ELSEVIER

Speech Communication 31 (2000) 141–154

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Speaker verification over the telephone [☆]

L.F. Lamel [*], J.L. Gauvain

Spoken Language Processing Group, LIMSI–CNRS, B.P. 133, 91403 Orsay, France

Received 20 September 1998; received in revised form 20 September 1999

Abstract

Speaker verification has been the subject of active research for many years, yet despite these efforts and promising results on laboratory data, speaker verification performance over the telephone remains below that required for many applications. This experimental study aimed to quantify speaker recognition performance out of the context of any specific application, as a function of factors more-or-less acknowledged to affect the accuracy. Some of the issues addressed are: the speaker model (Gaussian mixture models are compared with phone-based models), the influence of the amount and content of training and test data on performance; performance degradation due to model aging and how can this be counteracted by using adaptation techniques; achievable performance levels using text-dependent and text-independent recognition modes. These and other factors were addressed using a large corpus of read and spontaneous speech (over 250 hours collected from 100 target speakers and 1000 imposters) in French, designed and recorded for the purpose of this study. On these data, the lowest equal error rate is 1% for the text-dependent mode when two trials are allowed per verification attempt and with a minimum of 1.5 s of speech per trial. © 2000 Elsevier Science B.V. All rights reserved.

Résumé

L'authentification automatique du locuteur a été le sujet d'actives recherches durant de nombreuses années, et malgrés ces efforts et des résultats prometteurs en laboratoire, le niveau de performance sur le réseau téléphonique reste inférieur au niveau requis pour de nombreuses applications. L'étude expérimentale, dont les principaux résultats sont présentés dans cette article, avait pour objectif de quantifier en dehors de toute application l'influence de facteurs plus ou moins reconnus pour leur effet sur les performances des systèmes d'authentification du locuteur. Les questions addressées sont: le choix du modèle (mélange de gaussiennes ou modèle phonétique); la connaissance ou non du texte prononcé par le locuteur; l'importance de la quantité et de la nature des données d'apprentissage et d'authentification, en particulier l'influence du contenu linguistique des énoncés sur le niveau de performance pour des textes lus et de la parole spontanée; la dégradation des résultats due au vieillisssement des modèles et la manière de le compenser avec des techniques d'adaptation. Les résultats expérimentaux ont été obtenus sur un corpus téléphonique conçu et enregistré pour cette étude qui comprend plus de 250 heures de parole pour un total de 100 locuteurs abonnés et 1000 imposteurs. Sur ces données le taux d'égale erreur est de l'ordre de 1% dans le mode dépendant du texte lorsque deux essais sont autorisés par tentative d'authentification avec une durée minimale de 1,5 s par essai. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Text-dependent and text-independent speaker verification and speaker identification; Hidden Markov models

[☆] This work was carried out in collaboration with the Vecsys company in the context of a research contract with France Telecom.

[*] Corresponding author. Tel.: +33-1-69858063; fax: +33-1-69858088.

E-mail addresses: lamel@limsi.fr (L.F. Lamel), gauvain@limsi.fr (J.L. Gauvain).

0167-6393/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(99)00075-8

1. Introduction

Speaker verification has been the subject of active research for many years, and has many potential applications where propriety of information is a concern (Atal, 1976; Rosenberg, 1976; Doddington, 1985; Naik, 1990; Rosenberg and Soong, 1992; Furui, 1994; Gish and Schmidt, 1994; Boves and den Os, 1998). Despite these efforts and promising results using laboratory data, speaker verification performance over the telephone remains below that required for many applications (Boves and den Os, 1998).

The speaker recognition problem is closely related to the speech recognition one. Both problems share the same basic speech generation model $f(\boldsymbol{x}|\lambda, w)$. The speech signal $\boldsymbol{x}$ conveys both linguistic information (w, the word sequence) and non-linguistic information (λ, the speaker identity). Obtaining the former is the goal of speech recognition, whereas the non-linguistic information is more relevant to the speaker recognition problem. Typical approaches attempt to extract one of the types of information, treating the other as a source of noise. Performance is acknowledged to be dependent upon the linguistic content of the speech data. For speaker recognition there are varying degrees of control ranging from fixed prompt texts, text-dependent (variable prompts or user selected texts), to free text or text-independent. In fact, even when there are no restrictions on the text, users tend to say the same or a similar text, which implies that a text-dependent system may be appropriate for many applications.

The text mode has direct implications on model estimation. The more control there is on the speech input, the less there is a need for acoustic training data. In general, since application designers want to limit the time required for user enrollment, it is essential to use a discriminative approach to reduce the need for training data. Judicious design of prompts can help in optimizing performance for a given amount of data (for example, it is generally considered that voiced speech contains more information about the speaker than does unvoiced speech).

There have been a wide spectrum of proposed approaches to speaker verification starting with very simplistic models such as those based on long term statistics (Furui et al., 1972). The most sophisticated methods rely on large vocabulary speech recognition with phone-based HMMs (Newman et al., 1996). Intermediary approaches make use of phone or phone-class based models (Rosenberg et al., 1990; Lamel and Gauvain, 1992; Matsui and Furui, 1993; Gauvain et al., 1995; Lamel and Gauvain, 1995; Carey et al., 1996). As for the training data, systems making use of linguistic information (prior knowledge about the text or the output of a high quality speech recognizer) typically require less data for authentication than is required by systems not making use of this information source (where the linguistic information is seen as a kind of noise). So in general, text-independent systems require longer speech segments in order to properly identify the speaker. Modeling the linguistic content is certainly more accurate, but requires substantially more development work and data for training the models. This type of approach also is inherently language-dependent and assumes that the language spoken is known in advance.

The best compromise between accuracy and complexity is likely to be dependent upon the particular application. For example, the widely used cepstral-based GMM models have been quite successful for speaker identification of conversational speech. This task, introduced by NIST in 1996, makes use of the Switchboard corpus (Przybocki and Martin, 1998). On these data the phone and word-based modeling approaches have not out-performed systems based on GMMs (Reynolds, 1995, Carey et al., 1996; Newman et al., 1996; Lamel and Gauvain, 1997). The NIST framework is very attractive, particularly in enabling participants to compare technologies on a common task and corpus. However, the corpus type, which is evidently quite interesting for defense and criminal applications, may not be representative of many potential speaker verification applications. It is quite likely that many telecom applications (Boves and den Os, 1998) will involve a human interacting with a machine and not with another human.

The objective of this research is to assess the performance of state-of-the-art methods for speaker verification to determine if high enough performance levels could be obtained to support the development of telecom applications. This experimental study aimed to quantify more-or-less well-known trends in speaker recognition out of the context of any specific application. Some questions addressed are: how does the amount and content of training and test data affect performance; how much degradation of performance can be anticipated due to model aging and how can this be counteracted by using adaptation techniques; what performance levels are achievable using text-dependent and text-independent recognition modes. These and other factors were addressed using a large corpus of read and spontaneous speech in French designed and recorded for the purpose of this study.

At the time this work was started (Gauvain et al., 1995), and even today, there are no publicly available corpora for speaker verification of the size and content used in this work. (See (Campbell and Reynolds, 1999) for a compilation of available corpora.) The most widely used corpora for speaker verification are the TIMIT corpus (and derivatives), Yoho, Polycost and the portions of the Switchboard Corpus (Godfrey et al., 1992) used in the NIST evaluations. The TIMIT corpus, while offering data from a relatively large number of speakers was not designed for speaker verification and has the default that all the data for a speaker were recorded in a single-session. The Yoho corpus was recorded with a high quality microphone and is much smaller in terms of numbers of speakers. Polycost, which is closest in style to the corpus used here, contains telephone data in non-native English and European languages, but does not have imposter data. The Switchboard corpus contains only conversational speech, whereas many applications are more likely to use prompted speech or spontaneous responses as in a human–machine dialog context.

Our goal was not to determine a particular setup with the best performance, but to investigate key parameters that affect performance in the context of various telecom applications.

In the next section, the corpus and methodology used in this work are presented. Sections 3–5 provide experimental results for different training configurations, data content and speaking style. Section 6 provides observations based on these experiments and some conclusions concerning the use of this approach for telephone applications.

2. Corpus and methodology

For these experiments, we make use of a corpus especially designed to evaluate speaker recognition algorithms. [1] This corpus contains over 250 hours of speech data from 100 target speakers (or users), and from 1000 imposters (Gauvain et al., 1995). Each user completed 10 training calls, and 25 verification calls, from a variety of telephone handsets and calling locations over a period of 2 years. Each imposter completed a single verification-type call. The training calls took about 25 min to complete, producing about 12 min of speech data. The verification calls each resulted in about 2.5 min of speech data. The recordings are similar to the Polyphone recordings being collected in several languages (Bernstein et al., 1994; Godfrey, 1994). Each call provides a variety of speech data, including read speech material, and elicited and spontaneous speech so as to be able to assess the effects of data type on the verification accuracy. The read speech data consist of three types: digit strings, five phonetically controlled sentences (SEPT), [2] and sentences from the *Le Monde* newspaper selected to cover a large number of phonetic contexts. The spontaneous speech data contain responses to fixed questions (such as the type of handset, calling environment, calling area code, dates, times, etc.) and to more general open questions designed to obtain short monologues.

A statistical modeling approach is taken, where the talker is viewed as a source of phones, modeled by a fully connected Markov chain (Gauvain and

[1] The corpus, conceived and designed jointly by CNET and LIMSI, was recorded over the French telephone network and transcribed by the Vecsys company.

[2] The SEPT sentences were specified by the Service d'Etudes commun de la poste et télécommuncations. They are short easy to pronounce sentences containing almost only voiced phonemes.

Lamel, 1993; Lamel and Gauvain, 1993, 1995) for text-independent verification. [3] The lexical and syntactic structures of the language are approximated by local phonotactic constraints, and each phone is in turn modeled by a three state left-to-right HMM. For text-dependent identification, a left-to-right HMM is built by concatenating phone models according to the lexical pronunciations of words in an orthographic transcription.

When this approach is applied to speaker identification (Gauvain and Lamel, 1993; Lamel and Gauvain, 1993, 1995) a set of phone models is trained for each speaker and identification of a speaker from the signal $\boldsymbol{x}$ is performed by computing the phone-based likelihood $f(\boldsymbol{x}|\lambda)$ for each speaker λ. The speaker identity corresponding to the model with the highest likelihood is then hypothesized. The same speaker model can be applied to speaker verification by comparing the likelihood ratio $f(\boldsymbol{x}|\lambda)/f(\boldsymbol{x})$ to a single speaker-independent threshold in order to decide acceptance or rejection.

Speaker-specific models are generated from a set of speaker-independent (SI) seed models using maximum a posteriori (MAP) estimation. The speaker-independent seed models provide estimates of the parameters of the prior densities and also serve as an initial estimate for the segmental MAP algorithm (Gauvain and Lee, 1994).

Assuming no prior knowledge about the speaker distribution, the a posteriori probability $\Pr(\lambda|\boldsymbol{x})$ is approximated by the score $L(\boldsymbol{x};\lambda)$ defined as

$$L(\boldsymbol{x};\lambda) = f(\boldsymbol{x}|\lambda)^{\gamma} \Big/ \sum_{\lambda'} f(\boldsymbol{x}|\lambda')^{\gamma},$$

where the λ' are the speaker-specific models for all speakers known to the system and the normalization coefficient γ was empirically determined as 0.02. (This coefficient is needed to compensate for independency approximations in the model.) Calculating the denominator of this expression can be very costly as the number of operations is proportional to the number of speakers used in the calculation, or as in our case, the number of target speakers. We can significantly reduce the required computation by using a Viterbi beam search on all the speakers' models in parallel. This decoder, which was developed for speaker identification and the identification of other non-linguistic speech features (Gauvain and Lamel, 1993; Lamel and Gauvain, 1995), provides not only the likelihood of the most probable speaker, $f(\boldsymbol{x}|\lambda)$, but the likelihoods for the N most probable speakers. The necessary computation is reduced by approximating the above summation by a summation over a short list of the most probable speakers. In our implementation, the Viterbi algorithm is used to compute the joint likelihood $f(\boldsymbol{x},\mathrm{s}|\lambda)$ of the incoming signal and the most likely state sequence instead of $f(\boldsymbol{x}|\lambda)$.

If a verification attempt is unsuccessful, it is common practice to allow a second trial in order to reduce the false rejection of known users. A straightforward approach is to base the decision only on the score of the second attempt, ignoring the preceding trial. This approach can be justified on the ground that the actual test data are potentially invalid. An alternative it is to base the decision on the scores of both trials. [4] Making use of this second approach reduced the speaker identification error rate by 21%, compared to a 13% error reduction using only the score of the last attempt.

Fig. 1 shows the distribution of scores for 2221 trials each for target speakers and imposters (truncated at 300 attempts). 87% of the attempts by imposters have a score of essentially 0, and 51% of the attempts by target speakers have a score of essentially 1. However, there is a substantial overlap in the distributions, and it is apparent from these histograms that the main source of error comes from a low score for certain target speaker attempts. (Almost 2% of the attempts by target speakers have a score almost equal to 0.)

[3] This phone-based approach is also compared with Gaussian mixture models in Section 3.

[4] It is evidently possible to allow more than two trials per attempt, in which case the score would take into account scores from all previous trials.

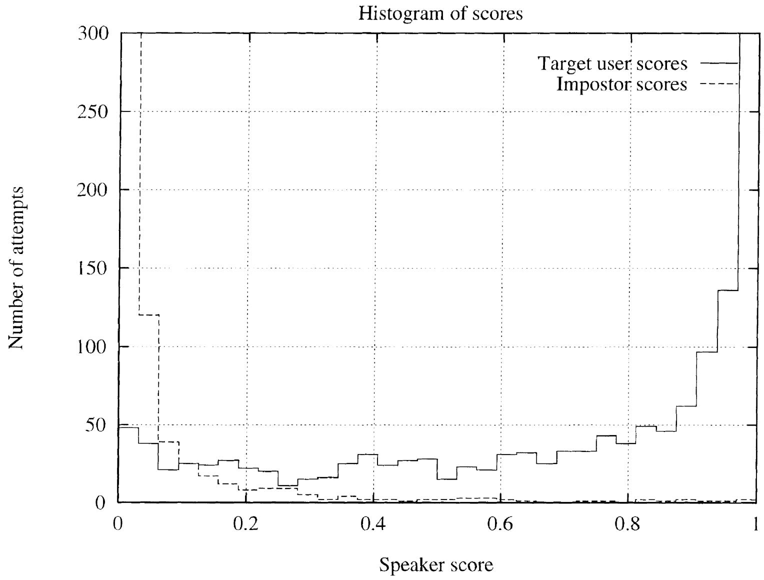

Fig. 1. Distribution of scores for target speakers and imposters (2221 attempts each), y-axis truncated at 300 attempts.

From the histograms it can also be seen that setting a threshold in the range of the equal error rate (EER) is not too problematic. Almost any value between 0.1 and 0.4 will keep the average error rate near this point. It should be recalled that the EER is obtained by selecting the decision threshold a posteriori such that the two types of errors are equal. In practice, however, the decision threshold should be fixed a priori based on a development corpus so as to minimize the cost function specific to the application. For example, for very secure applications a much higher cost will be associated to false acceptances than to false rejections. Without an appropriate cost function, it is common practice to select the decision threshold so as to minimize the EER. We compared average error rates using two decision thresholds: one a priori (determined using development data) and the other determined a posteriori. On these data although the average error rate varies only slightly, the rates of the two error types vary more. With a single authentication attempt, the EER with the a posteriori threshold is 2.61, compared to an average error of 2.64 with the a priori decision threshold. The corresponding false rejection and false acceptation rates are 2.97 and 2.30 respectively (instead of 2.61 for the a posteriori setting of the threshold).

3. Contrastive experiments

A series of baseline experiments were carried out to quantify speaker recognition performance as a function of parameters generally acknowledged to affect performance. This section summarizes the experiments and presents performance results on the corpus described above. Results are reported for speaker identification since this is easy to measure, and are strongly correlated with speaker verification error rates. However, since we are interested in assessing speaker verification performance, the equal error rates are computed for the configurations of greatest interest (those where the speaker identification error is not too high).

In all experiments reported in this paper, the acoustic feature vector containing 13 cepstrum coefficients derived from a Mel-frequency spectrum (0–3.5 kHz bandwidth) and their first-order derivatives was computed every 10 ms. In order to minimize effects due to channel differences, cepstral-mean removal was performed for each sentence.

3.1. Gaussian mixture versus phone-based models

Experiments were carried out to compare speaker verification performance using phone-based models with a baseline system using Gaussian mixtures. Two mixtures of 32 Gaussians are used, one for silence/noise (common for all speakers) and another for the speech, specific to each speaker. For the phone-based approach, text-dependent and text-independent modes are compared, for one and two verification trials. When two verification trials are authorized (for target speakers and imposters), there are on average 1.1 trials per user attempt.

Fig. 2 gives some baseline receiver operating characteristics (ROC) curves for different model types and operational modes for a subset of the telephone data. The ROC curve for the Gaussian mixture model is shown in (a). This can be compared with (b) the ROC of the phone-based approach in text-independent mode. The phone-based approach is seen to perform significantly better than the Gaussian mixture model (7.3% versus 9.0% EER) with only one trial per attempt and an average of 3.2 s of speech per trial. If the text is known, the EER is reduced to 5.1% (curve c). It should be noted that with the phone-based approach, knowing the text does not imply the use of a fixed text. The user can be prompted to read any text. In (d), two verification trials are allowed per attempt, reducing the EER to 4.4% with 1.1 user trials on average. Curve (e) shows the ROC if a minimum amount of 2 s of speech is required for each trial. For the sentences having this minimal duration, the EER is reduced to 3.5%.

For the remainder of the experiments reported in this paper, the phone-based approach is used.

3.2. Authentication utterance duration

Speaker recognition performance is known to be dependent upon the duration of the test utterance. This was illustrated in Fig. 2 where a significant performance improvement was obtained by ensuring a minimal test signal duration. One problem in designing systems is to ensure that the talker will supply the needed amount of speech data. This is particularly true for digit strings which tend to be quite short. In this section, speaker verification performance is assessed as a function of the duration of the test utterances. There is a strong correlation between the test utterance type and the test utterance duration. Only 10% of the digit strings and SEPT utterances are

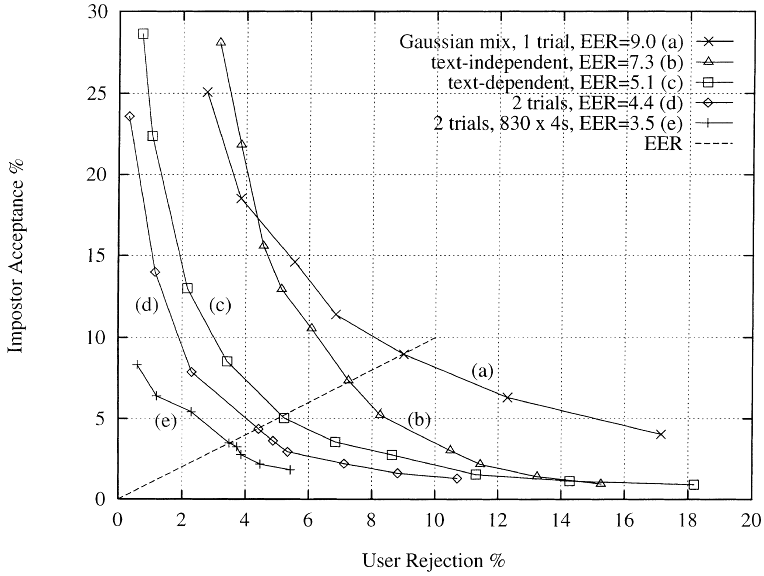

Fig. 2. ROC curves for different model sets and operational modes based on 21 775 user attempts and 10908 × 91 imposter attempts: (a) multi-Gaussian model; (b) 35 phone models, text-independent; (c) 35 phone models, text-dependent; (d) same as (c) with two trials; (e) same as (d) with exactly 4 s of speech. The dotted line shows the points of equal error (false acceptance/false rejection).

Table 1
Equal error rates (EER) for different test data types based on 21 775 user attempts and 10908×91 imposter attempts (the text is known)

Conditions	Average	Digits	SEPT	Sentences
1 Trial,	3.3	4.2	2.3	2.6
2 Trials,	2.7	3.1	1.7	2.0
2 Trials, $\geqslant 1.5$ s	1.8	1.4	1.0	1.9
1 Trial, 1.2 s	–	3.6	2.4	4.9
2 Trials, 1.2 s	–	2.8	1.9	3.2

longer than 2 s in duration, whereas 95% of the sentences are at least this long. Imposing a minimal duration of 1.2 s eliminates almost 25% of the digit strings and 10% of the SEPT utterances.

Table 1 gives the known-text equal error rates for the different types of test data. Results are given for 1 and 2 user attempts, with and without a minimal duration constraint. With one trial per attempt, the average EER is 3.3%. This is reduced to 1.8% if a miminal duration of 1.5 s is required and two trials per attempt are allowed. Allowing multiple attempts and requiring a minimum amount of authentication data can significantly reduce the EER.

In order to eliminate the dependence on test utterance duration, the last two table entries show the EERs using a fixed test duration of 1.2 s, for one and two trials. With the restricted duration, better performance is obtained for the SEPT sentences and digit strings than for the sentences.

3.3. Amount and recency of training data

The amount and recency of the training data are well-known factors that influence speaker verification performance. It is also known that better performance can be obtained with training data recorded in multiple sessions reflecting conditions of real use. Obtaining the necessary data can require a long enrollment procedure which is usually undesirable from the users' viewpoint. A related known problem is that of model aging: typically as the time between training and test increases, performance gradually degrades unless adaptation is used to keep the models up-to-date.

The aim of these experiments was to quantify the effects of limiting the training data on speaker recognition performance, and for a fixed number of training utterances, different means of obtaining it (single versus multiple session training). The performance is measured as a function of the quantity of data used to train the models. Three-session training is compared with single-session training (the last session of the 3), and with 1/3 of the training data taken from each of the three training sessions. [5] These training sessions are the last training sessions recorded for each speaker, so the first of the 3 was made 6 calls before the first authentication call. The latter comparison enables us to investigate the effects of single-session and multi-session training for a fixed amount of data, which can influence the choice of enrollment procedure.

Table 2 gives the speaker identification error rates (left) and equal error rates (right) as a function of the amount of training data, and the proximity to the test data. [6] Three session training results in the lowest error rates, which is expected as the acoustic models are trained on the most data. If the training data are to be reduced to one-third, the best performance is obtained by keeping 3 training sessions, but reducing the amount of data in each session. Single session training results in identification error rates substantially higher than multi-session training, even when this session is temporally closer to the test data. The EER is seen to significantly increase when the training data are reduced to one session, with an over 60% increase for the digits and SEPT sentences. It can

[5] Due to the training call length, alternate training sessions contained complementary data types. The odd sessions consisted of 25 digit strings and 25 journal sentences, whereas the even sessions consisted of the 25 SEPT sentences and spontaneous responses to 25 questions.

[6] These experiments were carried out before the corpus was completed, and therefore have a fewer number of user and imposter trials than reported in the other tables.

Table 2
Speaker identification error rates (left) and equal error rates (right) for different type-specific training conditions (known text, 1 trial per attempt, based on 1375 user attempts and 675 × 91 imposter attempts)

	Identification error rate			Equal error rate		
Training	Digits	SEPT	Sentences	Digits	SEPT	Sentences
3 Sessions	4.8	2	6.7	2.9	1.9	3.2
1/3 of 3 sessions	6.4	4.1	6.6	3.8	2.3	3.2
1 Session (last)	10.8	6.3	8.3	4.8	3.1	3.5

be noted that the performance on the *Le Monde* sentences is relatively insensitive to the training configuration. This may be related to the total amount of training data: since the *Le Monde* sentences have on average a longer duration (4.5 s each) than the digit strings and SEPT sentences (1.6 s each), even in the reduced training condition enough data are available with which to estimate the model parameters. These results support the need for multiple training sessions.

3.4. Model aging and adaptation

The well-known effects of model aging can be illustrated by the different performances observed for single-session training. If speaker-specific models are trained only on the first of the three-sessions instead of on the last one (given in the last line of Table 2), the identification error rates almost double: 19.7%, 11.5% and 16.7%, respectively for the digits, SEPT and sentences, compared with 10.8%, 6.3% and 8.3%. (On average several weeks passed between the first and last of the three training sessions.) Speaker-adaptation techniques can be used to reduce the effects of model aging. We experimented with MLLR-based adaptation (Legetter and Woodland, 1994) using data from all but the last two test sessions per speaker. Without adaptation, the EER obtained on the last two test sessions is 2.5%. This error rate is significantly higher than the EER of 1.6% obtained on the first two test sessions (the first two calls subsequent to the training sessions). After adapting the speaker models on data from the intervening session, the EER on the last two sessions is reduced to 1.7%. This indicates that adaptation is crucial to maintaining system performance over time.

3.5. Discussion of contrastive results

From these comparative results we can make the following conclusions:

- On this telephone corpus the phone-based approach outperforms the simpler approach based on a mixture of Gaussians. (This is different from what we observed in the 1996 NIST evaluation using a conversational speech corpus (Lamel and Gauvain, 1997.))
- As expected, a significant gain is observed when the text is known a priori. [7] So for telecom applications where a cooperative user is expected, this difference is big enough by itself to justify the use of known text recognition which is both more performant and less complex.
- Allowing a second verification trial reduces the EER without significantly increasing the number of trials for the target speakers. A second trial is needed in only 1 in 10 user attempts. (Evidently all imposter attempts have two trials, but this is not a concern.)
- Requiring 4 s of speech signal duration reduces the error rate substantially (about 20% comparing curves d and e of Fig. 2). Therefore, the verification procedure should ensure that a minimum of 4 s of speech is collected in each authentication attempt.
- It is preferable to acquire the training data in several sessions, than in a single session. Multiple session training is less sensitive to channel conditions and intra-speaker variability. The relative reduction in EER is between 10% and 25%.

[7] Although this condition made use of an orthographic transcription of the speech, in a contrastive experiment using the prompt text no difference in performance was observed.

- As the time between the training calls and the authentication call increases, performance tends to decrease. This model aging can be successfully counteracted with unsupervised adaptation so as to maintain performance over time.

In the following sections, the experimental setup is restricted to the phone-based approach. Most of the experiments are for the known-text condition, with the exception of the spontaneous speech where results are provided for both the known and unknown text conditions.

4. Choosing the prompt linguistic content

One important factor to be addressed is the influence of the linguistic content of the training and test material on speaker identification and verification performance. To investigate this factor, experiments were carried out using different subsets of the corpus for training and different types of test material.

The left side of Table 3 shows text-dependent speaker identification error rates as a function of the utterance type and the training condition (multi-style or type-specific). Multi-style training makes use of all types of read-speech training data for the 10 training calls. Type-specific training makes use of only one of these data types in training, i.e. digits, SEPT sentences or *Le Monde* sentences. For the training data, the average duration of the digit strings and the SEPT sentences are 1.6 s, and the average duration of the *Le Monde* sentences is 4.5 s. The type-specific models are trained with only one-third of the utterances used to train the multi-style models. Using multi-style training, the speaker identification error rate ranges from 9.5% for the digit strings to 4.5% for the SEPT sentences.

When type-specific training is used, and testing is carried out on the same type of data (the diagonal entries in the lower part of Table 3), the speaker identification error rates are seen to be slightly lower than with multistyle training for the digits and the SEPT sentences, even though the acoustic models have been trained with significantly less data. The lowest identification error rate (3.6%) is still obtained with the SEPT sentences.

Exactly the same pattern of performance is observed for speaker verification in terms of the equal error rate for both multistyle and type-specific conditions, with the lowest EER of 2.3% for the SEPT sentences (cf. right part of Table 3).

To assess how important it is to have matched conditions in the linguistic content of the training and test data, speaker identification performances was also measured under crossed-type training and testing conditions (the off-diagonal entries in Table 3). Such mismatch results in a dramatic performance degradation, with the best training material under mismatched conditions being the *Le Monde* sentences. This result was to be expected as the sentences have the largest variety of phonetic contexts.

Several conclusions can be drawn from this experiment. Some types of linguistic content are seen to clearly result in better speaker recognition performance than others, showing the importance of this aspect in system design. Comparing the three types of data, it is not evident to identify a

Table 3
Speaker identification error rates (left) and equal error rates (right) as a function of test data type with multi-style training and with type-specific training based on 21 775 user attempts and 10908 × 91 imposter attempts (one trial per authorization attempt with no minimal duration constraint, the text is known)

	Test data					
	Identification error rate			Equal error rate		
Training data	Digits	SEPT	Sentences	Digits	SEPT	Sentences
Multistyle	9.5	4.5	5.5	4.2	2.3	2.6
Digits	**8.6**	68.6	35.6	4.1	–	–
SEPT	64.1	**3.6**	24.3	–	2.3	–
Sentences	21.1	14.3	**5.9**	–	–	2.7

single differentiating factor that can explain the observed performance differences. Some characteristics of the prompt texts that can affect performance are the linguistic content (in terms of lexical coverage and phonological characteristics), the easiness to pronounce, the familiarity of the words and the utterance duration.

Limited phonetic contexts are desired as better acoustic models can be estimated with limited amounts of training data. Both the SEPT sentences and digit strings have limited linguistic contents, but quite different phonemic contents. They are both easy to pronounce, but the familiarity of users with digit strings can result in sloppy articulation (reduced pronunciations and short durations). The SEPT sentences are comprised of almost only voiced sounds which are widely acknowledged to contain more information about the vocal tract of the talker than unvoiced sounds. Another factor is utterance duration. The test digit strings range from 3 to 5 digits and some of the digit strings can be very short. As can be seen Table 1, even if a minimal duration is required, the SEPT sentences outperform the digit strings.

We can thus conclude that for text-dependent speaker recognition the choice of text used for training and for test, has a major impact on the performance. It is important that the training and test texts are of the same style. Simple, easy to pronounce texts, containing predominantly voiced sounds will result in the best performance, particularly when the training data are limited. Digits strings are often used in applications because they have no special meaning and can correspond to a speaker code, but are not optimal in terms of linguistic content. The SEPT sentences are more phonetically balanced and contain predominantly voiced sounds. They are also easy to pronounce and remember, however they may be awkward to use in an application, as they serve only for speaker verification and do not correspond to any natural data input.

5. Using spontaneous speech

There are a variety of applications where only spontaneous speech is available for speaker recognition. Applications in the domain of criminology often come to mind, but other applications concern the transcription of radio and television broadcasts or of meetings and conferences. In this case, automated methods may be used to partition the data into speaker turns and to identify the speakers. The identification can be used to enhance the transcription as well as to decode the speech signal with speaker-specific acoustic models. Other applications can be envisioned such as transparent, continual speaker recognition during a conversation with a human or a machine. In this case, the aim is to avoid fraudulent access via prerecorded speech or to detect any change of speaker during the transaction.

The responses to the fixed and open questions were used for the experiments with spontaneous speech. The fixed questions correspond to the type of data that could be used in a spoken dialog system where there is a desire to restrict access or ensure the identity of the caller supplying the information. These questions are of the type: “Say and spell the name of the town you are calling from.”, “What is the zip code in the town you are calling from?”, “What time is it?”. The average utterance durations in response to the fixed and open questions are shown in Table 4. The open questions were designed to incite the caller to say a short monologue. Some example questions are: “Describe the last movie you saw.”, “Describe your last vacation.”, “What is your favorite meal?”, “What advantages are offered by public transportation?”. The responses to the fixed questions were much shorter (1.5 s on average) than to the open questions (8.2 s on average). The

Table 4
Average durations for fixed and open questions

Questions	Average duration (s)
Calling place	1.3
Telephone type	1.7
Handset type	1.3
City/country	0.9
Postal code	1.3
Telephone no.	1.6
Date	2.3
Time	1.6
Open questions	8.2

response duration for the open questions was quite variable from a few seconds to very long monologes when the caller was interested in the question.

Fig. 3 shows the ROC curves for the spontaneous responses to the fixed and open questions. An automatic phone transcription was generated for each utterance using the speaker-independent seed models (35 context-independent phone models) and phone bigrams estimated on the BREF corpus (Lamel et al., 1991). This automatic transcription was then used in place of the true phone sequence in computing the likelihood for all the speakers' models. The same multi-style speaker models were used as in Section 4. The EERs for the fixed responses and open responses are 6.4% and 7.3%, respectively, with a maximum of two trials per attempt. These EERs are quite a bit higher than those that observed for the read speech data.

A logical question then is how much of this degradation is due to the differences in speaking style, and how much is due to the use of an imperfect phone transcription. There are two ways in which this question can be answered. The first is to carry out speaker verification for the read speech data without knowledge of the transcription. The second is to assess the verification performance on the spontaneous speech if an oracle were to provide the correct orthographic transcription (and the phone transcription via the lexicon).

The ROC curves for the digits, SEPT and *Le Monde* sentences in the unknown text condition using the same phone models are also shown in Fig. 3. The EER for the digits is 4.0%, whereas the EER for the SEPT sentences is 2.4% and the lowest EER of 1.7% is for the *Le Monde* sentences. This indicates that in the presence of phone errors, longer test utterance durations result in better verification performance. For the digits and the SEPT sentences the EERs are doubled compared to the known text condition (see Fig. 4). However, the performance on the read-speech data remains substantially better than on the spontaneous data.

In an attempt to further understand the large performance differences, the ROC curves for all five data types under the known-text condition are shown in Fig. 4. Even in this unrealistic condition for the spontaneous speech, the error rates for both error types are significantly higher than for

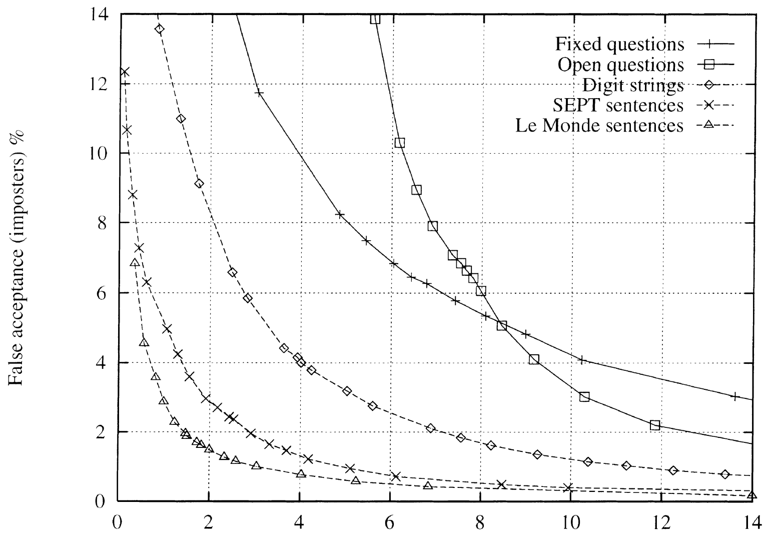

Fig. 3. ROC curves for spontaneous speech fixed responses r and open questions q without transcriptions (unknown text, phone recognition). Multi-style training. (Fixed questions: 8823 user attempts, 794 070 imposter attempts (simulated); open questions: 4691 user attempts, 422 190 imposter attempts (simulated).) Maximum of two trials allowed for each attempt with an average of 1.1 trials/attempt. ROC curves for the digits, SEPT and *Le Monde* sentences are given for comparison.

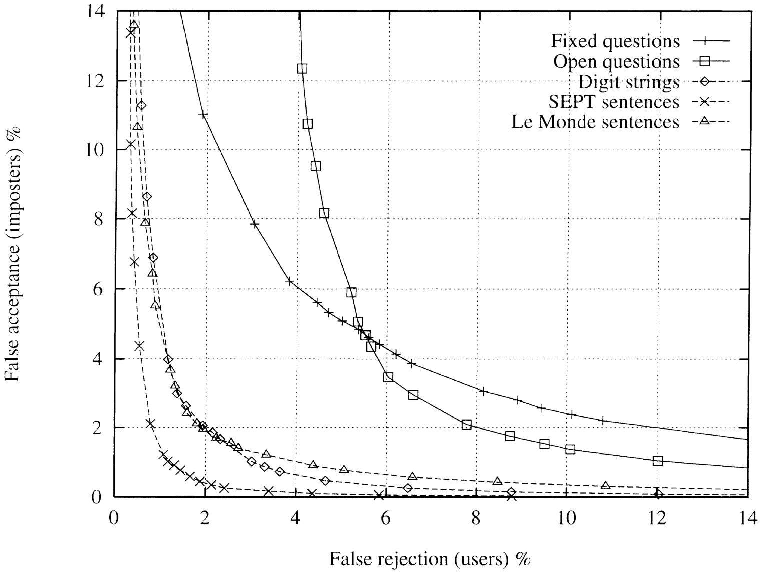

Fig. 4. ROC curves for spontaneous speech fixed responses r and open questions q using transcriptions (text known). Multi-style training. (Fixed questions: 8823 user attempts, 794 070 imposter attempts (simulated); open questions: 4691 user attempts, 422 190 imposter attempts (simulated).) Maximum of two trials allowed for each attempt with an average of 1.1 trials/attempt. ROC curves for the digits, SEPT and *Le Monde* sentences are given for comparison.

the read texts. We can therefore conclude that most of the performance difference is due to the nature of the spontaneous speech – more varied linguistic content, less fluent, less well articulated – and not to the errors in the phone transcriptions. Knowing the correct phone sequence only reduces the EER to 5.0%. Comparing the curves in Figs. 3 and 4, the degradation due to imperfect phone recognition can be estimated. The *Le Monde* sentences are the least affected by recognition errors, which is probably due to their longer average duration, and that the phone bigrams are well adapted to these data. The larger degradation observed for the SEPT sentences and digits strings can likely be attributed to a mismatch between their linguistic content and the phone bigram.

From the spontaneous speech ROC curves it appears that it is easier to reduce the false acceptances than to reduce false rejections. For the open questions in Fig. 3 the false rejection rate remains higher than 6%, whereas the false acceptances can be reduced to 2%.

These results confirm that speaker recognition using unconstrained spontaneous speech is significantly more difficult than with known prompts. This higher error rate can be partly attributed to a larger variation in speaking style and the larger variability in phonetic contexts.

6. Conclusion

With the recent advances in speech technologies, there has been increasing interest in developing interactive telephone-based services using voice. Some of these services could benefit from speaker verification technology in order to provide additional access security. The purpose of the experiments reported in this paper was to quantitatively assess performance as a function of system design choices, without constraints linked to a particular application. The main factors considered were the type of speaker model, the amount of test data, the amount and recency of the training data, the linguistic content, and speaking style.

Several observations can be made concerning these experiments. As expected, there is a correlation between the amount of training data and the system performance, with more data yielding higher performance. Similarly, for comparable amounts of training data, better performance is obtained when the data are taken from several training sessions, as opposed to all from a single call. Concerning the amount of data needed, estimation of speaker-specific models requires a minimum of about 1 min of speech. For the test utterances, the results indicate that it is advantageous to ensure a minimal duration of at least 1.5 or 2 s. An equal error rate of 1% was obtained on the SEPT sentences, in the text-dependent mode with two trials per verification attempt and with a minimum of 1.5 s of speech per trial.

It is evidently preferable that the linguistic content of the training data and test data are closely matched. If the test data are different in linguistic content (or uncontrolled), multi-style training is to be preferred, however type-specific training results in better performance when the same type of test data are used. The importance of phonetic content was illustrated for the crossed-type conditions, which led to significant degradation in performance.

Better performance is obtained for the SEPT sentences, with controlled linguistic content, than for digit strings or the more variable *Le Monde* sentences. This can be partially attributed to the smaller number of phonetic contexts, for which more accurate acoustic models can be estimated for a given amount of training data. Another contributing factor is that they are easy to remember and pronounce. As a result, speakers tend to say these naturally without hesitation. In contrast, reading aloud the *Le Monde* sentences sometimes caused difficulty for the callers. Verification performance using unrestricted spontaneous speech is significantly worse than for prompted speech. This can be partly attributed to a larger variation in speaking style and the larger variability in phonetic contexts.

References

Atal, B.S., 1976. Automatic recognition of speakers from their voices. Proc. IEEE 64 (4), 460–475.

Bernstein, J., Taussig, K., Godfrey, J., 1994. MACROPHONE: An American English Telephone Speech Corpus for the Polyphone Project. In: Proceedings IEEE ICASSP-94, Adelaide, Australia, Vol. 1, pp. 81–84.

Boves, L., den Os, E., 1998. Speaker recognition in telecom applications. In: Proceedings IEEE IVTTA-98, Torino, pp. 203–208.

Campbell, J.P., Reynolds, D.A., 1999. Corpora for the evaluation of speaker recognition systems. In: Proceedings IEEE ICASSP-99, Phoenix, AZ, pp. 829–832.

Carey, M.J., Parris, E.S., Bennett, S.J., 1996. Speaker Verification. In: Proceedings Institute of Acoustics (Speech & Hearing), Windermere, UK, pp. 99–106.

Doddington, G.R., 1985. Speaker recognition – Identifying people by their voices. Proc. IEEE 73 (11), 1651–1664.

Furui, S., 1994. An overview of speaker recognition technology. In: Proceedings ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification, Martigny, pp. 1–9.

Furui, S., Itakura, F., Saito, S., 1972. Talker recognition by longtime averaged speech spectrum. Trans. IECE 55-A (1), 549–556.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech & Audio 2 (2), 291–298.

Gauvain, J.L., Lamel, L.F., 1993. Identification of non-linguistic speech features. In: Proceedings ARPA Human Language Technology Workshop, Plainsboro, NJ, pp. 96–101.

Gauvain J.L., Lamel, L.F., Prouts, B., 1995. Experiments with speaker verification over the telephone. In: Proceedings ESCA Eurospeech'95, Madrid, pp. 651–654.

Gish, H., Schmidt, M., 1994. Text-independent speaker identification. IEEE Signal Process. Mag., October, 18–32.

Godfrey, J., 1994. Multilingual speech databases at LDC. In: Proceedings ARPA Human Language Technology Workshop, Plainsboro, NJ, pp. 23–26.

Godfrey, J., Holliman, E., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings IEEE ICASSP-92, San Francisco, CA, Vol. 1, pp. 517–520.

Lamel, L.F., Gauvain, J.L., 1992. Continuous speech recognition at LIMSI. In: Proceedings DARPA Artificial Neural Network Technology Speech Program, Final review, Stanford, CA.

Lamel, L.F., Gauvain, J.L., 1993. Identifying non-linguistic speech features. In: Proceedings Eurospeech'93, Berlin, Germany, Vol. 1, pp. 23–28.

Lamel, L.F., Gauvain, J.L., 1995. A phone-based approach to non-linguistic speech feature identification. Comput. Speech Language 9 (1), 87–103.

Lamel, L.F., Gauvain, J.L., 1997. Speaker recognition with the switchboard corpus. In: Proceedings IEEE ICASSP-97, Munich, pp. 1067–1070.

Lamel, L., Gauvain, J.L., Eskénazi, M., 1991. BREF, a large vocabulary spoken corpus for French. In: Proceedings ESCA Eurospeech'91, Genoa, pp. 505–508.

Legetter, J.C., Woodland, P.C., 1994. Speaker adaptation using linear regression. Technical Report, CUED/F-INFENG/TR.181.

Matsui, T., Furui, S., 1993. Concatenated phoneme models for text-variable speaker recognition. In: Proceedings IEEE ICASSP-93, Minneapolis, MN, Vol. II, pp. 391–394.

Naik, J.M., 1990. Speaker verification: A tutorial. IEEE Commun. Mag., 42–48.

Newman, M., Gillick, L., Ito, Y., McAllister, Peskin, B., 1996. Speaker verification through large vocabulary continuous speech recognition. In: Proceedings ICSLP'96, Philadephia, PA, pp. 2419–2422.

Przybocki, M.A., Martin, A.F., 1998. NIST speaker recognition evaluation – 1997. In: Proceedings RLA2C, Avignon, pp. 120–123.

Reynolds, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication. 17, 91–108.

Rosenberg, A.E., 1976. Automatic speaker verification: A review. Proc. IEEE 64 (4), 475–487.

Rosenberg, A.E., Soong, F.K., 1992. Recent research in automatic speaker recognition. In: Furui, Sondhi (Eds.), Advances in Speech Signal Processing, Marcel Dekker, NY, Chapter 22.

Rosenberg, A.E., Lee, C.H., Soong, F.K., 1990. Sub-word unit talker verification using hidden Markov models. In: Proceedings IEEE ICASSP-90, Albuquerque, NM, Vol. S5.3, pp. 269–272.