SPEAKER VERIFICATION OVER THE TELEPHONE*

L.F. Lamel, J.L. Gauvain
LIMSI - CNRS, B.P. 133, 91403 Orsay, France
{lamel,gauvain}@limsi.fr http://www.limsi.fr/TLP

RÉSUMÉ

Dans cet article nous présentons une étude sur l'authentification du locuteur à partir d'un signal téléphonique en mode dépendant et indépendant du texte. L'approche retenue consiste à modéliser le locuteur par une source markovienne de phones associée à des contraintes phonotactiques (mode indépendant du texte) ou à un lexique (mode dépendant du texte). Dans les deux cas les phones sont représentés par des modèles de Markov cachés gauche-droite à 3 états. Une serie d'experiences a été effectuée sur un corpus téléphonique enregistré spécifiquement pour l'évaluation d'algorithmes d'authentification du locuteur. Les résultats expérimentaux sont présentés pour différentes quantités de données d'apprentissage et de test, et pour de la parole spontanée et des textes lus. Sur ces données, le taux d'égale erreur le plus faible est 1% dans le mode dépendant du texte lorsque deux essais sont autorisés par tentative avec une durée minimale de 1.5s de parole par essai.

ABSTRACT

In this paper we present a study on speaker verification using telephone speech and for two operational modes, i.e. text-dependent and text-independent speaker verification. A statistical modeling approach is taken, where for text-independent verification the talker is viewed as a source of phones, modeled by a fully connected Markov chain and for text-dependent verification, a left-to-right HMM is built by concatenating the phone models corresponding to the transcription. A series of experiments were carried out on a large telephone corpus recorded specifically for speaker verification algorithm development assessing performance as a function of the type and amount of data used for training and for verification. Experimental results are presented for both read and spontaneous speech. On this data, the lowest equal error rate is 1% for the text-dependent mode when 2 trials are allowed per attempt and with a minimum of 1.5s of speech per trial.

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 [6, 7]. Despite these efforts and promising results using laboratory data, speaker verification performance over the telephone remains below that required for many applications. In this paper, we present an experimental study on speaker verification over the telephone for two operational modes, i.e. text-dependent and text-independent verification. Achievable performance levels are given for both the known and unknown-text conditions, using a large corpus of read and spontaneous speech.

A statistical modeling approach is taken, where the talker is viewed as a source of phones, modeled by a fully connected Markov chain[1, 3]. The lexical and syntactic structures of the language are approximated by local phonotactic constraints, and each phone is in turn modeled by a 3 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[1, 3] a set of phone models is trained for each speaker and identification of a speaker from the signal $\mathbf{x}$ is performed by computing the phone-based likelihood $f(\mathbf{x}|\lambda)$ 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 $f(\mathbf{x}|\lambda)/f(\mathbf{x})$ to a speaker independent threshold in order to decide acceptance or rejection.

METHODOLOGY

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[2]. This approach allows a large number of parameters to be estimated from a small amount of speaker-specific adaptation data. A set of context-independent phone models are built for each speaker.

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

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

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

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

Conditions	*Multi-style training*				*Type-specific training*		
	Average	*Digits*	*SEPT*	*Sentences*	*Digits*	*SEPT*	*Sentences*
SID rate	93.5	90.5	95.5	94.5	91.4	96.4	94.1
1 trial,	3.3	4.2	2.3	2.6	4.1	2.3	2.7
1 trial, $\geq$1.2s	2.6	2.9	1.8	2.6	2.9	1.8	2.7
2 trials,	2.7	3.1	1.7	2.0	3.2	1.8	2.2
2 trials, $\geq$1.2s	2.0	2.0	1.2	2.0	2.2	1.3	2.2
2 trials, $\geq$1.5s	1.8	1.4	1.0	1.9	1.6	1.1	2.1

Table 1: Speaker identification rate (single trial) and equal error rates (EER) for different test data types with multistyle training (left) and type-specific training (right), based on 21775 user attempts and 10908×91 imposter attempts. The text is known.

the model.) Calculating the denominator of this expression is 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 [1, 3] provides not only the likelihood of the most probable speaker, $f(\mathbf{x}|\lambda)$, but the likelihoods for the N most probable speakers. The neccessary 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 $f(\mathbf{x}, \mathrm{s}|\lambda)$ of the incoming signal and the most likely state sequence instead of $f(\mathbf{x}|\lambda)$.

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 straight-forward approach is to base the decision only on the score $L(\mathbf{x};\lambda)$ of the second attempt, ignoring the preceding trial. This approach can be justified on the ground that the actual test data is potentially invalid. An aternative it is to base the decision on the scores of both trials.[1] Making use of this second approach reduced the error rate by 21%, compared to a 13% error reduction using only the score of the last attempt.

For these experiments we make use of a corpus especially designed to evaluate speaker recognition algorithms.[2] This corpus contains data from 100 target speakers or users, and from 1000 impostors[8]. Each user completed 10 training sessions, and 25 verification calls, from a variety of telephone handsets and calling locations. Each call provides a variety of speech data, including a variety of read speech material and ellicited and spontaneous speech so as to be able to assess the effects of data type on the verification accuracy. The read test data consist of three types: digit strings, 5 phonetically controlled sentences (SEPT), 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 so as to obtain short monologues. The acoustic feature vector contains 13 cepstrum coefficients derived from a Mel-frequency spectrum (0-3.5kHz 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.

MULTISTYLE VS TYPE-SPECIFIC TRAINING

Experiments were carried out to assess the influence of the amount and type of data used for training speaker-specific models and for the authorization attempts.[3] The first row in Table 1 compares text-dependent speaker identification 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.

Using multi-style training, the average identification rate across the 21775 test samples from the 25 test sessions is 93.5%. If a minimal duration of 1.2s is required (not shown in table), the average identification rate is 94.7%, and about 10% of the data (mostly 3 digit sequences) are not used. For longer durations (minimum 2s) the average identification rate is 95.6%. Comparing the different types of test data, the highest identification rates are obtained for the SEPT sentences. When type-specific training is used, and testing is carried out on the same type of data, the speaker identification rates are slightly higher for the digits and the SEPT sentences, and slightly lower for the *Le Monde* sentences.

The lower part of Table 1 gives the known-text equal error rates (EER) for the different data types for multistyle and type-specific training. 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

[1] It is evidently possible to allow more than 2 trials per attempt, in which case the score would take into account scores from all previous trials.

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

[3] Results are reported for speaker identification or for verification somewhat interchangably, as we have found that speaker identification error rates are directly related to speaker verification error rates, yet the speaker identification error is much easier to measure.

	Test data		
Training data	*Digits*	*SEPT*	*Sentences*
Digits	8.6	68.6	35.6
SEPT	64.1	3.6	24.3
Sentences	21.1	14.3	5.9

Table 2: Speaker identification error rates with type-specific training for same-type and cross-type test data.

is reduced to 1.8% if a miminal duration of 1.5s is required and two trials per attempt are allowed. Under all conditions the SEPT sentences have the lowest EER. This performance is attributed to the limited phonetic contexts found in the SEPT sentences, enabling them to be well-modeled using the 25 repetitions occurring in the complete set of training data.

In order to evaluate the importance of the linguistic content of the training data, we investigated the performance under crossed training/test conditions. The speaker identification error rates are given in Table 2. There is a large degradation in performance when the training and test data are of different types. The *Le Monde* sentences have the least degradation as they have the largest variety of phonetic contexts. The highest errors occur between the SEPT sentences and the digit strings, where there is a large difference in linguistic content.

AMOUNT & RECENCY OF TRAINING DATA

The amount and recency of the training data are well known factors that influence speaker verification performance. These effects are quantified here by comparing 3 session training with single session training (first or last session), and with 1/3 of the training data taken from each of the 3 training sessions. The entries in Table 3 correspond to known-text EER results for the 3 session training, single (last) session training, and training on one-third of the data from each of 3 sessions. As expected, the EER is seen to significantly increase when the training data is reduced to one session. Training on the same amount of data (1/3 from each of 3 sessions) reduces this performance degradation.

Conditions	*Digits*	*SEPT*	*Sentences*
1 trial,	2.9/4.8/3.8	1.9/3.1/2.3	3.2/3.5/3.2
1 trial, $\geq$1.0s	2.8/4.5/3.6	1.9/3.1/2.3	3.2/3.5/3.2
2 trials,	1.7/3.1/2.6	1.3/2.2/1.7	2.5/2.6/2.6
2 trials, $\geq$1.0s	1.8/3.0/2.5	1.3/2.3/1.7	2.5/2.6/2.6
2 trials, $\geq$1.5s	1.7/2.5/2.5	1.3/1.9/1.6	2.5/2.6/2.6

Table 3: Equal error rates as a function of the amount of training data, using type-specific training. For each condition the 3 EERs correspond to training on 3 sessions, the last session, and 1/3 of each of 3 sessions. The text is known.

Table 4 gives the speaker identification error as a function of the amount of training data, and the proximity to the test data. Although not new, these results quantify the need for multiple training sessions. Additionally, speaker-adaptation techniques can be used to reduce the effects of model ageing.

Training	*Digits*	*SEPT*	*Sentences*
3 sessions	4.8	3.2	6.7
1/3 of 3 sessions	6.4	4.1	6.6
1 session (last)	10.8	6.3	8.3
1 session (first)	19.7	11.5	16.7

Table 4: Speaker identification error rates for different training conditions. Known text.

SPONTANEOUS SPEECH

Experiments were carried out to measure the speaker identification and verification rates on spontaneous speech using the fixed and open questions. The responses to the fixed questions were much shorter (1.5s on average) than to the open questions (8.2s on average). These experiments compare known and unknown text conditions. In the unknown text mode, a speaker-independent phone recognizer is used to provide a phone transcription of the utterance, which is then used for identification or verification.

Questions	*Avg. duration*	*Known text*	*Unknown text*
calling place	1.3s	74.4	66.4
telephone type	1.7s	84.1	72.6
handset type	1.3s	77.7	66.0
city/country	0.9s	65.5	53.6
postal code	1.3s	75.2	61.7
telephone no.	1.6s	86.0	79.8
date	2.3s	82.8	73.7
time	1.6s	73.2	61.3

Table 5: Speaker identification rates for the fixed questions with multistyle training.

Table 5 shows the speaker identification rates for the different types of fixed questions (about 1200 trials for each type). There is a clear correspondance between the average duration of the response and the identification rate, with the lowest rates being obtained on the shortest responses, such as "city/country" where the callers often gave a single word response.

Figure 1 compares the ROC curves with (left) and without (right) the use of transcriptions. The ROC curves for the digits, SEPT and *Le Monde* sentences are given for comparison. The error rates are significantly higher for the spontaneous speech than for the read texts. The equal error rate when the transcription is known is about 5.0% compared to about 6.5% in text-independent mode for spontaneous speech.

DISCUSSION AND CONCLUSION

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 com-

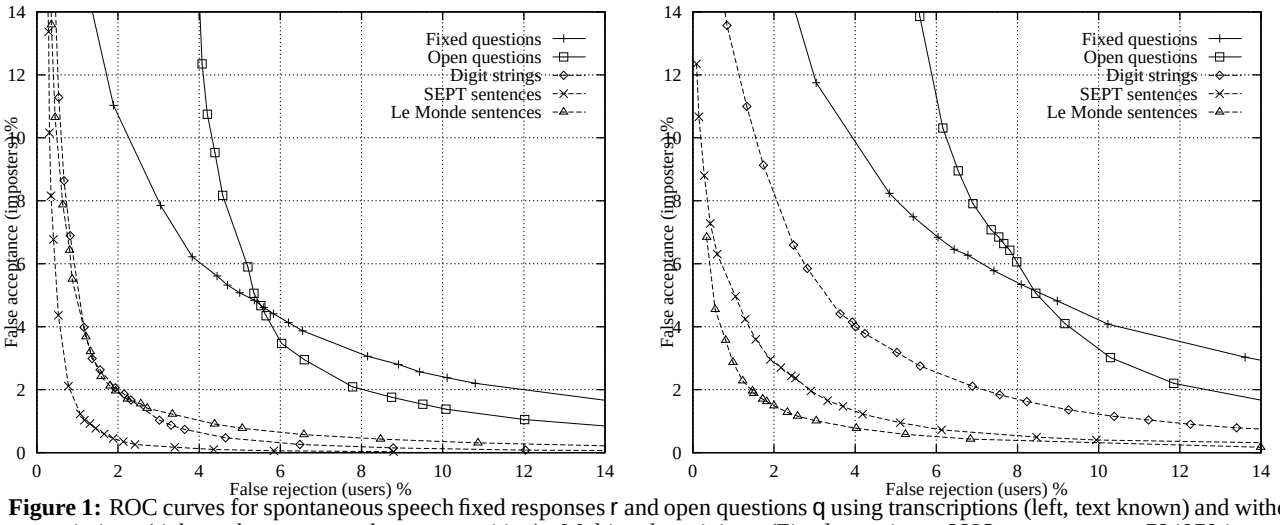

Figure 1: ROC curves for spontaneous speech fixed responses r and open questions q using transcriptions (left, text known) and without transcriptions (right, unknown text, phone recognition). Multi-style training. (Fixed questions: 8823 user attempts, 794070 imposter attempts (simulated); Open questions: 4691 user attempts, 422190 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.

parable amounts of training data, better performance is obtained when the data is taken from several training sessions, as opposed to all from a single call. Type-specific training results in better performance when the same type of test data is used. If the test data is different in linguistic content (or uncontrolled), multistyle training is to be prefered. The importance of phonetic content was illustrated for the crossed-type conditions, which led to significant degradation in performance (see Table 2).

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 there are only a few (5) different forms and 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 users.

Identification and verification performances on spontaneous speech (for both text-dependent and text-independent modes) are substantially worse than performance on read speech. This significantly higher error rate can be partly attributed to a larger variation in speaking style, the short duration of the responses, and the larger variability in phonetic contexts. Another factor which has not yet been investigated is that the acoustic models were trained only on the read-speech data. Training on spontaneous speech may reduce the performance difference. Although better performance is obtained in the known-text condition, that is using the transcription of the data, this is not very realistic, as in general one cannot assume that the transcriptions of spontaneous speech are available.

Concerning the amount of data needed, estimation of speaker-specific models requires a minimum of about 25 sentences, corresponding to about 1 minute of speech. For the test utterances, performance is better for longer durations, indicating that it is advantageous to ensure a minimal duration of at least 1.5 or 2s. An equal error rate of 1% was obtained on the SEPT sentences, in the text-dependent mode with 2 trials per verifcation attempt and with a minimum of 1.5s of speech per trial.

REFERENCES

[1] J.L. Gauvain, L.F. Lamel, "Identification of Non-Linguistic Speech Features," *Proc. ARPA Human Language Technology Workshop*, March 1993.

[2] J.L. Gauvain, C.H. Lee, "Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech & Audio*, **2**(2), April 1994.

[3] L.F. Lamel and J.L. Gauvain, "A Phone-based Approach to Non-Linguistic Speech Feature Identification," *Computer Speech and Language*, **9**(1), pp. 87-103, Jan. 1995.

[4] A.L. Higgins, L. Bahler, J. Porter, "Speaker Verification Using Randomized Phrase Prompting," *Digital Signal Processing*, **1**, 1991.

[5] A.E. Rosenberg, "The Use of Cohort Normalized Scores for Speaker Verification." *ICSLP-92*.

[6] J.M. Naik, "Speaker Verification: A Tutorial," *IEEE Communication Magazine*, pp.42-48, Jan 1990.

[7] H. Gish, M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, pp. 18-32, Oct 1994.

[8] J.L. Gauvain, L.F. Lamel, B. Prouts, "Experiments with speaker verification over the telephone," *Eurospeech'95*.

[9] J.L. Gauvain, L.F. Lamel, B. Prouts, Final report Marché France Telecom No. 94 6M 714, "Authentification vocale du locuteur à travers le réseau téléphonique", May 1997.