

4.2 Results

The result of our investigations is that the FAR does not increase linear with the number of tests. Nevertheless, the number of tests should be limited with respect to a reasonable trade-off between FAR and average duration till verification.

Example 3:

- Speaker *A* is recognized at $t = t_1$ for DTD or $t = t_{max}$ for the conventional approach.
- FAR (DTD) = 2/N (speaker B, C are impostors with PIN of A)
- FAR (conventional) = 2/N (speaker B, C are impostors with PIN of A)



Example 3: No performance degradation for DTD-approach due to statistically correlated recognition results.

Our system gave best performance with four tests after 1, 2, 5 and 10 seconds. Compared to the non-DAD system, the FAR increases from 1,3 % to 1,7 % but the average testing duration goes down from 10 seconds to only 2,4 seconds. For some applications the massive reduction of the medium test duration is a key argument which opens new business fields.

5. CONCLUSIONS

This paper reports on a study, which incorporates some customer-oriented considerations into the assessment of speaker recognition systems. The analysis of own field tests showed that the test durations are a major point of criticism. Therefore, a new approach for speaker recognition has been introduced, based on Dynamic Test Durations. The classification is performed by exploiting a rank-based representation of the recognition results, which supports a dynamic way of speaker verification with several tests instead of using a fixed, predefined test duration.

The results are quite remarkable as they are described by a reduction of the mean test duration from 10 down to 2,3 seconds while the false recognition rate increases slightly from 1,3 % to 1,7 %. Nevertheless, these results have to be verified on other databases and speaker recognition algorithms.

Thanks to this study, we were able to improve the system's performance towards closing the gap between technology- and business-driven goals in the field of speaker recognition.

6. 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.*
3. Glaeser, A., Bimbot, F. "Steps toward the integration of speaker recognition in real-world telecom applications", *Proceedings ICSLP, Sydney, 1998.*