CAVE – Speaker Verification in Banking and Telecommunications

David James[†] Hans-Peter Hutter[†] Fréderic Bimbot[‡]

[†]Ubilab, Union Bank of Switzerland,
Bahnhofstr. 45, CH-8021 Zurich

[‡]Ecole Nationale Supérieure de Télécommunications,
46 Rue Barrault, 75634 Paris, France

e-mail: {David.James, Hans-Peter.Hutter}@ubs.com

Banks, amongst other businesses, will soon be under a great deal of pressure to improve the range and friendliness of their telephone-based services without compromising transaction security. Currently, automatic telephone services authenticate callers using a PIN-code and consequently do not offer a sufficient level of security for significant transactions; more sophisticated services, which depend on a human agent, demand lengthy authentication procedures. The technology of speaker verification has the potential to deliver fast and secure caller authentication for automated or agent-assisted telephone services. This paper gives an overview of the work of the CAVE consortium, a European-wide research project in speaker verification in which Ubilab is involved. Results obtained in experimental work so far compare favourably with the state of the art in speaker verification.

1. Introduction

It is clear that increasing numbers of businesses, and banks in particular, are becoming less reliant on traditional "bricks-and-mortar" infrastructure. Credit-card purchases and all manner of other financial transactions are currently being made by telephone, and the recent unparalleled growth of the Internet has now given rise to the first World-Wide Web-based financial services [Sfn96][Sch96]. The future range and depth of the services which will be offered over these channels, or those that evolve from them, is only hinted at by what is currently available. However, whereas the providers of these Internet services are careful to emphasise the level of cryptographic security with which transactions can be made, many telephone services are at the moment secured only by information which is relatively easy to snoop, such as a personal identification number (PIN) code, or, in more sophisticated cases, a password and the answers to various questions. Calling-card services, in which cardholders are able to make telephone calls and charge them to a personal account with the card-issuing company, are a particularly vulnerable form of telephone service; they are subject to billion-dollar levels of fraud, since it is a simple matter for a fraudster to "shoulder-surf" calling-card users at public telephones and obtain all the information necessary to charge calls to a user's account.

Telephone direct banking services have not yet fallen victim to fraud on such a scale; however, they may find themselves under attack from fraudsters as they become more prevalent and begin to offer a wider range of transactions to callers. A potential solution to this problem is telephone speaker verification -- the authentication, by the sound of the voice alone, of a caller's identity claim, in order to grant that caller access to sensitive data or services. The European Caller Verification (CAVE) consortium, of which Ubilab is a member, has been set up in response to the current threat to telephone calling-card services, and in anticipation of future threats to telephone banking services; it is currently researching and developing speaker verification technology, which it intends to deploy into field test systems.

This paper aims to give a brief overview of the underlying technology of speaker verification, the current market and applications for the technology, and the research which has been undertaken so far within the framework of the CAVE project. The subsequent sections of this paper address these areas in turn.

2. Speaker Verification

The human voice is a biometric – a physical or physiological characteristic of a person which distinguishes him or her, to a greater or lesser extent, from other people. The goal of speaker verification (SV), is to make a Yes/No decision in response to a speaker's initial claim of identity [Gis94]. It is therefore relatively constrained compared to speech recognition, in which the sequence of words that were spoken must be detected. However, SV is based on many of the same fundamental concepts as speech recognition; the parametrisation of the speech signal to extract appropriate coefficients for pattern matching, the training of acoustic models on appropriate spoken training data, and the subsequent matching of acoustic models against new, "unseen" speech data. In speech recognition, the output of this stage is the hypothesised sequence of words uttered; in speaker verification, it is, generally speaking, a measure of similarity, be it a Euclidean distance or some probabilistic measure, between the new speech and the speaker-specific model or models which corresponds to the claimed identity. This score is typically thresholded and the speaker's identity claim consequently accepted or rejected. There are therefore clearly two types of classification error which a verification system can make; the false rejection of a speaker making a honest identity claim, and the false acceptance of a speaker making a dishonest one. The performance of a laboratory-based SV system is usually measured by the single figure representing the point at which the rates of false acceptance and false rejection are equal; this single figure measure is termed the Equal Error Rate (EER). Of course, the actual point at which the trade-off between false acceptance and false rejection should be set in a real-world system need not necessarily be the point of equal error, but would have to take into account a number of factors, such as expected cost of false acceptance, customer tolerance of false rejection, and so on.

Although simple *template-matching* schemes can give reasonable verification performance for comparatively small computational effort [Dod85], and neural network techniques are starting to show promise [Sha95], the most flexible technology on which to build SV systems is arguably the Hidden Markov Model (HMM) [Rab93]. The HMM is a statistical model of the speech signal, and has achieved great success recently in speech recognition. An HMM consists of a set of states, each modelling some short period of speech, and a set of probabilities corresponding to the allowable transitions, over time, between states. Individual models can be built for acoustic units as small as the *phoneme* (the smallest unit of speech where substitution of one such unit for another leads to a change in meaning) and as large as an entire phrase. Phoneme-sized HMMs can be concatenated to form word models, so long as the pronunciation of a word is available, and word models concatenated together to form a network, to constrain the output of the recogniser only to valid sentences from some grammar. HMMs can model the variation with which a single speaker utters a particular sound or word, and also model such variation across multiple speakers (although speaker verification is clearly only concerned with the former). In addition, the choice of model *topology* - the number of states and the network of allowable interconnections between them - is highly flexible and can be chosen in advance according to the given task and the amount of available training data.

Differing levels of dependence on the lexical content of the speaker's utterance can be engineered into an SV system. At one end of the scale, SV can be made totally independent of the sequences of words uttered during enrolment - the period during which the potential user supplies utterances from which the model of his voice is to be trained, and access – when the system is in operation and the speaker wishes to be authorised. Such a text-independent SV system can be trained with and used on any utterance from a speaker; since it places no constraint on the utterance, it could be used to perform non-intrusive verification, for example to detect fraudulent use of a mobile phone where it might not be acceptable to undergo a separate authentication stage before dialling a number. In contrast, text-dependent SV specifies that not only must the speaker's voice match a stored voice model, but also that the speaker must utter some specific piece of personal information, such as a PIN-code or other password. A related form of text-dependent SV is referred to as text-prompted; here, the speaker does not know what he is required to say in order for authentication to take place. Instead, the required utterance (for example, a randomly generated short digit sequence) is used to prompt the speaker, who must then repeat the sequence. The primary advantage of text-prompted SV is that it is far less vulnerable to attack by the replay of recorded speech than either text-independent or password-based text-dependent verification.

A technique which was only recently applied to SV and is now commonplace is so-called cohort modelling; this refers to the normalisation of a verification score with respect to scores obtained for other speakers in the speaker collection [Ros92]. At its most basic, this normalisation removes the effect of non-speaker-related characteristics (for example, extraneous noise) from the verification process. Additionally, if, for each speaker X, cohort speakers are chosen for their similarity to X, then the process can also improve the discrimination of the speaker from these "soundalikes". In this case, the soundalike group is referred to as the cohort of speaker X. If the normalisation speakers are the same for all speakers, a single model can be trained from their speech; this model is known as a world model.

3. The application of Speaker Verification to Telephone Banking

Speaker Verification is lagging significantly behind speech recognition, in terms of deployment of the technology in wide-scale, publicly-accessible telephone services such as those offered by banks or other businesses. In America, where new technologies are typically adopted far earlier than in Europe, there has still only been one major application of SV technology, namely the Sprint Voice FONcard [Mar95]. This is a telephone calling card in which the caller's identity claim and the verification of that claim is made through one single utterance of the caller's social security number. No such services are available in Europe; instead, telephone services, such as UBS's "LibertyLine", use touch-tone sequences to authenticate callers and allow them to navigate through various options.

Even before the phenomenal growth of the World-Wide Web, it was estimated that direct banking, in the forms of telephone and videotex systems, would account for more than half the banking transactions of 30% of the population of Europe by the turn of the century [Dat95]. 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 far more popular than videotex banking, since the most demanding piece of equipment it requires in the home is a touch-tone telephone. However, the evolution of the Internet into a place where business can be done, and projections of the future prevalence of the low-cost domestic "Internet appliance", suggest that a large number of banks will use the Internet as a channel for the delivery of a wide range of banking services in the not-too-distant future. It therefore seems likely that banks will begin to offer telephone and Internet banking alongside each other, thereby giving the customer a broader range of options for accessing personal financial data and making transactions.

Since totally-automated telephone banking services rely on touch-tone entry of the customer number and PIN-code, the functionality of these services is limited, and they do not attempt to provide a comprehensive alternative to existing bank services which can be obtained by post or through branch visits. More advanced services, offering a broad range of transactions such as bill payments, standing orders, and other forms of cash transfer, must currently be offered via a call centre, in which a large number of human agents deal with customer enquiries round-the-clock. In these services, customer authentication is currently performed through the use of a "security handshake"; this involves the asking of a number of questions to which only the authorised caller should know the answers. For example, one of Europe's market-leading telephone banks, interviewed by CAVE, currently authorises customers by asking them to state 2 letters (chosen at random) from a secret password, plus any one of the following pieces of information:

- Date of Birth
- Mother's maiden name
- Place of Birth
- "Special" date (not birthday)
- "Special" address (but not the customer's own)

This method was reported to work well but was also felt to be inconvenient and time-consuming. Since the password is never uttered in its entirety, it is not possible for a casual caller to overhear it, although it would eventually be possible to work it out if a number of these authentications were overheard.

The case for SV in future direct banking services is, in our view, quite powerful. Since authorisation for the use of Internet banking services will be a swift and secure process involving only the typing of a password, existing strong authentication methods for telephone services (as opposed to the weak methods exemplified by PIN-codes) will be judged unacceptably complicated and time-consuming. Moreover, advanced telephone banking services must still authenticate major transactions by confirming them in writing; this would clearly be unthinkable in the world of Internet banking!

Therefore, if customer expectation of telephone-based services increases, PIN codes will cease to be a sufficient authentication procedure for these services. Moreover, the PIN code is already proving itself to be an insufficient security measure for some existing telephone services. As more and more companies and organisations issue cards, there is a greater burden on the cardholder to memorise randomly-selected digit strings corresponding to differing cards from various issuers; moreover, if it is possible for the cardholder to change their PIN-codes, the customer can be tempted to adopt the same PIN-code for more than one card. PIN-codes remain popular with card-issuers, since they are seen as a method of "out-sourcing" the security problem to the cardholder; the small print signed by new cardholders ensures that any breach of PIN-code security is legally the fault, and consequently the responsibility, of the cardholder. However, much of the US calling-card fraud reported annually can be attributed to the use of sophisticated techniques, such as video equipment to spy on unwitting cardholders using payphones in public areas. With such activity threatening to invalidate the "PIN-code assumption", it seems appropriate to see how current security measures may be reinforced. Speaker verification would be one such reinforcement method.

The two ways in which SV could be applied can be referred to as the "gatekeeper" or the "alarm-bell", each one applicable to fully-automatic and agent-assisted services. In the gatekeeper method, the caller must initially be challenged to prove his identity in order to gain access to the service. The appropriate SV technologies here are text-dependent (for example the utterance of a PIN-code or a password, although this may be too easily snooped) and text-prompted (in this case, overhearing or recording the utterance is of no use since this utterance will never be used to access the service again). In alarm-bell mode, the caller can be weakly authenticated at the beginning of the call and continuously verified throughout a transaction. When a major transaction is finally processed, the SV system must be sufficiently confident that the caller is who he claims to be. The strength of this approach would be that the result of the verification could be used, along with various auditing information such as a record of recent transactions, to generated a combined decision on whether to carry out the transaction. Too low a final authentication score would trigger the "alarm"; at this point, a further attempt to verify the caller could be made. So-called "audit trail" information is already used to root out fraudulent use of companies' private telephone exchanges or mobile telephones; tell-tale signs of fraud include, in the PBX case, long overseas calls outside office hours, or the apparent use of a mobile telephone from two different parts of a country at the same time.

4. 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 PTT Telecom NL) 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 Telecommunications, 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 the 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.

Research so far has involved the YOHO corpus [Cam95], since it is currently the only comprehensive and freely available SV data collection. It contains enrolment and verification sessions for each of 138 speakers (106 male and 32 female). Utterances were collected in a quiet office environment via a telephone handset connected to a workstation and sampled at 8 kHz, the standard sample rate for telephone speech. For each speaker, there are 4 enrolment sessions of 24 utterances each, and 10 verification sessions of four utterances each. Each utterance is a so-called "combination lock" phrase consisting of three two-digit numbers (for example, "twenty-six, thirty-four, sixty-one"). No phrase is common to the enrolment and verification sessions for any speaker, and the experiments described in this paper can consequently be thought of as text-prompted.

State-of-the-art performance on the YOHO corpus has recently reached an Equal Error Rate (EER) of less than 0.5%. Colombi *et al* [Col96] achieved an EER of 0.282% with an

HMM-based verification system by optimising the selection of cohort speakers against whom the verification score was normalised. Enrolment was performed on 24 of the combination-lock triplets (a quarter of the available enrolment data) and tests performed using groups of 4 triplets. Setlur *et al* [Set96] performed experiments only on the male speakers and, by combining standard speaker-dependent modelling with some speaker-dependent data derived from speaker-independent speech models, achieved an EER of 0.255%. Although their speaker-specific models were trained on 96 combination-lock phrases – 100% of the available enrolment data[1], verification was only performed on individual triplets. Use of the maximum possible amount of data for both enrolment and verification unsurprisingly reduces the error rate. Che *et al* [Che96] obtained a 0.09% false acceptance rate at a false rejection rate of 0% by using all 96 available triplets in enrolment and a group of 4 triplets for each verification attempt. This result suggests an EER of somewhere between these two rates.

An experimental environment for SV research has been developed on Unix platforms using the HTK V2.0 Hidden Markov Modelling toolkit [Ent95]. This environment has allowed the investigation of a wide variety of text-dependent and text-independent approaches to verification. The first stage of CAVE project work using this platform has been aimed at calibrating verification performance between research sites, and has also allowed an initial investigation of optimal HMM topologies for SV. So far, simple "world modelling" has been investigated; out of the 106 male and 32 female speakers, 10 speakers of each sex were chosen and their utterances used to create a single world model. These speakers were not subsequently used in the initial SV tests. Varying amounts of enrolment speech were used to train speaker-specific models, ranging from just the data from a single enrolment session of that speaker, to all the data from all four sessions. The particular form of model depended on the SV approach taken; in text-independent mode, a single model was trained for each speaker using many differing utterances; in text-prompted mode, a set of 17 speaker-specific sub-word models, as illustrated in Table 1, was required, to allow verification on arbitrary combination-lock strings.

one	two	three	four	five	six
seven	nine	ty	Twen	Thir	Four
Fif	Six	Seven	Eigh	Nine	

Table 1: Sub-word recognition units in text-prompted SV experiments on YOHO

The speech waveform data supplied in the YOHO collection had to be parametrised to extract appropriate coefficients for model training. Each waveform file was blocked into 100 overlapping frames per second, each frame of length 25.6 ms. Each frame was parametrised to yield a vector of 12 *mel-frequency cepstral* (MFC) coefficients. The use of MFC coefficients is common in speech modelling; amongst other reasons, the coefficients within each frame are relatively decorrelated, which reduces the need to reflect such correlations, or *covariances*, in the HMM. "Differential" coefficients, reflecting the rate of change of each MFC, were also added to the vectors; this is a common technique which compensates, at

[1] Although YOHO is divided into enrolment and verification subsections, there are no standards for the proportion of enrolment data used and the exact nature of testing – such as whether speakers are excluded from the test to be used in cohort modelling, or the relative number of "impostor attempts" made compared to the number of true accesses. This must be borne in mind when comparing results.

least to some extent, for the in-built HMM assumption that successive speech vectors are statistically independent. In all experiments described here, individual models were trained using the widely-used Baum-Welch estimation algorithm on which the HTK training tools are based. Differing SV experiments were performed by changing a number of model parameters, such as the model topology, the number of model states allocated to each phoneme and the number of statistical Gaussian distributions modelling the acoustic parameters in each state. In addition, the amount of enrolment data used to train speaker models was also varied.

Verification tests were performed using a speaker-specific model, X, corresponding to a claimed identity, a world model W, and a silence model S. The silence model was trained on a random sample of silent portions of enrolment data across all YOHO speakers, and used in verification to take optional inter-word pauses into account. For an unseen utterance Y, which may or may not have been spoken by speaker X, the HTK recogniser was used to force an optimal alignment between the speaker-specific model X and the utterance Y. This alignment yielded a probabilistic score, denoted as $\Lambda(Y|X \cup S)$, measuring the closeness of the fit between the model and the unseen utterance. Additionally, an alignment score $\Lambda(Y|W \cup S)$ was generated using the world model W. The two scores were combined to generate a so-called *log likelihood ratio*

$$LR_{X,W,S}(Y) = \log\left[\frac{\Lambda(Y|X \cup S)}{\Lambda(Y|W \cup S)}\right]$$

For each speaker, the acceptance/rejection threshold $\Theta(X)$, to which the likelihood ratio score is compared, was set *a posteriori*; this allowed laboratory EER figures to be calculated.

40 honest access attempts were simulated for each speaker, by performing verification individually on each of the 40 available combination-lock digit strings. Additionally, impostor break-in was simulated, for each speaker X, by selecting 40 impostor utterances, 30 from speakers of the same gender as X and 10 from the opposite gender, and trying to verify the speech of each impostor using the model of speaker X. A total of 9440 (that is, 118*(40+40)) separate access attempts were therefore performed for each HMM topology tested. A subset of the results obtained in initial experimentation are shown in Table 2. The "S/P" and "G/S" columns give the number of model states allocated per phoneme, and the number of Gaussian distributions per state, respectively; also, where multiple enrolment sessions were used to train speaker-specific models, training utterances were taken equally from each session. All results given here are for the so-called “left-right” HMM model topology, in which a transition is only allowed from a model state to itself or to its immediate successor; performance here was consistently better than for more flexible model topologies.

Covariance Modelling	S/P	G/S	Number of Training Utts	Number of Sessions	Equal Error Rate
No	2	1	12	1	4.58
No	2	1	24	1	3.15
No	2	1	12	4	2.63
No	2	1	24	4	2.24
No	2	2	96	4	0.82
Yes	1	1	96	4	0.36

Table 2: Results on varying HMM topologies and parameters in CAVE SV experiments.

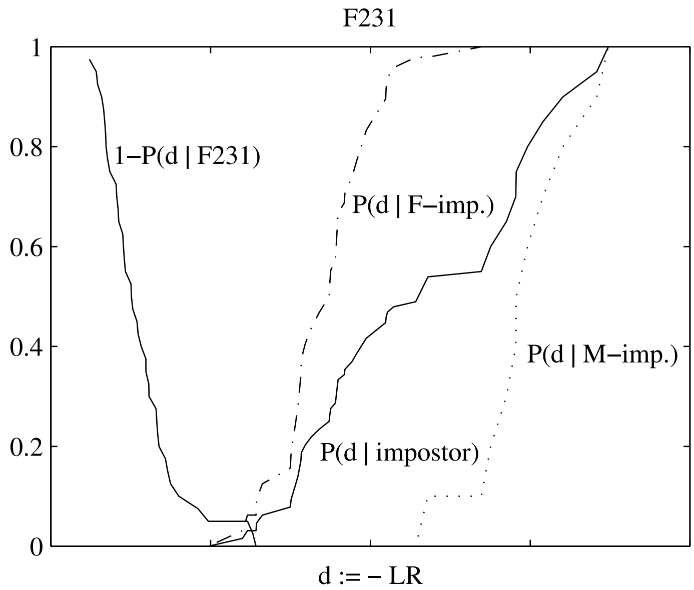

Figure 1: SV performance for YOHO speaker F231.

As can be seen from Table 2, the best result obtained was an EER of 0.36%. Equal Error Rates here have been calculated in a gender-balanced manner to correct the imbalance between the numbers of male and female speakers in the YOHO collection.

The best performance has been achieved where the covariances between MFC coefficients have been modelled, despite the relative decorrelation of these coefficients. For methods without covariance modelling, it is unsurprising to note that the use of more data improves verification performance; in addition, it can be seen that training with data from a number of enrolment sessions leads to a significant reduction in error rate, from 4.58% to 2.63% for the EER between rows 1 and 3 of Table 2.

Recently, a new set of results has been obtained with the latest release of the HTK front-end analysis software; although they have not yet been ratified across all CAVE research sites, they do point to a major improvement in performance through the use of an increased number of Gaussian distributions per state. In particular, an EER of 0.124% has been achieved using 4 Gaussians per state, and without the need for full covariance modelling. This figure compares favourably with the state of the art in speaker verification on YOHO.

Figure 1 shows, for YOHO female speaker F231, the distributions of the SV distances, d, from the HMM verification system, for all the honest and dishonest accesses attempted for this speaker. The inversion of the "honest access" curve, $P(d|F231)$, allows the point of equal error to be shown, as the intersection of $1-P(d|F231)$ with $P(d|impostor)$. $\Theta(F231)$, the acceptance-rejection threshold, is therefore set to the value of d at this point of equal error. It can be seen from the intersection of the honest access curve with the curve $P(d|F-impostor)$ that consideration only of the female impostors would lead to a higher point of equal error and a different value of Θ. Correspondingly, the EER with male impostors only is zero.

5. Future Challenges

The CAVE consortium faces a number of future challenges, technical and implementational, in reaching its goal of demonstrating speaker verification in field tests corresponding to real services. One of the most important is telephone variability. Since the YOHO collection was recorded from a standard telephone handset connected directly to a workstation, it does not feature any of the acoustic variability associated with the use of differing telephones, or with the transmission of the speech signals over differing networks (most notably the mobile network). A new round of laboratory experiments is currently beginning using a Dutch SV database which exhibits a great deal of channel and handset variability.

Outside the laboratory, a partial solution to the channel variability problem is to specify that real users of a speech-enabled telephone service should enrol into the service a number of times using the telephones they anticipate using to access the service in the majority of cases. Although in the case of a banking service, it is fair to assume that home or work telephones will be used to make the majority of accesses, such an assumption is not possible in a calling-card application, where it can be more or less assumed that the caller is definitely not using his own telephone. The enrolment question is an important one since it is clearly not enough to achieve good verification performance regardless of how much enrolment data is required; instead, the enrolment strategy must be carefully selected to ensure acceptable performance without placing too great a burden on the customer. What actually constitutes acceptable performance is, of course, a further question, which can only be investigated in field tests.

6. Conclusions

Banks may soon have to deploy technologies like 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 anticipate 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 within the project framework is already state-of-the-art and the design and implementation of field-test systems which will allow the exploitation of this research has already begun.

Acknowledgements

The CAVE Project is supported in the European Union by Grant LE-1930 in the Telematics Application Programme and in Switzerland by the Office Fédéral de l'Education et de la Science (Bundesamt für Bildung und Wissenschaft). The authors wish to acknowledge the help of all the CAVE partners in performing the experiments described in this paper. We also wish to thank Cedric Jaboulet, Beat Pfister and Prof. H.-P. Frei for their very helpful comments on earlier versions.

References

[Cam95] Campbell JP: *Testing with the YOHO CD-ROM voice verification corpus*, Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc., 1995, pp. 341-344.

[Che96] Che C, Lin Q, Yuk DS: *An HMM Approach To Text-Prompted Speaker Verification*, Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc., 1996, pp. 673-676.

[Col96] Colombi JM, Ruck DW, Rogers SK, Oxley M, Anderson TR: *Cohort Selection and Word Grammar Effects for Speaker Recognition*, Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc, 1996, pp. 85-88.

[Dat94] Datamonitor: *European Telebanking Report* 1994; *Case Studies in European Direct Banking*, 1995.

[Dod85] Doddington GR: *Speaker Recognition – Identifying People by their Voices*, Proc. IEEE, Vol. 73, No. 11, November 1985.

[Ent95] Entropic Research Laboratory Inc.: *HTK Hidden Markov Model Toolkit V2.0*, 1995. http://www.entropic.com.

[Gis94] Gish H, Schmidt M: *Text-Independent Speaker Identification*, IEEE Signal Processing Magazine, October 1994, pp. 18-32.

[Mar95] Markowitz JA: *Voice ID: Applications and Sources for Speaker Recognition*, TMA Associates, 1995.

[Rab93] Rabiner LR, Juang BH: *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[Ros92] Rosenberg AE, DeLong J, Lee CH, Juang BH, Soong FK: *The Use of Cohort Normalized Scores for Speaker Verification*, Proc. Int. Conf. Spoken. Lang. Proc (ICSLP), 1992, pp. 599-602.

[Sch96] Schwab Online Trading: http://www.schwab.com.

[Set96] Setlur AR, Sukkar RA, Gandhi MB: *Speaker Verification using Mixture Likelihood Profiles Extracted From Speaker Independent Hidden Markov Models*, Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc, 1996, pp. 109-112.

[Sfn94] Security First Network Bank: http://www.sfnb.com.

[Sha95] Sharma M, Mammone R: *Subword-based Text-Dependent Speaker Verification System with User-Selectable Passwords*, Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc, 1995, pp.93-96.