

CAVE - Speaker verification in bank and telecom services

Johan Lindberg, Mats Blomberg, Håkan Melin
Department of Speech, Music and Hearing, KTH, Stockholm

Abstract

The CAVE project aims at implementing and assessing two demonstration systems that are using speaker verification (SV) technology. In order to provide scientific grounds for the technological choices that are made for building these demonstrators a large number of available SV techniques have been investigated. The SV techniques have been evaluated using the YOHO and SESP databases and the results obtained are summarised in this paper. These experiments show promising results for the implementations to be made.

Introduction

CAVE, an acronym for caller verification, is a European project started in order to investigate and implement Speaker Verification (SV) into existing telecom services. The project is funded by the European Commission and runs two years starting from December-95. It addresses one of the key issues of telematics transaction services, viz. security and fraud prevention. In order to address this problem the consortium has joined three groups of organisations, namely technology providers, service providers and research labs. The project consist of 9 partners from 5 countries. The partners are shown in table 1.

Table 1. Partners participating in CAVE

PTT Telecom B.V.	The Netherlands
Catholic University of Nijmegen (KUN)	The Netherlands
Royal Institute of Technology (KTH)	Sweden
École National Supérieure des Télécommunications (ENST)	France
Union Bank of Switzerland (UBS)	Switzerland
Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP)	Switzerland
Vocalis Limited	United Kingdom
Telia Research AB	Sweden
Swiss PTT	Switzerland

The Cave partners have specified, and is currently implementing pilot versions of two SV protected telematics transaction services, one in the realm of telephone network operation, the other in the field of banking. These pilot applications will be used to test SV technology in realistic environments and to investigate human factors aspects of services protected by SV.

In order to provide the core technology for these systems, the academic partners of CAVE have been working on laboratory research in the area of speaker verification. This research has led to state-of-the-art performance on database evaluations on different SV approaches. The results from the laboratory-based work are feeding directly into the SV system being implemented for use in the demonstration systems.

Purpose of Demonstration Systems

Speaker Verification, by which a telephone caller's identity claim (currently represented by the combination of an account number and PIN-code) is verified, can be augmented with information derived from the sound of the caller's voice and could then allow for fast and secure authentication of callers.

As part of its work in investigating speaker verification, the CAVE project has carried out a major survey of the market for speaker verification in direct banking and calling card services. This survey indicates that there are many different methods by which SV could be deployed in existing and future telephone services.

In parallel with this survey, CAVE drew up the functional and technical specifications of two automatic demonstration systems, a telephone banking service and a calling card service. These systems are currently being built at the premises of Ubilab and KPN Research respectively. Each system will enter its initial, laboratory-test phase, with a relatively small number of users, at the beginning of 1997. After going through two further iterations, major field-tests of these systems with hundreds of callers will be carried out in the summer of 1997.

We are confident that the planned tests will give us a major new insight into the issues surrounding the deployment of speaker verification technology, such as how to handle the question of user enrolment into the SV system, and generally how best to trade off technological concerns against human-factors issues in order to provide the most secure and usable systems.

Research

The academic partners of the consortium are working on the development of reliable laboratory tests for SV technology, as well as on improvements of the technology as such. In order to do this the research partners have built a laboratory reference system for SV (Bimbot et al., 1997). This has been used for running a systematic test campaign, which have been achieving state-of-the-art performance on two different databases of speech recordings for SV. One is the widely used YOHO (Campbell, 1995) database and the other is a real life telephone speech database called SESP (Boogart et al., 1994).

Reference system

The CAVE generic speaker verification software package is based on HTK 2.0 (Entropic, 1995). Even though HTK is mainly intended for experiments on speech recognition using Hidden Markov Modelling (HMM), CAVE used the same platform to codify a variety of SV algorithms in such a way that they can be expressed under a common formalism and with a common notation. The system is comprised of Unix shell-scripts and c-programs.

The decisions of the SV system is based on likelihood normalisation (Rosenberg et al., 1992). In this normalisation process the verification relies on 3 models: the claimed speaker model, a speaker independent world model and a speaker independent silence model (inter-word model) which is shared by both client and world models during the verification.

When training the world, client and silence models, a word boundary segmentation of training sequences is needed. At test time the system makes its own segmentations.

For performance evaluation of the different SV algorithms a so called dynamic scoring technique was used. Setting of rejection thresholds is done *a posteriori* in order to allow the calculation of an Equal Error Rate (EER). EER is the rate for which false acceptance and false rejection rates are equal.

The use of this common system, assessment procedures and speech databases have allowed for immediate reproducibility of experiments across research partner sites, allowing for improvements to be incorporated at all sites and efficiently used as a starting point for new rounds of tests.

Experiments on YOHO

The first test campaign of CAVE was conducted on the YOHO corpus. It contains American-English speakers uttering combination lock sequences, i.e. phrases like “twenty-six, thirty-four, sixty-one”. The data was collected in a quiet office environment via a telephone handset connected to a workstation and sampled at 8 kHz. In the CAVE experiments this data was filtered to the telephone band-width of 300-3400 Hz. 20 speakers from the database were used in order to train the world model. These speakers were randomly selected and were not used as clients nor as impostors in the tests performed. This left 118 speakers, 22 female and 96 male, for use as clients and impostors.

YOHO experiments concentrated on the influence of the amount of enrolment data, thus studying the number of utterances per enrolment session as well as the number of sessions needed. The impact of the HMM topology in terms of number of states (p) per phoneme and number of Gaussian mixtures per state (q) was also investigated. A strict left-right HMM topology was found to consistently give the best performance (James et al., 1997).

Results

The best performance on YOHO was obtained with a left-right HMM with one state per phoneme and five mixtures per state. Using 16 weighted LPCC coefficients and energy, their deltas and delta-deltas, this approach reached a gender balanced sex independent equal error rate (GBSI-EER) of 0,028 %. This was done using all the enrolment material as defined in (Campbell, 1995) and the result competes well with other published tests performed on YOHO.

The experiments on YOHO also showed that it is better to have a larger number of shorter enrolment sessions rather than fewer sessions each with more data. Unsurprisingly, the best performance is obtained on as much enrolment data as possible.

As regards the choice of the model topology, it seems that the product of states (p) and number of mixtures (q) governs the performance with a maximum for pq around 4-5. For a constant product pq and a large amount of enrolment data, configurations with fewer states and more mixtures perform better than the opposite. However if the enrolment data is reduced then having more states and less mixtures will improve performance instead.

Experiments on SESP

The second test campaign was conducted on the SESP corpora. This database was collected by KPN research and contains native Dutch speakers uttering 14 digit telephone calling card numbers and PIN codes. The SESP data is far more realistic since it contains authentic speech collected over the telephone network from people calling from a variety of places. SESP contained only 46 speakers, 24 male and 22 female, thus it did not make sense to remove SESP speakers for the purposes of world modelling. Instead the world model was trained on a gender-balanced 48-speaker subset of the Dutch Polyphone database (Boogart et al., 1994).

The results obtained on YOHO were cross checked with SESP. Further experiments on different parameterisations were performed in order to find the optimal speech parameterisation for the SV task on telephone speech.

Results

The best performing experiment so far has been obtained using a left-right HMM with 2 states per phoneme and 3 mixtures per state. Using 16 weighted LPCC coefficients, energy, their deltas and acceleration coefficients, this method reached an GBSI-EER of 0,225 % which competes well with today's state-of-the-art techniques.

This was obtained on 8 enrolment utterances obtained over 4 sessions. If this is reduced to only 4 enrolment utterances obtained over 2 sessions the GBSI-EER is almost doubled to 0,436 %. Even though there are no other published results to compare with on this database we still feel that this is a very good result given the realistic nature of the database.

As with YOHO the SESP experiments showed that for large amount of training data the topology of the model did not have such a major impact on performance. Thus the EER was mostly unaffected by differing values for p and q with the same amount of enrolment data. However a slight gain in performance can be noted for topologies with more mixtures and less states when there is large amount of enrolment data available. If the enrolment data is reduced, however the HMM topology again becomes an important factor and configurations with more states and fewer mixtures outperform other configurations.

Conclusion

By use of SV technique in telecom services the speed, functionality and user friendliness of the systems can be improved while improving the security of these systems at the same time. Experimental results gained on YOHO and SESP are encouraging and test of real world systems has already begun.

Acknowledgement

CAVE is supported by grant LE-1930 of the European Union Telematics Application Programme and by the Swiss Federal Office of Education and Science. The authors would also like to thank the many "CAVErs" that have been working on the SV research within the project.

References

- Melin H. 1996. Gandalf - A Swedish Telephone Speaker Verification Database. *Proceedings of the International Conference on Spoken Language Processing, ICSLP-96 PA, USA*, pp 1954-1957.
- Campbell J. 1995. Testing with the YOHO CD-ROM Voice Verification Corpus. *Proceedings of Int. Conf Acoust. Speech. Sig. Proc. ICASSP 1995* pp 341-344
- Entropic Research Laboratory Inc. 1995. HTK - Hidden Markov Model Toolkit V2.0. <http://www.entropic.com>
- James D., Hutter H.-P., Bimbot F. 1997. The CAVE Speaker Verification Project - Experiments on the YOHO and SESP Corpora. *Proceedings of the first international conference on Audio- and Video-based Biometric Person Authentication (AVBPA)*. pp 385-394
- Bimbot F., Hutter H.-P., Jaboulet C., Koolwajj J., Lindberg J., and Pierrot J.-B. 1997. The CAVE Project: Caller Verification for Telephone Applications. *To be published in Eurospeech-97*.
- Rosenberg A., DeLong J., Lee C.-H., Juang B.-H. and Song F. 1992. The use of Cohort Normalised Scores for Speaker Verification, Proc. Int. Conf. Spoken Lang. Proc. (ICSLP). pp. 599-602.
- Boogart T.I., Bos L. and Boves L. 1994. Polyphone project overview. *Proceedings of IVTTA. Kyoto. Japan*