The CAVE Speaker Verification Project – Experiments on the YOHO and SESP corpora*

David James[1], Hans-Peter Hutter[1] and Frédéric Bimbot[2]

[1] Ubilab, Union Bank of Switzerland, Bahnhofstrasse 45, CH-8021 Zürich, Switzerland
[2] Ecole Nationale Supérieure de Télécommunications, 46 Rue Barrault, 75634 Paris, France

Abstract. Many businesses, such as banks, will soon be under pressure to improve the functionality and friendliness of their telephone-based services without compromising security. Current automatic services authenticate callers with a PIN-code and consequently are not secure enough to offer callers the chance to carry out significant transactions; more sophisticated services necessarily depend on a human agent and demand lengthy authentication procedures. The technology of speaker verification (SV) has the potential to deliver fast and secure caller authentication for automatic or operator-assisted telephone services. This paper gives an overview of the work of the CAVE consortium, a pan-European group researching the application of SV to automatic telephone-based banking and calling-card services. Results obtained in experimental work so far compare favourably with the state-of-the-art.

1 Introduction

It is clear that some sectors of business are becoming less reliant on traditional "bricks-and-mortar" infrastructure. For example increasing numbers of credit-card purchases and all manner of other financial transactions are now being made by telephone, and the recent growth of the Internet has now given rise to the first financial services available over the World-Wide Web. However, whereas providers of these Internet services are careful to emphasise the encryption methods by which transactions can be kept hidden from prying eyes, automatic *telephone* services are at the moment secured only by relatively "snoopable" information, such as a personal identification number (PIN) code, spoken or transmitted using the telephone keypad. This is especially true of two kinds of telephone services; calling-card services and automatic telephone banking. Calling-cards, which allow cardholders to charge calls made from any telephone to a personal account with the card-issuing company, are extremely vulnerable to simple methods of fraud, which costs the industry and customers billions of

* This paper will appear at the 1st International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA), Crans-Montana, Switzerland, March 1997. Proceedings to be published by Springer-Verlag.

dollars every year. As regards telephone banking services, while they have not yet fallen victim to such levels of fraud, this is because automatic services typically offer extremely low functionality, and it is for the most part only through operator-assisted telephone banking services, with their lengthy authentication procedures, that a reasonable level of functionality can be obtained.

The European Caller Verification (CAVE) consortium has been set up to investigate the potential of speaker verification (SV) in order to protect telephone services against fraud whilst ensuring authentication speed and friendliness. It has been researching and developing speaker verification technology and has already begun to deploy this technology into field test systems. This paper gives a brief overview of the market areas which CAVE has been investigating and the research which has been undertaken so far within the framework of the project. The subsequent sections of this paper address these areas in turn.

2 Applications of Speaker Verification

2.1 Telephone Direct Banking

Even before the phenomenal growth of the World-Wide Web, it was estimated that direct banking, in the forms of telephone and PC-based systems, would account for more than half the banking transactions of 30% of the population of Europe by the turn of the century [2]. In addition, in the same survey, 79 out of 83 European banks surveyed predicted that they would be offering telephone banking by the year 2000. So far, telephone direct banking has proved popular since the most demanding piece of equipment it requires in the home is a touch-tone telephone.

However, since totally-automated telephone banking services can currently only authenticate callers through keypad entry of the customer number and PIN-code, their functionality is correspondingly limited. A full range of transactions such as bill payments, standing orders, and other forms of cash transfer, can at the moment only be offered over the telephone through a call centre, in which a large number of human agents deal with customers' telephone enquiries round-the-clock. In these services, caller authentication is currently performed through the use of a "security handshake"; this involves the asking of a number of questions to which only the true account-holder should know the answers. For example, one of Europe's market-leading telephone banks, interviewed by CAVE, authenticates customers by asking them to state 2 letters (chosen at random) from a secret password, and then any one of the following pieces of information:

- Date of Birth
- Mother's maiden name
- Place of Birth
- "Special" date (but not the caller's date of birth)
- "Special" address (but not the caller's home or work address)

Although the method works well, the bank feels it to be inconvenient and time-consuming. In addition, it is common practice for telephone banks to confirm major transactions before they are carried out by writing to the account holder's registered address; this introduces a delay of at least 48 hours (at least one working day for a letter to arrive, and another day to give the account holder a chance to react).

Currently, European banks with a telephone service offer either an automatic service or an operator-assisted one; however, there is increasing interest in combining the two approaches, offering a fast automatic service for those callers who only want to find out their balance (by far the most popular enquiry made of telephone banks and very frequently the sole purpose of a call) and an operator-assisted service for more complicated transactions. These two approaches could be combined extremely successfully using speaker verification, since it would allow a stronger authentication in the automatic service and no need for extra authentication if the caller chose to connect to a human agent. This would allow a greater range of functionality, such as bill payment, to be handed over to the automatic system.

2.2 Telephone Call Theft

Calling-cards are an obvious example of an existing telephone-based service in which caller authentication is too weak to secure the value of the transaction. In the United States, calling-card fraud cost $1 billion in 1995. The fraud arises from fraudsters observing telephone key-presses over the shoulders of legitimate card-holders, or using more sophisticated methods of surveillance, as they use their cards in places like airport lounges or hotel lobbies, and then selling this information on to third parties. Calling-card services are, generally speaking, automated, with operators always available to assist customers, for example when the card-holder is dialling from a rotary telephone. Whereas the annual figure of $1 billion calling-card fraud may sound large, it is undoubtedly tiny in comparison to the total yearly value of long-distance telephone traffic in the United States. However, since the card-holder is legally held responsible for the first $50 of unapproved calling charges, there is a very strong case for improving the security of these services. Speaker Verification has an obvious application to the reduction of calling-card fraud.

In addition, there are other areas hit hard by telephone call theft; companies' own private branch exchanges (PBXs), and mobile telephones. PBX fraud is carried out when fraudsters dial in to an exchange and break the access codes which restrict the ability to make outgoing calls, and costs hundreds of millions of dollars on both sides of the Atlantic every year. In a 1995 high-profile case, the UK's Metropolitan Police admitted that more than £1 million worth of calls had been made illegally on its internal PBX. AT&T has estimated that the crime costs an average of $20,000 (all borne by the owner of the PBX) every time a PBX is broken into. As regards the mobile network, this is also an extremely lucrative target for telephone fraudsters in Europe. In Great Britain alone, fraud losses due to analogue mobile telephone "cloning" (rewriting memory on a stolen telephone with signature information eavesdropped from another telephone) trebled to $150 million between 1994 and 1995, and the number of individual cloning instances *sextupled* in the 12 months to August 1995. However, significant human-factors and technical questions surround the potential application of SV in these cases.

3 CAVE SV Experiments

The CAVE Project is a two-year research project, supported by the European Union, which aims to research speaker verification technology and deploy it into field trials of real telephone services in the fields of telecommunications (represented in the project by Dutch PTT) and banking (represented by UBS). Other project partners include Vocalis, the UK speech technology company, KTH (Royal Institute of Technology, Stockholm), Telia, KUN (Catholic University of Nijmegen, Netherlands), ENST (Ecole Nationale Supérieure de Télécommunications, Paris) and IDIAP, (Institut Dalle Molle d'Intelligence Artificielle Perceptive, Martigny, Switzerland).

Effort so far has concentrated on determining the potential for the application of SV into existing services, developing specifications at the functional and technical level for our prototype systems which will be built, developing the experimental environment for SV research and trying to achieve state-of-the-art performance on available SV data collections. Further information about CAVE can be obtained by visiting our Web site, `http://www.ptt-telecom.nl/cave`.

3.1 Technology

Our verification systems implement text-dependent and text-prompted SV with vocabularies composed either of digits or of 2-digit numbers. In our target markets, verification utterances can be sequences of digits, either known in advance by the caller (a card number and/or PIN code), or randomly-generated digit sequences which the caller is prompted to repeat. Our approach has been based on the Hidden Markov Model (HMM), undoubtedly the start-of-the-art technology for speaker verification at the moment, with continuous Gaussian density modelling techniques; a wide variety of techniques can be thought of as differing forms of HMM, such as Gaussian Mixture Modelling (GMM), Vector Quantisation (VQ) and Dynamic Time-Warping (DTW), depending on which model parameters are fixed *a priori* or constrained during training. In this methodology, speaker enrolment consists only of learning the parameters for the models for each individual vocabulary term and each individual enrolled speaker, using standard HMM *isolated-word* training techniques.

HMM technology normally used for speech recognition is relatively simple to apply to SV, simply by knowing in advance the text of the utterance, whether imposed by the verification system or not. Using a speech recogniser to perform a word-level *alignment* results in the acoustic likelihood of the speech segment, for the sequence of digits or numbers known to compose the linguistic content of the utterance. More formally, if a speaker claims the identity X, and utters some speech S with the expected linguistic content W, the alignment procedure yields a likelihood value $L(S|X,W)$. Further, if a speaker-independent model Ω, approximating a model for $\bar{X}$ (i.e., speakers other than X), is used to obtain $L(S|\Omega,W)$, then the resulting quotient, the *likelihood ratio*, provides a more reliable score on which to make a verification acceptance/rejection decision. [4]

3.2 Methodology and Scoring

A common experimental environment for SV research has been developed on Unix platforms using the HTK V2.0 Hidden Markov Modelling Toolkit [3] and distributed to all the CAVE research partners. This has involved the codification of a wide variety of SV algorithms in such a way that they could be expressed under a common formalism and with a common notation. This method has allowed the investigation of a wide variety of approaches to verification. The reference system is comprised of Unix shell-scripts written in tcsh, and programs written in the C programming language. The common reference system, assessment procedures and speech databases have allowed for the immediate reproducibility of experiments across research partner sites, allowing for a significant improvement found at one site to be incorporated at all sites and efficiently used as a starting point for a new round of tests.

A basic set of scoring procedures was developed in order to standardise performance evaluation using so-called *dynamic* evaluation, in which setting rejection thresholds *a posteriori* allows the computation of a single Equal Error Rate (EER), the rate for which false acceptance and false rejection rates are equal. This EER is *gender-balanced*, meaning that any numerical bias towards male or female speakers in the test collection is not reflected in the result. While the EER can give an accurate diagnostic of global trends, it can also lack resolution; for example, an EER of 2.5% could either refer to the same level of error distributed over the entire population, or a rate of 50% error for 5% of the speakers. Of course, the luxury of *a posteriori* thresholding will not be available in a real service, and we are already working on the problem of how to estimate and adapt real-world SV rejection thresholds.

3.3 YOHO Experiments

The first test campaign was conducted on the YOHO corpus [1], still the only comprehensive freely-available SV data collection. It contains 138 American English speakers (106 male and 32 female), each of whom uttered 24 "combination-lock" number sequences (for example, "twenty-six, thirty-four, sixty-one") in 4 distinct enrolment sessions, and 40 other such sequences over 10 verification sessions. Utterances were collected in a quiet office environment via a telephone handset connected to a workstation and sampled at 8 kHz (downsampled from 32kHz). In order to obtain a telephone-bandwidth speech signal, the speech was filtered to a bandwidth of 300-3400Hz. The combination-lock strings were decomposed into 18 sub-word units, such as “For”, “five”, “ty”, and so on. (Only 56 of the numbers between 1 and 100 appear in the YOHO combination-lock phrases.)

The acoustic analysis used in the first set of experiments had the following characteristics:

- pre-emphasis of 0.97
- Hamming windowing, with window size 25.6 ms and shift 10 ms.
- Application of 20 Mel-scale triangular filters between 300 and 3400 Hz.
- Computation of cepstral coefficients $c_i(t)$ for $0 \leq i \leq 12$
- Subtraction of long-term cepstral mean vector
- Computation of delta coefficients $c_i'(t)$ from a window of 5 cepstral coefficients $c_i(t-2) \ldots c_i(t+2)$
- Computation of delta-delta coefficients $c_i''(t)$ as the deltas of 5 successive deltas.

A world model was built using the enrolment speech of 10 randomly-chosen speakers of each sex. These speakers were not used as “callers” in our subsequent SV experiments. This left 118 speakers, each with 40 combination-lock test utterances, allowing 4720 true verification attempts. For each speaker, 40 randomly-chosen impostor accesses were also made; 30 from a same-sex impostor, 10 from an opposite-sex impostor. Therefore, in each test of an SV method, 9440 accesses were carried out, each on a single combination-lock utterance.

YOHO experiments concentrated on the influence of the number of enrolment sessions (s) and the number of utterances per session (u) used in enrolment; also, the impact of the choice of the HMM topology, in terms of the number of model states per phoneme (p), and the number of Gaussian mixtures per state (q), was investigated. A strict left-right HMM topology, was adhered to throughout since this consistently gave best performance.

Figure 1 shows the gender-balanced speaker-independent (GBSI) EERs for different enrolment configurations. It can be seen from this figure that, for a given amount of enrolment data, it is better to have a larger number of shorter sessions rather than fewer sessions each with more data. It can be seen that doubling the amount of data from a single session halves the EER, but spreading the 24 utterances over four sessions causes an additional improvement by a factor of two. Likewise, enrolling with 12 utterances over four sessions does as well as 24 utterances from a single session. Also, where two results are given for each of the configurations 1x12 and 1x24, the better performance comes from the last-recorded enrolment session, and the poorer from the first. Unsurprisingly, the best performance is obtained when all the available data is used for enrolment.

Since CAVE is concerned with the application of SV technology to real services, rather than just the maximisation of experimental results with available corpora, the impact of this result on a real enrolment strategy must be considered. It would obviously be difficult and unwelcome to try to force prospective callers to an SV-secured service to undergo a lengthy and complicated enrolment procedure; the most appropriate strategy would appear to be to limit the number of enrolment sessions to 1 or 2, and to try to retrain the speaker models as soon as possible on speech data obtained during use of the service.

As regards the choice of the model topology, Figure 2 shows the impact of varying the values of p and q. It seems that the product pq governs the performance of the system, with an optimum value of pq of 4–5. For a constant product pq, configurations with fewer states perform consistently better for some fixed amount of enrolment data. When using only a quarter of the enrolment data (24 utterances), the performance differences of varying p, q, and pq-products was small within the ranges investigated.

The above comparisons were all done with an MFCC parameterisation. In the meantime, however, liftered linear predictive cepstral coefficients (LPCCs) have been shown to yield better performance than MFCCs. Our best EER so far on YOHO, using all the training material, was obtained with a system based on 16 weighted LPC-coefficients liftered with a sin-lifter of size 16, and with an HMM topology of 1 state per phoneme and 5 Gaussian mixtures per state. This system had a speaker-independent EER of 0.061%. At this level of performance, it is difficult to see any significant improvements with the given test set size.

3.4 SESP Experiments

Our second test campaign, which is still running at the time of writing, is based on the SESP database. SESP, which was collected by KPN Research, is a database of 46 native speakers of Dutch (24 male, 22 female) uttering 14-digit telephone calling-card numbers and PIN codes, in several sessions, over the telephone. SESP is a far more realistic database for CAVE experiments, since it contains authentic telephone speech, with data from different locations around the world, transmitted over differing telephone networks (including mobile) and with differing telephone handsets. The utterance of account numbers and PIN codes is unconstrained and exhibits a great deal of variation, such as, for example, whether numbers are spoken in terms of digit "groups" or one at a time. It was decided, for the time being, only to use, in our experiments, the 2973 utterances of account numbers in which digits were spoken one-at-a-time.

The enrolment set was defined, for each speaker, as consisting of four sessions with a quiet recording environment, with two distinct handsets being used, and correct pronunciation of the account number throughout. As for YOHO, varying subsets of this data were taken for experiments and the raw waveform data subjected to the identical acoustic parametrisation. However, it did not make sense to remove SESP speakers for the purposes of world modelling, due to the relatively low number of speakers; instead, a world model was trained from a gender-balanced 48-speaker subset of the Dutch *Polyphone* database. Tests were performed with 1658 true accesses, distributed as evenly as possible among speakers, and 1016 impostor attempts; once again, the ratio of same-sex to opposite-sex impostor attempts was 3:1. Each verification was performed on a single account number utterance.

Figure 3 shows that for 8 enrolment utterances over 4 sessions, the HMM topology does not have such a major impact on performance, with the EER mostly unaffected by differing values of p and q. This is a similar finding to the YOHO experiment with 24 utterances, i.e. with about about the same amount of enrolment data per HMM. If the enrolment data is further reduced, however, the HMM topology again becomes an important factor, as can be seen from Figure 4, which shows the EERs for different numbers of states per phoneme using a fixed pq-product of 4. With less enrolment data, the configurations with more states and fewer mixtures clearly outperform other configurations.

Again, our best performance so far, using 4 enrolment sessions with 2 utterances in each, on the SESP database is an EER of 1.27%. This has been obtained from a system with 2 states per phoneme and 2 Gaussians per state, using the same liftered LPCC parameterisation as decribed above. Also we have no results on this database to compare with, we feel that this is a very good result given the extremely realistic nature of the data collection, and we believe that this figure will drop further as the CAVE research campaign continues.

4 Conclusion

Providers of telephone-based services will soon have to deploy speaker verification to improve the speed, friendliness and functionality of their telephone services without compromising security. One of the goals of the European CAVE consortium is to meet this need, through a combination of high-level requirements analysis and system building, and algorithm-level basic research to deliver technologies suitable for incorporation into such systems. Experimental research performed on the SESP and YOHO databases is already state-of-the-art and the design, implementation and test of real-world systems which will allow the exploitation of this research has already begun.

5 Acknowledgements

CAVE is supported by Grant LE-1930 of the European Union Telematics Application Programme and by the Swiss Federal Office of Education and Science. Of the many "CAVErs" working on our SV research, the authors especially thank Cedric Jaboulet, Johan Koolwaaij, Johan Lindberg, and Jean-Benoît Pierrot. Cedric also gave helpful comments on early drafts, as did Beat Pfister and Prof. Hans-Peter Frei. Finally, Aran Lunzer provided invaluable help with LaTeX.

References

[1] J. Campbell, *Testing with the YOHO CD-ROM Voice Verification Corpus*, Proc. Int. Conf. Acoust., Speech., Sig. Proc. (ICASSP), 1995, pp. 341–344.

[2] Datamonitor, *European Telebanking Report*, 1994.

[3] Entropic Research Laboratory Inc., *HTK Hidden Markov Model Toolkit V2.0*, 1995. `http://www.entropic.com`.

[4] A. Rosenberg, J. DeLong, C.-H. Lee, B.-H. Juang and F. Soong, *The Use of Cohort Normalised Scores for Speaker Verification*, Proc. Int. Conf. Spoken Lang. Proc. (ICSLP), 1992, pp. 599–602.

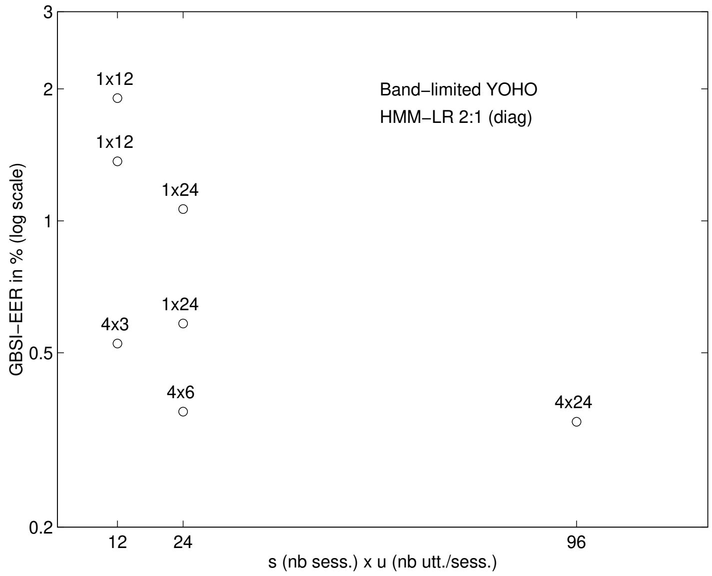

Fig. 1: Band-limited YOHO: influence of number of training sessions (s) and number of utterances per training session (u) on EER, plotted on log-scale.

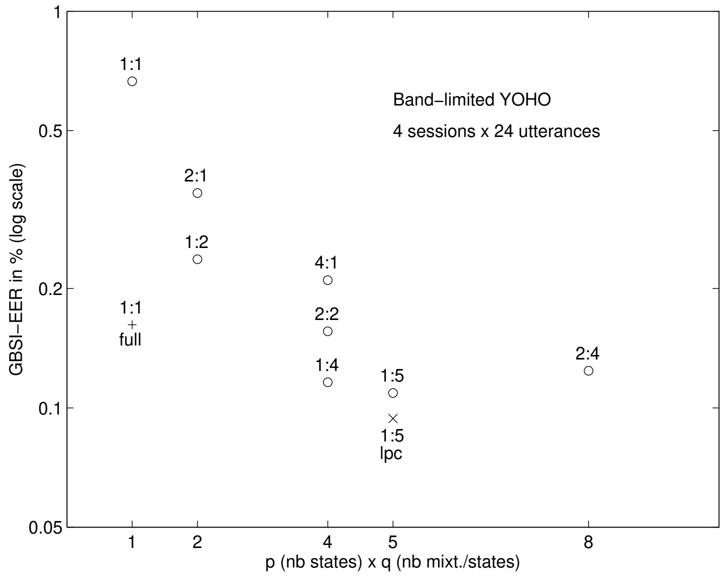

Fig. 2 Band-limited YOHO: influence of HMM topology (p states/phoneme, q Gaussians/state, diagonal covariance) on EER, plotted on log-scale. 4 sessions, 24 enrolment utterances per session.

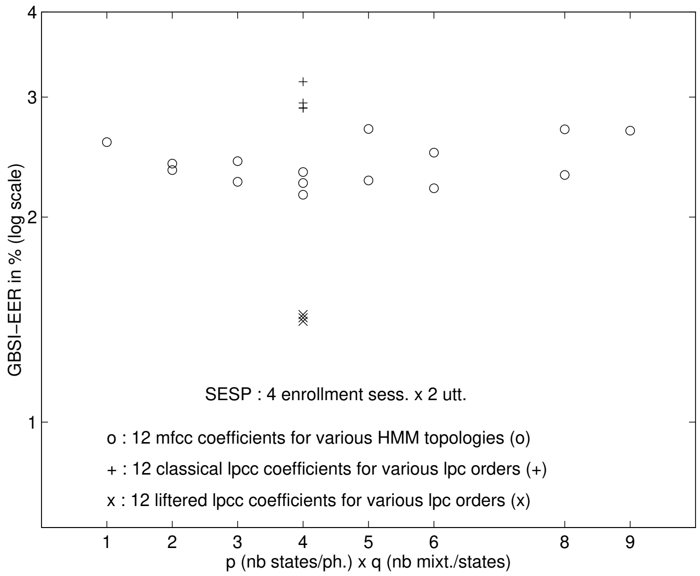

Fig. 3 SESP: influence of HMM topology (p states/phoneme, q Gaussians/state, diagonal covariance), on EER, plotted on log-scale. 4 sessions, 2 enrolment utterances per session. Also, improvement for $pq = 4$ with liftered LPCC coefficients.

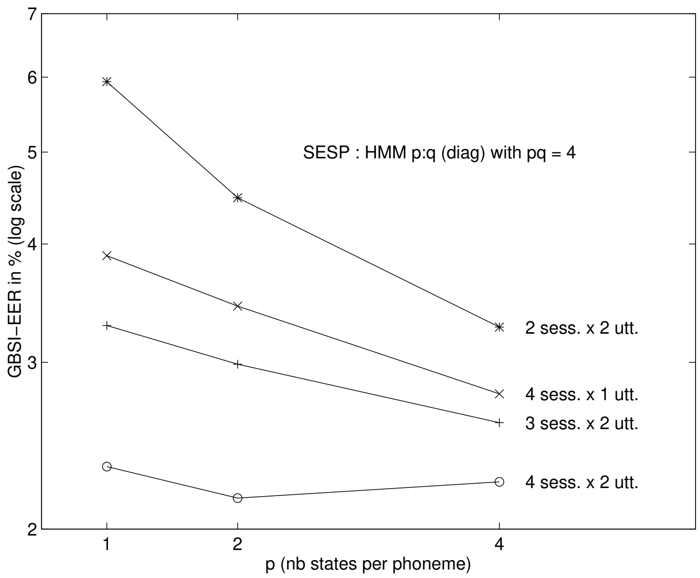

Fig. 4 SESP: Influence of p, number of states per phoneme for HMMs with $pq = 4$. For limited training material, $p > q$ is preferable.

This article was processed using the LaTeX macro package with LLNCS style