SPEAKER VERIFICATION IN THE TELEPHONE NETWORK : RESEARCH ACTIVITIES IN THE CAVE PROJECT

Frédéric BIMBOT [1] *Hans-Peter HUTTER* [5] *Cédric JABOULET* [2]
Johan KOOLWAAIJ [4] *Johan LINDBERG* [3] *Jean-Benoît PIERROT* [1]

(1) ENST / CNRS (2) IDIAP (3) KTH (4) KUN (5) Ubilab-UBS

bimbot@sig.enst.fr hans-peter.hutter@ubs.com jaboulet@idiap.ch
koolwaaij@let.kun.nl lindberg@speech.kth.se pierrot@sig.enst.fr

http://www.PTT-Telecom.nl/cave

ABSTRACT

This paper summarizes the main results from the Speaker Verification (SV) research pursued so far in the CAVE project. Different state-of-the art SV algorithms were implemented in a common HMM framework and compared on two databases : YOHO (office environment speech) and SESP (telephone speech). This paper is concerned with the different design issues for LR-HMM-based SV algorithms which emerged from our investigations and which led to our current SV system, which delivers Equal Error Rates below 0.5 % on a very realistic telephone speech database.

1. CONTEXT

The CAVE project (CAller VErification in Banking and Telecommunications) is a 2 year project supported by the Language Engineering Sector of the Telematics Applications Programme of the European Union, and for the Swiss partners by the Office Fédéral de l'Education et de la Science (Bundesamt für Bildung und Wissenschaft). The partners are Dutch PTT Telecom, KUN, KTH, ENST, UBILAB, IDIAP, VOCALIS, TELIA and Swiss Telecom PTT. It started on December 1^{st}, 1995.

The technical objectives of the CAVE project are to design, implement and assess 2 telephone-based systems which use speaker verification technology. Work Package 4 (WP4) of this project focuses on the research and development aspects. This paper describes the methodology, experiments and results of the work performed so far within WP4.

[1] ENST - Dépt Signal, CNRS - URA 820, 46 Rue Barrault, 75634 Paris cedex 13, FRANCE-EU

[2] IDIAP, Rue du Simplon 4, Case Postale 592, CH-1920 Martigny, SWITZERLAND

[3] KTH, Department of Speech, Music and Hearing, Drottning Kristinas Väg 31, S-100 44 Stockholm, SWEDEN-EU

[4] KUN, Dept of Language & Speech, Erasmusplein 1, NL-6525 HT Nijmegen, THE NETHERLANDS-EU

[5] Ubilab, Union Bank of Switzerland, Bahnhofstrasse 45, CH-8021, Zürich, SWITZERLAND

2. THE CAVE GENERIC SPEAKER VERIFICATION SOFTWARE

Initially, the partners devoted some effort to building and validating a common suite of speaker verification software. The CAVE generic speaker verification software package is based on HTK (Hidden Markov Modelling Toolkit), version 2.0 [1]. Though HTK is intended for Hidden Markov Modelling (HMM), we used it to simulate a wide variety of text-dependent, text-prompted and text-independent approaches to SV.

In addition to Left-Right and Ergodic HMMs, we implemented Vector Quantization, Gaussian Mixture Modelling and even a particular form of Dynamic Time Warping, all under the single HTK framework. The formalisation of such a variety of algorithms within a common platform is an additional guarantee that variations in performance for differing approaches can only be attributed to the model.

In practice, the main difference in the implementation of text-dependent, text-prompted and text-independent approaches is the nature of the speech segment that is modelled : a particular word (or sentence) for text-dependent approaches, a set of words or subwords for fixed-vocabulary text-prompted approaches, or the entire speech material for text-independent approaches. Once this is specified, 4 main characteristics are sufficient to determine a particular speaker verification algorithm : the number of states for each speech segment, the number of Gaussian densities in each state, the covariance matrices (full or diagonal, fixed or learnable) and the transition matrix (fixed or learnable).

For instance, a text-dependent VQ algorithm using Euclidean distance is implemented as a word-based Ergodic HMM, with the probability density function of each state modelled by a single Gaussian distribution with an identity covariance matrix, and equal probability of transition to all other states. Only the state means are trained during the enrolment. Similarly, a word-based DTW approach is implemented as an HMM model whose number of states is equal to the number of frames in the enrolment utterance. When several training utterances are used, the corresponding model is composed of several branches in parallel.

3. VERIFICATION ALGORITHM

Let Y_1^N be a test speech token and $\mathcal{X}$ the speaker model for the claimed speaker. The verification algorithm depends on 3 models; the (claimed) speaker model, $\mathcal{X}$, a (speaker-independent) world-model, Ω, and a speaker-independent silence model, $\mathcal{S}$. Each registered speaker model is trained on enrolment speech data from that speaker. The world-model is trained on a distinct set of speakers, who are neither registered speakers, nor used for impostor accesses. The silence model is trained on non-speech portions from the enrolment data. After enrolment, a *client+silence* model ($\mathcal{X} \cup \mathcal{S}$) is constructed by merging the speaker model and the silence model. The *client+silence* model provides a global likelihood score $\mathcal{L}\,(Y_1^N \mid \mathcal{X} \cup \mathcal{S})$. Similarly, the world-model is merged with the silence model, and the *world+silence* model is used to compute a likelihood which can be denoted as $\mathcal{L}\,(Y_1^N \mid \Omega \cup \mathcal{S})$. Ultimately, the log likelihood ratio :

$$LLR\,(Y_1^N) \;=\; \log \left[\frac{\mathcal{L}\,(Y_1^N \mid \mathcal{X} \cup \mathcal{S})}{\mathcal{L}\,(Y_1^N \mid \Omega \cup \mathcal{S})}\right] \qquad (1)$$

is calculated, and compared to a speaker-dependent threshold $\Theta\,(\mathcal{X})$ for the acceptance/rejection decision.

In theory, the score of equation (3.) can be considered as an approximation of the likelihood ratio on speech portions only. In fact, under the hypothesis that most non-speech frames are simultaneously assigned to the same state of $\mathcal{S}$ in the merged models $\mathcal{X} \cup \mathcal{S}$ and $\Omega \cup \mathcal{S}$, the corresponding terms in the likelihood ratio LR cancel out. In practice however, the alignments with the *client+silence* model and with the *world+silence* model often differ quite significantly, and if these alignments are forced to be the same, the level of performance degrades. It seems therefore that model $\mathcal{S}$ plays the more general role of *garbage* model, by preventing very low likelihood values for portions of the utterance which do not fit the *client-* and/or *world-* models.

4. SCORING PROCEDURE

Together with the CAVE generic system, a basic set of scoring procedures was developed in order to standardise performance evaluation across the partners. In the experiments reported here, we measure the performance in terms of Gender-Balanced Sex-Independent Equal Error Rate (GBSI-EER), following EAGLES recommendations [2]. The GBSI-EER is obtained in the following way :

a) for each registered speaker X_i (male or female), a threshold value Θ_i is computed (a posteriori) so as to equalise the False Rejection Rate and the False Acceptance Rate, taking into account male and female impostors with an equal contribution,

b) the corresponding EER (E_i) for speaker X_i is then computed,

c) all E_i are averaged across male clients to generate E_M, across female clients to generate E_F, and then these 2 separate scores are themselves averaged : $E = (E_M + E_F)/2$.

This score corrects possible imbalance between the number of female and male speakers in the tested population. It assumes that impostors do not know the sex of the genuine client in advance.

5. VALIDATION

A first test campaign was set up using the YOHO database. In this campaign, each partner was in charge of testing particular configurations of the generic system on a common set of tasks. This set of experiments was carried out in order to validate the generic system on a well-known task.

YOHO is a database for text-prompted speaker verification [3]. Out of the 138 speakers in this database, we randomly selected 10 male and 10 female speakers to create the world-model. Each of the remaining 96 male and 22 female speakers was used as a registered speaker and as an impostor against some set of YOHO speakers. Various amounts of enrolment material were used, and several HMM topologies and complexities were investigated. The tests were carried out on the 9440 YOHO test utterances, which were used both as true claims and as random impostor attempts.

We report only briefly on the best configurations tested on YOHO in text-dependent mode, which were variants of Left-Right (LR) HMMs. The HMM-LR topology is characterised by the number p of states per phoneme and the number q of Gaussian mixtures per state. In all experiments reported in this paper, covariance matrices are diagonal. We train each speaker model with speech material recorded in 4 distinct sessions. The first experiment uses 6 utterances from each training session, while the second is based on all the available training material, i.e 24 utterances per session. Tests are performed on one utterance only.

In this series of experiments, acoustic features are 12 LPC-derived Cepstral Coefficients (LPCC)[6], together with the log-energy, plus the first and second derivatives of these features, leading to a total of 39 coefficients per frame. Cepstral mean subtraction is used, to simulate the processing that is classically used on telephone speech, even though it probably degrades the performance on YOHO.
The results given in Table 1 show that the CAVE generic speaker verification system yields state-of-the-art performance on the YOHO database.

[6] The window size is 25.6 ms, window shift is 10 ms, preemphasis factor is 0.97, a Hamming window is used, and the LPC model is of order 16.

$p = 1$ $q = 5$
enrolment : 4 sess. $\times$ 6 utt. 0.21 %
enrolment : 4 sess. $\times$ 24 utt. 0.03 %

Table 1. Gender-Balanced Sex-Independent (GBSI) Equal Error Rate on the YOHO database, for different enrolment conditions (test on one utterance only)

6. TELEPHONE SPEECH EXPERIMENTS

Following the positive results of the first test campaign on YOHO, WP4 switched to a real telephone speech database, the SESP database, whose contents can be considered as very representative of real-world telephone speech.

7. THE SESP DATABASE

SESP is a database collected by KPN Research. It contains telephone utterances of 24 male and 24 female speakers calling with different handsets (including some calls from mobile phones) from a wide variety of places (such as restaurants, public phones and airport departure lounges). All the recordings were made between March and May 1994. A substantial proportion of the calls was made from foreign countries. In our experiments, the 21 male and 20 female speakers for whom there is sufficient speech material, are used as clients.

The speech material under focus in this paper is "Scope" (telephone calling-card) numbers; sequences of 14 digits uttered in a more or less continuous fashion. A full session contains 2 utterances of such items. For each speaker, speech recorded in 4 distinct sessions was selected as enrolment material[7]. The other sessions were considered as test sessions.

No obvious factor makes the SESP data significantly different from those that could be expected from a field test data collection, except for the lack of intentional impostor attempts.

8. EXPERIMENTAL PROTOCOL

As discussed in section 3, the CAVE generic system is based on likelihood normalisation using a world-model. For estimating the parameters of this model for the SESP experiments, we used a small subset of the Dutch Polyphone database, corresponding to 24 male and 24 female speakers, and consisting of 6 sequences of digits, the length of which ranges from 4 to 16. All speakers are distinct from SESP speakers. In our experiments, the world-model has the same topology as the speaker models.

As for the YOHO experiments, we report the results obtained with HMM-LR topologies, since they yielded the best results for text-dependent verification. We studied the influence of the HMM topology in terms of number of states per phoneme and number of (diagonal) mixtures per state (respectively p and q). We also investigated the impact of the number of enrolment sessions (between 2 and 4) and the effect of the acoustic features.

Tests were carried out on a single utterance of the card number. Each trial consisted of 1658 genuine accesses and 1016 impostor attempts.

9. RESULTS

The first set of results, reported in Table 2 illustrates the impact of both the HMM topology and the number of enrolment sessions. Here, we use the same acoustic features as those described in section 5. Again, we use cepstral mean subtraction to compensate for channel variation between calls.

$p \times q$	1	2	3	4	5	6
$p : q$	1:1	1:2	1:3	1:4	1:5	1:6
4 sess.	1.29	0.86	0.58	0.56	0.50	0.54
3 sess.	1.27	0.80	0.69	0.52	0.62	0.44
2 sess.	1.39	0.92	0.76	0.67	0.80	0.69
$p : q$		2:1		2:2		2:3
4 sess.		0.72		0.50		0.30
3 sess.		0.86		0.52		0.50
2 sess.		0.88		0.78		0.61
$p : q$			3:1			3:2
4 sess.			0.47			0.47
3 sess.			0.63			0.32
2 sess.			0.79			0.44
$p : q$				4:1		
4 sess.				0.45		
3 sess.				0.58		
2 sess.				0.85		
$p : q$					5:1	
4 sess.					0.49	
3 sess.					0.54	
2 sess.					0.78	
$p : q$						6:1
4 sess.						0.39
3 sess.						0.55
2 sess.						1.00

Table 2. GBSI Equal Error Rate (in %) on the SESP telephone speech database, as a function of the HMM topology $p : q$ (p : number of states, q : number of mixtures per state), for different numbers of enrolment sessions (4, 3 or 2) - LPCC coefficients..

As has been observed in other publications [4], it is mainly the product $p \times q$ that governs the level of performance in our experiments. Optimal results with LPCC coefficients are obtained with (p:q) = (2:3) or (3:2), depending on the number of enrolment sessions. However, with the overall EER being so low, and the

[7] As enrolment sessions, we chose 4 sessions with a low level of background noise.

$p \times q$	1	2	3	4	5	6
$p:q$	1:1	1:2	1:3	1:4	1:5	1:6
4 sess.	2.07	0.86	0.53	0.51	0.31	0.33
3 sess.	2.46	1.34	0.87	0.74	0.59	0.58
2 sess.	2.97	1.58	0.94	1.04	0.80	0.89
$p:q$		2:1		2:2		2:3
4 sess.		0.95		0.38		0.43
3 sess.		1.18		0.83		0.65
2 sess.		1.47		0.87		0.86
$p:q$			3:1			3:2
4 sess.			0.52			0.35
3 sess.			0.91			0.65
2 sess.			1.05			0.79
$p:q$				4:1		
4 sess.				0.55		
3 sess.				0.65		
2 sess.				0.95		
$p:q$					5:1	
4 sess.					0.49	
3 sess.					0.53	
2 sess.					0.89	
$p:q$						6:1
4 sess.						0.41
3 sess.						0.68
2 sess.						0.90

Table 3. GBSI Equal Error Rate (in %) on the SESP telephone speech database, as a function of the HMM topology $p:q$ (p : number of states, q : number of mixtures per state), for different numbers of enrolment sessions (4, 3 or 2) - MFCC coefficients..

differences so small, it is difficult to draw definite conclusions on the significance of this optimum.

Unsurprisingly, the performance generally degrades with fewer enrolment sessions. However, the error rate remains in a very acceptable range, even with only 2 enrolment sessions.

Table 3 reports equivalent results obtained with 12 MFCC (Mel Frequency Cepstrum Coefficients), on similar test configurations. Here, the 12 coefficients are obtained from a bank of 20 filters arranged along the Mel scale over the full 0-4000 Hz band. As for LPCCs, the log-energy of the signal is also used, and the set ot 13 acoustic features is augmented with delta and delta-delta parameters, leading to a total of 39 coefficients. Here, the advantage seems in favour of LPCC coefficients, but the differences are still very small when p and q are optimal.

We finally report on some additional experiments, comparing results obtained with 12 and 16 LPCC coefficients (from the same 16th order LPC model). The results are given in Table 4. These results show that an additional gain in performance can be expected from this larger set of acoustic features. However, these results need to be confirmed on a wider scale.

topol. (p:q)	nb enr. sess.	12 LPCC	16 LPCC
2:3	4	0.30	0.23
2:3	3	0.50	0.29
2:3	2	0.61	0.40
3:2	4	0.47	0.27
3:2	3	0.32	0.27
3:2	2	0.44	0.50

Table 4. Comparison of performances obtained with 12 and 16 LPCC coefficients, for a few HMM topologies (p:q) and enrolment conditions (number of sessions). GBSI-EER in %

A significant gain in the performance was achieved by a particular method for estimating the variance of the HMM states. It is not possible to describe the technique in detail here, but it will be the focus of a future paper.

10. CONCLUSIONS

Our experiments and results lead to the following conclusions. Firstly, the methodology adopted in the CAVE research activities is a very efficient one : the design of a generic system shared between the partners is an investment that yields very good returns from the scientific and technological viewpoints, as it allows fast and efficient communication between collaborating laboratories. Secondly, the level of performance obtained on the very realistic SESP database confirms our view that the technology is ready to be deployed in appropriate applications.

Beside the demonstration of two prototype systems at the end of the CAVE project, it is the intention of the CAVE-WP4 partners to make the generic system available in the public domain soon after the end of the project.

REFERENCES

[1] S. YOUNG, J. JANSEN, J. ODELL, D. OLLASON, P. WOODLAND *The HTK BOOK*, HTK 2.0 Manual. 1995.

[2] F. BIMBOT, G. CHOLLET *Assessment of speaker verification systems. In Spoken Language Resources and Assessment* EAGLES Handbook. 1995.

[3] CAMPBELL *Testing with the YOHO CD-ROM Voice Verification Corpus.*, Proc. ICASSP 95, vol. 1, pp. 341-344, Detroit, 1995.

[4] S. FURUI *An overview of speaker recognition technology.* ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, Martigny, 1994.