AN OVERVIEW OF THE CAVE PROJECT RESEARCH ACTIVITIES IN SPEAKER VERIFICATION

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

1. CONTEXT

The CAVE project (CAller VErification in Banking and Telecommunications) was 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 were Dutch PTT Telecom, KUN, KTH, ENST, UBILAB, IDIAP, VOCALIS, TELIA and SWISSCOM. The CAVE project terminated on November 30th, 1997.

The technical objectives of the CAVE project were to design, implement and assess 2 telephone-based systems which use speaker verification technology. Work Package 4 (WP4) of this project has been focusing on the research and development aspects. This paper describes the technology used and the results achieved by WP4.

2. THE CAVE-WP4 GENERIC SPEAKER VERIFICATION SYSTEM

Initially, the partners devoted some effort to building and validating a common software platform for speaker verification. The CAVE-WP4 Generic Speaker Verification software package is based on HTK (Hidden Markov Modelling Toolkit), version 2.x [1]. The complete description of the CAVE-WP4 Generic SV System is given in a companion paper [2].

3. VERIFICATION ALGORITHM

3.1. General approach

The systems on which we report in this paper are fixed-vocabulary text-dependent speaker verification systems. The vocabulary is composed of the 10 digits. The algorithmic approach is based on Hidden Markov Models, associated with a likelihood normalisation approach. We present in this section the main features that characterize the CAVE-WP4 configuration.

3.2. Speaker models

In the case of text-dependent speaker verification, the text of the utterance is known by the SV system as it goes together with the claimed identity. Therefore, the likelihood of the utterance for a given claimed speaker is computed as the likelihood of the speech segment for the sequence of word models or subword models of the claimed speaker that compose the expected linguistic content of the utterance. In other words, the speaker model of a speaker X yields, for a given speech utterance $\mathcal{S}$, and an expected linguistic content W, a likelihood value $\mathcal{L}(\mathcal{S}|X,W)$. In the rest of this paper, we will drop the dependence on W and refer to the client likelihood function as $\mathcal{L}(\mathcal{S}|X)$.

3.3. Likelihood normalisation

It was often observed that the acceptance/rejection decision, when based only on the likelihood value of the client model, is relatively unreliable. It is both theoretically and experimentally more efficient to base the decision on a normalised likelihood, also called *likelihood ratio*. In fact, this result is a direct consequence of Bayesian decision theory. With $\mathcal{P}(\mathcal{S}|X)$ denoting the probability that the observations in $\mathcal{S}$ have been produced by the claimed speaker (X), and $\mathcal{P}(\bar{X}|\mathcal{S})$ the probability that they have been produced by an impostor $(\bar{X})$, the Bayesian decision rule writes :

$$\frac{\mathcal{P}(\mathcal{S}|X)}{\mathcal{P}(\mathcal{S}|\bar{X})} \underset{\text{reject}}{\overset{\text{accept}}{\gtrless}} R \tag{1}$$

In practice, the probability $\mathcal{P}(\mathcal{S}|X)$ is approximated by the client-model likelihood $\mathcal{L}(\mathcal{S}|X)$ but there are several ways for estimating the term $\mathcal{P}(\mathcal{S}|\bar{X})$, the most classical two being the use of a *cohort model* [3] or the use of a *world-model*. Under the *world-model* approach, which we adopted in the CAVE experiments, a single model is trained from a pool of speech utterances produced by various speakers (usually

---

[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

non-client speakers). If we denote as Ω the set of these speakers, the world-model approach consists of the approximation $\mathcal{P}(\mathcal{S}|\bar{X}) \approx \mathcal{L}(\mathcal{S}|\Omega)$. Whereas the literature does not show a systematic advantage on one approach over the other in terms of performance, the world-model has many attractive properties :

- a world-model is much less time-consuming in terms of computation : only one likelihood value has to be computed for the non-speaker model likelihood.
- a world-model is much more economical in storage volume than a client-dependent cohort model.
- a world-model does not require a selection of cohort speakers. This is all the more desirable as there is no well-established procedure for doing so.

3.4. Enrollment

In the context of fixed-vocabulary Speaker Verification, the training procedure consists of learning the parameters of a model for each speaker and each word of the vocabulary (in our case, the 10 digits). In practice, the clients are asked to pronounce several sequences of digits.

3.4.1. Initialisation

Even though embedded training theoretically allows the training of word models from sequences of words without requiring a segmentation of the training material, this very procedure can not be applied safely in the case of small amounts of training material, as is the case for a realistic SV application. Therefore we used an automatic speech recognition system in forced alignment mode, in order to segment the enrollment material into words and non-speech portions. This alignment was produced at KUN with the PHICOS system, obtained from PHILIPS Research under the LRE project MAIS. In a real application, this alignment stage can easily be implemented by a dedicated speech recognition system, even if this system is relatively basic.

3.4.2. Modified-EM training

In the case of limited training data, some of the covariance matrices of the Gaussian distributions within the HMM states may be difficult to train with robustness (even when diagonal). In fact, in the case of very few data, the variance of the mixtures may become so small that the model overfits the training data and can not generalise to new data. As a consequence, the conventional EM training algorithm can not be applied as such.

In our case, we put some constraints on the Gaussian densities so that they have a minimum variance. Moreover, this minimum is scaled to the range of variation for each coefficient. In practice, if we denote as s_k^2 the overall variance of the k^{th} coefficient, the *adaptive variance flooring* proceeds as follows :

$$\text{if} \quad \sigma_{ijk}^2 < \gamma\, s_k^2 \quad \text{then} \quad \sigma_{ijk}^2 := \gamma\, s_k^2 \tag{2}$$

where γ corresponds to the *variance flooring factor*. In other words, if the variance of a particular coefficient in a given state becomes, during enrollment, $1/\gamma$ times smaller than the overall variance of that coefficient, overfitting is suspected and the variance is not allowed to decrease any further. Adaptive variance flooring is experimentally much more efficient than fixed flooring. A patent based on and generalizing this idea is currently under review.

3.4.3. Handling of untrainable models

Even on segmented training material, we observed occasional difficulties in training some word models for some speakers. For instance, when there were more mixtures than enrollment speech frames for a given word model. The approach which we adopted, is to replace the untrainable word model by the model of the same word in the world-model. This very simple approach is viable, provided a very limited number of words are untrainable.

3.5. Access

As opposed to the enrollment phase, the verification process does not require any segmentation, nor even an initial step of speech/non-speech detection.

3.5.1. Forced alignment

As the linguistic content of the (expected) verification utterance is known, the likelihood value of the test speech material is computed with a deterministic syntax, corresponding to the expected sequence of words (forced alignment mode).

3.5.2. Speech/non-speech detection

Silence portions or, more generally, non-speech portions, can be present between words and are not predictable. They are dealt with by an optional silence model which is inserted between successive words before the best path is searched for, during the computation of the utterance likelihood. This non-speech model is naturally speaker-independent and it is trained with signal portions labeled as non-speech, in the enrollment material.

By using the same non-speech model for both the client scoring and the world scoring, it can be expected that a vast majority of the non-speech frames will be assigned to the same states in both decoding processes, and will therefore have the exact same likelihood value. If so, these terms cancel out in the likelihood ratio (LR) and have no contribution to the decision. This is illustrated in equation (3), where the symbol # denotes non-speech.

$$\begin{aligned} LR &= \frac{\mathcal{L}(\# \oplus s_1 \oplus \# \oplus s_2 \oplus \ldots \oplus s_p \oplus \# \mid X)}{\mathcal{L}(\# \oplus s_1 \oplus \# \oplus s_2 \oplus \ldots \oplus s_p \oplus \# \mid \Omega)} \\ &= \frac{\mathcal{L}(\#|X)\,\mathcal{L}(s_1|X)\,\ldots\,\mathcal{L}(s_p|X)\,\mathcal{L}(\#|X)}{\mathcal{L}(\#|\Omega)\,\mathcal{L}(s_1|\Omega)\,\ldots\,\mathcal{L}(s_p|\Omega)\,\mathcal{L}(\#|\Omega)} \\ &\approx \frac{\mathcal{L}(s_1|X)\,\ldots\,\mathcal{L}(s_p|X)}{\mathcal{L}(s_1|\Omega)\,\ldots\,\mathcal{L}(s_p|\Omega)} \\ &= \frac{\mathcal{L}(s_1 \oplus s_2 \oplus \ldots \oplus s_p \mid X)}{\mathcal{L}(s_1 \oplus s_2 \oplus \ldots \oplus s_p \mid \Omega)} \end{aligned} \tag{3}$$

Equality would hold if paths in the non-speech portions were perfectly identical for both decodings with the speaker and the world models, which would happen with forced borders for non-speech portions, but this would require an explicit speech/non-speech detection. The approximated approach is much simpler to implement and a few experiments showed no significant differences between both.

It is a similar principle that justifies the substitution of a particular word model by the model of this word in the world-model : if, the untrained word for speaker X is modeled by the world-model instead of a speaker-dependent model, the likelihood ratio simplifies for this word (under the assumption of perfect synchrony) during the decoding process.

3.5.3. *Likelihood time-normalisation*

In practice, the likelihood terms $\mathcal{L}(\mathcal{S}|X)$ and $\mathcal{L}(\mathcal{S}|\Omega)$ are evaluated as the product of frame-based terms (each of them corresponding to the product of an emission probability and a transition probability). Therefore, the logarithm of the likelihood ratio LR can be written as the sum of T frame-based log likelihood ratios, corresponding to each of the T observations in the speech utterance $\mathcal{S}$. Some authors find it appropriate to normalise the LR by dividing it by the utterance length. In our experiments, we found a very little but slightly negative impact of this normalisation. Therefore we carried out most of our experiments with the non-normalised likelihood ratio.

3.6. 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 [4]. 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 unbalance 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.

All results reported in this paper correspond to the GBSI-EER, i.e. with speaker-dependent a posteriori threshold setting. However, other experiments were carried out in the CAVE project with a priori threshold setting. Details are given in a companion paper [5].

4. EXPERIMENTAL PROTOCOL

4.1. 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 placed 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 composed of "Scope" (telephone calling-card) numbers, namely sequences of 14 digits uttered in a more or less continuous fashion, some corresponding to the speaker's scope number, some others to other clients'. Each session contains 2 utterances of the speakers own's number. For each speaker, speech recorded in 4 distinct sessions was selected as enrolment material. 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.

4.2. Acoustic analysis

In the course of the project, we tested several sets of acoustic features. In all the experiments we used a pre-emphasis of 0.97, a frame size of 25.6 ms, a frame shift of 10 ms and a Hamming windowing.

4.2.1. *FFT-derived cepstrum coefficients*

The first step for the calculation of these coefficients is the computation of the magnitude of the Fourier Transform. Then, we use a set 20 of triangular filters on the full frequency range between 0 and 4000 Hz, following a Mel scale (MFCC coefficients). The cepstrum coefficients are obtained as the coefficients between rank 1 and m of the cosine transform of the log-energy computed in each filter. The logarithm of the frame energy is added to the feature vector.

4.2.2. *LPC-derived cepstrum coefficients*

The LPC Cepstrum coefficients are obtained from a (16th order) linear prediction analysis (using the auto-correlation method). The LPCC coefficients correspond to a 0-4000 Hz bandwidth and a linear frequency scale. Here also, a log-energy coefficient is added to the feature vector.

4.2.3. *Additional processings*

We apply long-term cepstral mean substraction (estimated separately on each utterance). We extend the feature vector with delta coefficients computed from 5 successive cepstral coefficients and with delta-deltas computed from 5 successive deltas.

4.2.4. *Size of the feature vector*

In summary, a frame is represented by a vector of m Cepstrum Coefficients (MFCC or LPCC) plus a log-energy coefficient, and (an approximation of) their first and second derivatives, namely a vector of $3 \times (m+1)$ coefficients. Typical values of m are between 12 and 16.

4.3. Experimental protocol

4.3.1. *World-model*

For estimating the parameters of the world-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 client models.

4.3.2. HMM topology

In this paper, we report the results obtained with HMM-LR topologies (p states per phonemes $\times$ q Gaussian densities per state), since they yielded the best results for text-dependent verification. All covariance matrices are diagonal.

4.3.3. Test configuration

Tests were carried out on a single utterance of the card number. Each trial consisted of 1658 genuine accesses and 1016 impostor attempts (75 % same-sex, 25 % cross-sex).

5. RESULTS

We studied the influence of the HMM topology in terms of number of states per phoneme and number of (diagonal) mixtures per state. We also investigated the impact of the acoustic analysis, the influence of the flooring factor and the effect of the number of enrolment sessions.

5.1. Impact of the HMM topology

In a series of experiments, we compare the impact of the HMM topology, characterised by p and q for an enrollment in 4 sessions. We use the adaptive variance flooring technique, with a flooring factor equal to 1.0. Figure 1 depicts the GBSI-EER as a function of the product pq, with LPCC coefficients. Figure 2 gives corresponding results with MFCC coefficients. With both parameterisations, it can be seen that the product pq is the main factor that governs the performance level, as noted in [3].

5.2. Impact of the type of acoustic analysis

The comparison of Figure 1 and Figure 2 shows that, while LPCC coefficients outperform MFCC coefficients for simple HMM topologies ($pq \leq 2$), this advantage disappears with more elaborate models, namely those for which $pq \geq 2$.

Figure 3 compares, for the 2×3 and 3×2 topologies, the level of performance with 12 and 16 LPC-16 cepstrum coefficients, for several enrollment configurations. The speech parameterisation based on 16 LPC-16 coefficients shows a slight advantage over 12 LPC-16 cepstrum coefficients.

5.3. Impact of the flooring factor

On Figure 4, we have depicted two series of results obtained with 2 enrollment sessions only : the first series corresponds to a fixed variance floor (of 0.005) on the individual variance of each coefficient (standard HTK setting) whereas the second series corresponds to an adaptive variance floor of 1.0 (found to be optimal on most SESP experiments). A typical situation of overfitting is observed with a fixed variance floor : when the number of Gaussian mixtures per states increases, the overall performance decreases, indicating that some mixtures are getting over-specialised on the enrollment data. With adaptive flooring, the picture is inverse. Increasing the number of parameters in the HMM-LR has a beneficial effect : the error rate decreases with the HMM complexity and reaches smoothly an asymptot, at a level of performance which is 5 to 10 times better than with fixed flooring.

5.4. Impact of the number of enrollment session

We report here performance obtained with 4, 3, 2 and 1 enrollment sessions, with the experimental configuration that usually performed well in our experiments : a 2-state, 3-mixture Left-Right HMM with diagonal covariance matrices, with 12 LPC-16 cepstrum coefficients (+ log energy + Δs + $\Delta\Delta$s) and a variance flooring factor of 1.0. The 1-session condition is a very critical configuration for at least two reasons : long-term speaker variability is not represented in the enrollment data and the volume of enrollment material is very limited (here, two utterances).

Table 1 gives the performance obtained for that particular experiment. The error rate with one enrollment session is about multiplied by almost 2.5 as opposed to two enrollment sessions. Even though the error rate can still be considered reasonably low, the degradation observed shows how important it is, if feasible, to collect enrollment material in 2 sessions.

4 enrollment sessions	0.30 %
3 enrollment sessions	0.50 %
2 enrollment sessions	0.61 %
1 enrollment session	1.45 %

Table 1. Performance obtained on SESP using 1 to 4 enrollment sessions. 12 LPC-16 cepstrum coefficients + energy in dB, and their first and second derivative. HMM-LR 2:3 (diag). Adaptive variance flooring factor = 1.0. Scores are GBSI-EERs in %

5.5. Conclusions and perspectives

The low error rates reached on the SESP database are undoubtebly more than competitive as compared to the state-of-the-art performance reported on telephone data in the literature. Even though SESP has not been calibrated yet in other laboratories, EERs of 0.3 % with 4 enrollment sessions and of 0.6 % with 2 enrollment sessions are certainly challenging values for further research.

Many tracks remain to be explored. In particular, the adjustment of the topology of the world-model, the design of new techniques for HMM enrollment with scarce data, the use of some techniques of frame selection for robust likelihood ratio evaluation, the implementation of incremental enrollment, etc... Most of these issues will be addressed in the context of the Telematics-LE PICASSO project, starting in January 1998.

REFERENCES

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

[2] C. JABOULET, J. KOOLWAAIJ, J.-B. PIERROT, J. LINDBERG, F. BIMBOT : *The CAVE-WP4 generic speaker verification system* Submitted to RLA2C workshop, Avignon, 1998.

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

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

[5] J.-B. PIERROT, J. LINDBERG, J. KOOLWAAIJ, H.-P. HUTTER, D. GENOUD, M. BLOMBERG, F. BIMBOT : *A comparison of a priori threshold setting procedures for speaker verification in the CAVE project.* Submitted to RLA2C workshop, Avignon, 1998.

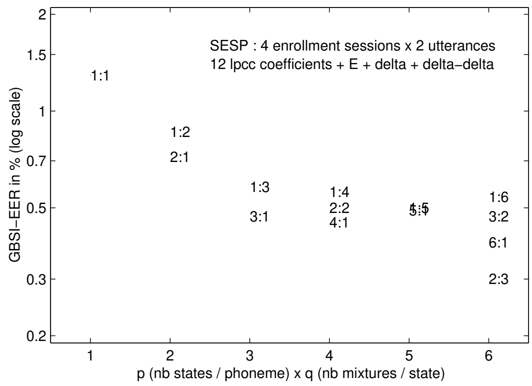

Figure 1. SESP : influence of the HMM topology on the Equal Error Rate. Acoustic features are 12 LPC-16 Cepstrum coefficients + log energy + Δs + $\Delta\Delta$s. *GBSI-EER in % (in log scale) as a function of pq.*

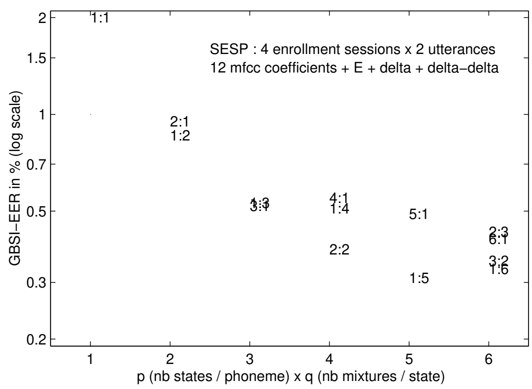

Figure 2. SESP : influence of the HMM topology on the Equal Error Rate. Acoustic features are 12 MFCC coefficients + log energy + Δs + $\Delta\Delta$s. *GBSI-EER in % (in log scale) as a function of pq.*

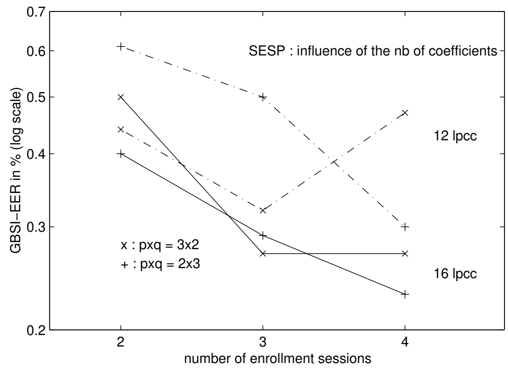

Figure 3. SESP : influence of the number of LPC-16 Cepstrum coefficients on the Equal Error Rate. Acoustic features are 12 or 16 LPC-16 Cepstrum coefficients + log energy + Δs + $\Delta\Delta$s. *GBSI-EER in % (in log scale) as a function of the number of enrollment sessions.*

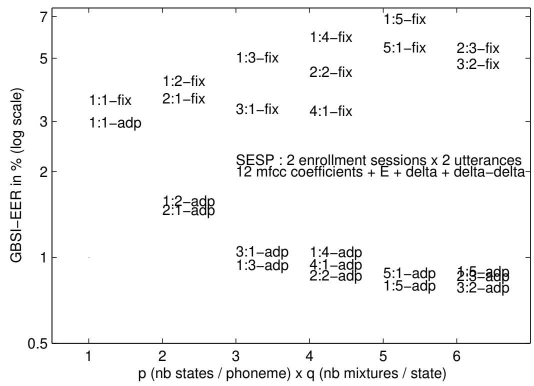

Figure 4. SESP : influence of the adaptive vs fixed variance flooring approach. Acoustic features are 12 LPC-16 Cepstrum coefficients + log energy + Δs + $\Delta\Delta$s. *GBSI-EER in % (in log scale) as a function of the HMM topology for fixed flooring (fix) and adaptive flooring (adp).*