SPEAKER VERIFICATION IN WWW APPLICATIONS

L. Boves & J.W. Koolwaaij

Department of Language and Speech, Nijmegen University
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
E-mail: {boves,koolwaaij}@let.kun.nl

RÉSUMÉ

Dans cet article, nous décrivons un site WWW qui utilise la reconnaissance du locuteur pour le contrôle d'accès. Le site a été installé pour acquérir une expérience réelle de l'application de reconnaissance du locuteur à travers Internet. Nous avons voulu obtenir des informations sur les aspects techniques, procéduraux et facteurs humains de la reconnaissance du locuteur sur le Web.

En ce qui concerne les aspects techniques, il s'avère que la qualité des microphones et des systèmes similaires des postes de travail PC sont le principal souci. Beaucoup de problèmes procéduraux restent à résoudre, la plupart provenant du fait que le WWW n'a pas été conu pour un traffic birectionnel. D'ailleurs les capacités graphiques du web ne sont pas prévues pour la complexité nécessaire à la transmission de signaux de parole. Il s'est avéré que la combinaison des problèmes techniques et procéduraux a placé assez haute la barrière pour le déploiement de la vérification du locuteur dans les applications multimédia sur le Web.

ABSTRACT

In this paper we describe a WWW site that uses speaker recognition for access control. The site was set up to acquire real-world experience with the application of speaker recognition technology in the Internet. We wanted to obtain information about technical, procedural and human factors aspects of speaker recognition on the Web.

With respect to technical issues it appears that the quality of the microphones and the analogue subsystems in PC workstations are the major concern. Many procedural issues remain to be solved, most of which are connected to the fact that WWW was not designed for two-way traffic. Moreover, the graphics based Web did not really anticipate the complexities involved in the transmission of speech signals. It appeared that the combined technical and procedural problems constitute a relatively high threshold for the deployment of speaker verification in multi media WWW applications.

1. INTRODUCTION

Speaker Recognition (SR) has been studied for use in protecting physical access to premises, electronic access to data and information and to monitor the whereabouts of persons who are not allowed to travel freely. A comparison of these applications shows that each makes specific requirements to the technology and especially to the way in which the underlying recognition algorithms must be tuned and integrated in the overall hardware and software architecture.

Recently, interest in yet another application domain has grown. Increasingly, service providers are interested in using biometric verification to protect access to Internet and World Wide Web services; thanks to the increasing availability of multi media workstations, SR is the option of choice, if only because the performance of face recognition still leaves very much to desire. In this paper we investigate the ways in which Web-based applications make specific requirements that have an impact on the integration, implementation and the performance of the SR technology. We will address technological, procedural and human factors issues.

The paper is based on our experiences with access to a Web site created by the Department of Language & Speech. This site does not contain confidential information, and visitors do not receive special rewards when they access the site, as true customers or as impostors. Thus, this paper does not aim to describe a formal test of the level of security that can be attained with the help of specific SR technology. Rather, we want to derive guidelines on how SV can be used in Internet services that require some degree of protection.

2. DESCRIPTION OF THE SITE

The Web site under discussion [5] offers two access modes, using either Speaker *Verification* or Speaker *Identification*. Both modes require that a customer must first be enrolled. To this end a prospective customer must fill in a form with his/her name, age and gender, and e-mail this to the Web server. Each applicant then receives an automatic reply containing a 14 digit private access number, modeled after the card numbers used in the *scope* calling card service of PTT Telecom, together with an enrollment access code, to ensure the correct connection between the applicant's identity and his/her enrollment speech. To enroll the applicant has to speak the 14 digit access number eight times, creating a separate speech file for each utterance. These files must then be transmitted to the Web server, where the speech is used to create speaker models for each digit in the access number. After creating the

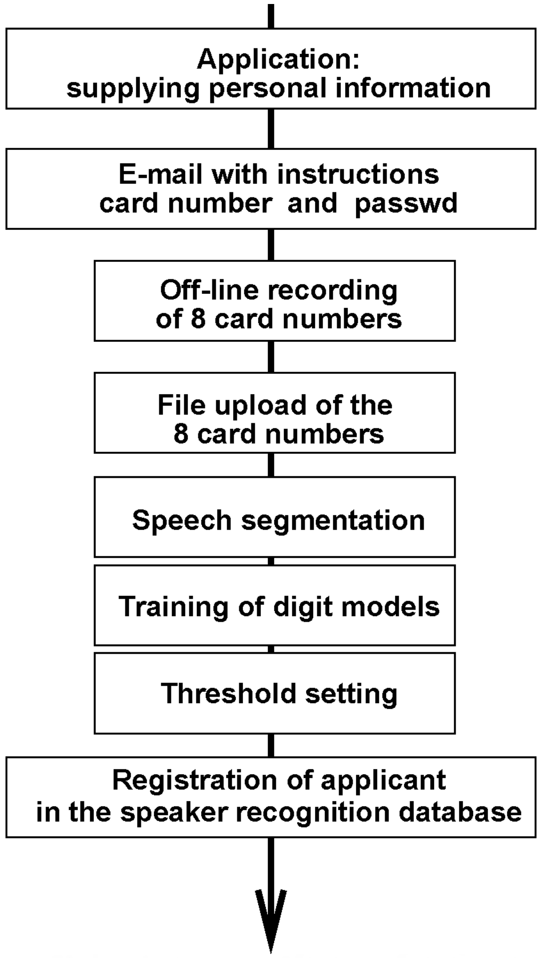

Figure 1. The enrollment procedure

digit models, a speaker dependent threshold is estimated and the applicant is registered in the speaker recognition database. This process is schematically depicted in Fig. 1.

2.1. Speaker Verification Access

The first access mode implements speaker verification (SV). On accessing the site, the applicant submits his/her name (see Fig. 2), from which the personal access number can be retrieved by the Web server. By displaying this number on the customer's screen the system then prompts the customer to speak the card number. (This is obviously not the way it will be implemented in real Internet services. Normally a customer would identify himself by speaking his name or personal access number, but the text prompted approach that displays the access number to every person who attempts to get access is used to enable imposter attempts.) Next, the access number must be recorded, and the speech file sent to the Web server, where the speaker's identity is checked.

2.2. Speaker Identification

The second access mode implements speaker identification (SI). In this mode the applicant is prompted to say a random fourteen digit access number. The speech is sent to the server, and matched against the models of all clients who are enrolled in the service. If none of the matches exceeds a minimal threshold, the applicant is rejected as an unknown speaker. Otherwise, the applicant is accepted as the speaker whose voice pattern matched best. This is the so called open set identification.

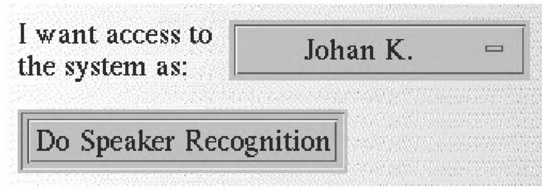

Figure 2. Example of the verification mode

2.3. The Web site's functionality

There is no real Internet service to be accessed via the demo. After having been admitted, the applicants are only asked to fill in a fixed questionnaire:

- Was the decision of the system correct?
- Did you pronounce the correct code and with your normal voice?
- Are you a native Dutch speaker?
- What is your gender and age?

The results of these questionnaires are used to better understand the conditions in which the SR system makes an erroneous decision.

Since the World Wide Web was initially designed to be able to download information (text, graphics, sound) from the Web server to the client, the procedures and protocols for uploading files from a client to a server have obtained much less thought and attention. Consequently, going the other way round (i.e., sending large amounts of information from the client to the server) is far from trivial. Most applications that require heavy two way traffic rely on proprietary protocols, that are shared by client and server. For our Web site a different approach was needed, because we could not impose a specific protocol on all interested visitors of the site. Therefore, we organized submission of the speech files to the Web server by means of form based file upload as described in Request For Comments (RFC) 1867 [6]. This implies that the speech is first recorded and saved on the client side of the Internet connection. Submitting speech in this way requires the applicant to perform several explicit actions, but for the time being it is the only way to cover a wide range of popular platforms (e.g. Unix, Windows and Mac) and Web browsers (e.g. Netscape and Internet Explorer). As long as one can stay within a single hardware or software platform more elegant solutions for uploading speech from client to server are available, but none of these works reliably across platforms (for instance, some procedures work under Windows NT, but not under Unix).

3. ALGORITHMS

Speaker models in this application are HMMs; separate client models are trained for all digits that occur in the client's card number. The HMM topology used is a left-to-right HMM, with two states per phoneme, two mixtures per state and a diagonal covariance matrix. Acoustic features are 12 liftered zero-mean cepstra (LPC based) together with the log energy and their first and second time derivatives. In addition to the client models there is a single set of gender independent world models, one for each of the ten digits. Finally, there is a silence model (or maybe better: non-speech model, since this model also captures background noise, mouth noise, etc.), that is shared by all clients and the world. For more details see [1].

3.1. A priori threshold estimation

As in any operational service accept/reject thresholds must be estimated from the enrollment data, possibly combined with the results of imposter attempts using previously recorded speech. In the Web application an interpolation technique is used to convert biased false reject thresholds estimated from enrollment speech to values which are more appropriate. This technique is described in another paper submitted for presentation in the workshop [2].

4. WWW ISSUES

For the time being, and for quite some time to come, the Internet is being used in other ways and in other circumstances than the Plain Old Telephone network. These differences have an impact on the way in which SR is best integrated in the overall application, and on the type and quality of the speech signals that are available for enrollment and decision making. However, Internet use, as well as the PC workstations, are evolving very rapidly. Therefore, the details of the picture sketched in this paper will probably change in the future.

4.1. Personal Workstations

Most Internet subscribers access services from a single personal workstation. This workstation can be in the office, or at home. A much smaller number of users access Internet services both from the office and from their private homes. An even smaller set of users access Internet services from many different locations. However, when people are on the move, they will often use remote log-in to access the Internet server 'at home'.

One of the advantages of WWW applications is that information identifying the workstation (or at least the Internet provider) from which access is attempted is almost always available. This is not yet the case with Calling Line Identification in the landline telephone network. Also, CLI information is not usually available from roaming handsets in the cellular GSM networks. The knowledge that most Internet users almost always access services from the same workstation (or the same Internet provider) certainly helps in increasing security.

4.2. Enrollment

One of the sticky problems for every biometric recognition technique is how to verify the identity of the applicant during enrollment. This problem is only worse in Internet services, where enrollment is essentially unsupervised. The possibility to identify the workstation helps somewhat, but it is certainly not enough to cater for high risk applications. Moreover, the remote host will not always be able to identify the workstation. For instance, if the customer is connected via the Serial Line Internet Protocol (SLIP), only the identity of the Internet service provider can be established, since the remote host changes each time the customer connects to the internet. Another possibility would be to set a cookie. (Cookies are a general mechanism which server side connections, such as CGI scripts, can use to both store and retrieve information on the client side of the connection.) This is more secure, because the cookie really identifies the workstation, but it still says nothing about the person who is using that workstation. An additional problem is that some people do not allow the Web server to set a cookie, but one might argue that accepting a one time cookie from a server that is trusted as secure enough for the application should be an acceptable prerequisite for obtaining access to a protected service.

Anyhow, for the 'empty' service in our experiment we could not warrant labourious verification of the identity of the users. Therefore, we request that the applicants fill in a form with name, age, gender and e-mail address. The applicant then receives an e-mail with among others the enrollment access code in his/her personal e-mail box. This access code code must be submitted together with the applicant's name and his/her enrollment speech and is used as identity verification during enrollment. Assuming that the applicant is cooperative and keeps the code secret, this method is almost 100% secure.

4.3. WWW vs Telephone

Compared to the Plain Old Telephone WWW has both advantages and disadvantages with respect to speaker recognition. WWW beats the telephone when it comes to multimodal interaction. Access to Web sites naturally involves graphical and textual prompts, which can easily be combined with typing or point-and-click actions. This eliminates the need for complex sequentially presented menu structures found in the interface to many telephone services. This advantage of the Web is offset by the much wider availability of telephone handsets. Moreover, and perhaps somewhat surprisingly, the signal-to-noise ratios and the distortion levels in telephone signals appear to be better than in speech recorded on the majority of multi-media workstations. The quality of the microphones which come with these workstations appears to vary dramatically, as does the quality of the sound boards they are connected to. Moreover, mouth to microphone distance varies over a much larger range in the case of workstations than with telephone handsets. When this distance is large, a low energy speech signal is recorded (with the attendant low signal-to-noise ratio), but when the distance is too small the speech signal can become highly peak clipped (which is equivalent to high levels of non-linear distortion). Last but not least, the analogue subsystems must work in extremely adverse environments, with high frequency radio noise emitted by the digital chips, combined with considerable dips in the DC power supply due to disk head movements.

From listening to the sound files for enrollment and access it was obvious that many files are very soft, using only a small part of the total amplitude range, while others are heavily clipped, causing considerable non-linear distortion. Also, background noise (speech and non-speech) is at least as bad as it is in telephone speech databases recorded in operational services. It shows that telephones have had a century long history of optimization for speech transmission, which the multi media workstations still lack.

Telephone speech transmission adheres to strict standards for coding and transmission protocols (even if different continents may follow different standards, like A-law and μ-law coding in Europe and the US). Despite SUN's promise to provide a completely platform independent speech API under JAVA, the actual situation shows a myriad of different and incompatible interfaces and protocols. Even if the speech is recorded locally and then transmitted as a file (which is quite clumsy) files may arrive at the server in a number of different formats, some of which use a very poor quality 8-bit pcm quantization.

5. RESULTS

The site was announced six months prior to submission of this paper. In six months time it has received 2447 visits, by 1235 different persons. 223 persons asked for enrollment information (about 10% of them are female and the average age is 30 years) and 59 persons actually completed the enrollment procedure. So 18% of the visitors is interested in enrollment but only 4% really completes enrollment. We think that the ratio between visitors and 'customers' is mainly due to the technical problems that many interested persons experience. The majority of the Internet population can only play sound files and not record them, because they lack appropriate hardware or software. Of course, due attention must be paid to this type of trouble when launching a WWW site that relies on speech input.

The fact that there is no real service or reward to be gained by enrolling and using the Web site has certainly contributed to the relatively small proportion of interested visitors who have taken the trouble to enroll. All this makes it understandable that only 4% of the visitors want to enroll themselves, only because they are interested in new technology.

We have sent a short and simple questionnaire to 86 people who expressed interest in enrollment, but never got round to actually doing it, asking why they did not complete the enrollment. We received 35 answers: 21 people said they lacked time, 8 experienced problems while trying, 6 lacked the necessary hardware (microphone or sufficiently fast modem) and nobody was having problems with the necessary software. So it appears that it is the lack of time (and maybe real interest) which is the most important reason for not realizing the initial intentions. But the proportion of persons who were stopped by technical problems is far from negligible, the more so if one takes into account that only persons who have a vested interest in new technology take the trouble to request the information necessary for enrollment. Unless hardware, software and communication procedures can be simplified substantially, one is likely to find that speaker recognition in Web services will have a very hard time catching on. Of course, the same will hold for applications that do speech recognition on a central server, instead of in the workstation (after which an ASCII coded message can be formed and transmitted over the Web using well understood technology, procedures and protocols). In any case, the problems encountered in transmitting raw speech data over the Web are an argument in favour of distributed processing, where speaker verification might be done locally on the workstation (perhaps even using speaker templates that could be downloaded from a central server that maintains the speaker recognition database).

Up to now 247 identification and 251 verification attempts have been recorded. For verification a False Reject Rate of 3.9% and a False Accept Rate of 0.3% has been observed (which depends on threshold setting), while the Identification Error Rate for the much more difficult identification task is 7%. Of course this latter percentage is heavily dependent on the size of the population of enrolled speakers [4]. This population size varied from one to about 40 during the six months that the demo is on line.

Despite the problems with the audio quality explained above, these results are comparable to the error rates obtained with a single enrollment session (also asking customers to repeat their card number 8 times) in a telephone calling card service. [3]

REFERENCES

[1] F. Bimbot, H.-P. Hutter, C. Jaboulet, J.W. Koolwaaij, J. Lindberg, J.-B. Pierrot, "Speaker verification in the telephone network: An overview of the technical development activities in the CAVE project" *Proceedings EUROSPEECH-97*, Vol.2, pp. 971-974, 1997.

[2] J. Lindberg, J.W. Koolwaaij, H.-P. Hutter, D. Genoud, M. Blomberg, F. Bimbot, J.-B. Pierrot, "Techniques for a priori decision threshold estimation in Speaker Verification" *Proceedings RLA2C*, Avignon, 1998.

[3] T. Moser, E.A. den Os, H. Jongbloed, L. Boves, "Field test of a speech driven calling card service" *Proceedings RLA2C*, Avignon, 1998.

[4] M. Sokolov, "Speaker Verification on the World Wide Web" *Proceedings EUROSPEECH-97*, Vol.2, pp. 847-850, 1997.

[5] The SR demonstration site at http://lands.let.kun.nl/TSpublic/cave

[6] Form-based File Upload in HTML The RFC archive at http://sunsite.auc.dk/RFC