A Speaker Authentication Module in TelCorreo

Leandro Rodríguez Liñares[1] and Carmen García Mateo[2]

[1] E.T.S.E. Informática
Campus As Lagoas
32004 - Ourense, Spain
leandro@tsc.uvigo.es
[2] E.T.S.E. Telecomunicación
Campus Universitario de Vigo
36200 - Vigo (Pontevedra), Spain
carmen@tsc.uvigo.es

Abstract. Our experiments in Speaker Recognition showed that the combination of Speaker Verification and Utterance Verification techniques is an efficient way of improving the performance of a Speaker Authentication System.
This technique is know implemented in the speaker authentication module of TelCorreo, an e-mail client that allows Internet users to read their e-mail using speech through the telephone.

1 Introduction

TelCorreo [1, 2] is a real-world application designed to give its users the ability to access their electronic mailbox and handle their electronic messages through a simple telephone call. It can be viewed as a bilingual (Spanish/Galician) e-mail client application which interacts with the user through a speech interface.

Traditionally, the security of systems like TelCorreo relied on the use of some private code or password. Given the fact that it has been designed as a dialog system, the basic technology it uses allowed us to implement a voice-based user authentication module.

Presently, there are very few speaker recognition systems to be deployed in real-world applications. The performance achieved by most of the techniques proposed in the literature dramatically drops for out-of-lab conditions. Scanning the literature, we can find two generic approaches for designing a speaker authentication system:

– Verifying the actual content of the speech as a mean of verifying the speaker identity. We call these systems “Utterance verifiers” [3].
– Verifying the speaker voice by itself as a way of verifying the identity. This approach is by nature speech content independent. We call these systems “Speaker verifiers” [4], and although these systems are not speech recognizers, they usually share with speech recognizers most of their technology, namely, most of them are based on a HMM (Hidden Markov Model) formulation with a front-end based on cepstral analysis.

Our experiments in Speaker Recognition showed that the combination of Speaker Verification and Utterance Verification techniques is an efficient way of improving the performance of a Speaker Authentication System. In [5] and [6] several ways of combining such systems are explored using our database called "TelVoice".

Among the explored combination techniques, the neural network combination outperforms the others due to its ability to learn the optimal operation point from the data. However, the real-world implementation of this scheme is only possible when a certain amount of experimental data has been collected.

For TelCorreo, we decided to use what we called the "restrictive criterion". This is equivalent to demand that both tests have to be passed simultaneously to accept the speaker.

The rest of this paper is structured as follows. Section 2 explains the module design from the users' point of view, while section 3 presents TelCorreoDB: an experimental frame adequate for testing the speaker authentication technology used in TelCorreo. Section 4 presents the implementation details. Finally, some conclusions are included.

2 Speaker Authentication Module Design

In this subsection, the speaker authentication module is presented from the user's point of view. When the user access to the system, he/she is asked to introduce the personal identification number using the telephone keypad. From this moment, the goal of the speaker authentication module is to decide if the speaker is actually who he/she claims to be.

The first thing the system does is to verify if the voice speaker authentication option is activated or not for this presumed speaker:

1. If not, the user is asked to introduce his personal user's password using the telephone keypad. If the password is correct, the rest of the call is managed by the system's kernel.
2. If the option is activated, there are two possible situations:
   (a) If the user has trained an acoustic model previously, he is asked to pronounce the personal user's password. These utterance is used by the speaker authentication module to determine if the accessing attempt has to be accepted or not.
   (b) If there is no model for this user, he is asked to introduce the personal user's password with the telephone keypad. After the password is verified, the system asks if the user wants to train his model or not in this moment. If he does, the user is asked to repeat a set of phrases for training a model that will be used in the following accessing attempts. If the user does not want to train a model in this moment, the control of the call is passed to the system's kernel.

3 TelCorreoDB

When transferring technology from laboratory to real-world, an important point is how to test and adjust the systems before their field deployment. This process must be made in conditions close to the ones the system is going to have to deal with in real-world.

With this purpose in mind, we recorded TelCorreoDB: a database designed as an experimental frame adequate for testing the speaker authentication technology used in TelCorreo. TelCorreoDB includes voice of 15 speakers recorded in the same conditions in which these speakers would access through the speaker authentication module. Each speaker recorded one training session, which consists of 8 fixed phrases up to a total of approximately 50 seconds and three test sessions, each one with a pronunciation of the personal user's password.

We used TelCorreoDB for the following tasks:

1. Testing the components of the speaker authentication module: parametrization, training, users' information mantainement...
2. Estimating thresholds for the speaker verification sub-module.
3. Training the world model.

4 Implementation Details

4.1 Front-end Block

The modules should share technology as much as possible due to computational reasons. This includes a front-end block that performs the parametrization of the input utterance. The used parametrization consist of the energy and 12 mel-cepstra coefficients with their first and second derivatives. This makes a total of 39 parameters per vector.

4.2 Utterance Verification Sub-module

The Utterance Verifier in TelCorreo is the same speech recognizer used for the dialog management. As the accessing codes consist of four digits, a simple grammar that allows all the possible combination is used. In [1], it is stated that TelCorreo is a bilingual system that can be used in Spanish or Galician indistinctly. Thus, the grammar must allow the presence of digits in both languages. The output of the utterance verifier consists in the sequence of five recognized digits.

4.3 Speaker Verification Sub-module

We decided to use 16-mixtures covariance-tied GMM's (Gaussian Mixture Models), which are a special case of continuous HMM's where the number of states is one. This type of model has proved to be effective for modeling the speaker identity in text-independent speaker recognition applications [4].

In a speaker verification problem the goal is to determine whether a person is who he or she claims to be. The most straightforward approximation would be to use the log-likelihood of the sequence of parameters given the supposed speaker's model $\log(P(\mathbf{O}/\lambda_k))$. This is what it is called *unnormalized log-likelihood score*. The speaker is accepted when the unnormalized score $S_{\text{unnor}}(\mathbf{O}, k)$ is above a certain threshold τ_k:

$$S_{\text{unnor}}(\mathbf{O}, k) = \log(P(\mathbf{O}/\lambda_k)) \ \geq \ \tau_k$$

In [4], it is stated that to avoid taking the verification decision on a relative score dependent on non-speaker utterance variations some normalization procedure must be used. This normalization alleviates effects like voice quality or speaker's vocal tract variations. One possible solution to these problems is to use another probability $P(\mathbf{O}/\widehat{\lambda_k})$ as a normalization factor to calculate the a normalized score:

$$S_{\text{nor}}(\mathbf{O}, k) = \frac{\log(P(\mathbf{O}/\lambda_k))}{\log(P(\mathbf{O}/\widehat{\lambda_k}))} \ \geq \ \tau_k$$

$\widehat{\lambda_k}$ is called the *antimodel* and represents the universe of possible speakers but the presumed one. There are typically two strategies to build it up: to train a world model shared by all the speakers or to use a set of speaker-dependent models called cohort or background models. It is normally accepted that the use of cohort models give better performance than using a world model.

When transferring technology from laboratory conditions to real-world applications, several restrictions must be applied. This is particularly true in the Speaker Verification Sub-module, where the data avalaible to training or estimation procedures is limited. Thus, we have applied several restrictions:

- The number of registered users in TelCorreo is not constant. Then, the use of cohort models for normalization purposes is not practical, as each new registration would imply to recalculate the cohorts for each registered speaker. We decided to use a *world model*, trained using TelCorreoDB. This model consists of a variance-tied 48-mixtures GMM.
- As previously stated, a set of thresholds must be estimated in order to decide if a utterance belongs to a user or not. Theoretically, there are two possibilities: to use one threshold shared by all the speakers or to use speaker-dependent thresholds. The latter option give better results in laboratory conditions. However, threshold estimation needs to be performed in the conditions where the verification process is going to be performed, and in a real world system this is not a practical approach. Then, shared threshold must be used, and these thresholds are a priori calculated using TelCorreoDB.

4.4 Sub-modules Combination

In summary, the system has to decide if a utterance belongs to a speaker or not. This decision has to be made based on a string of recognized digits and an acoustic score calculated by the Speaker Verification Sub-module.

It was previously said that we decided to use the “restrictive criterion”, that is, to demand that both tests have to be passed simultaneously to accept the speaker. To take into account variations of channel or voice quality that would degrade the verification performance, a two-level verification is performed based on the use of two thresholds in the Speaker Verification Sub-module. We call these thresholds τ_{low} and τ_{high}, where $\tau_{low} < \tau_{high}$.

The verification procedure is as follows. First, the recognized sequence of digits is studied. This can lead the system to three distinct situations:

1. If the sequence of digits is totally correct, $S_{\mathrm{nor}}(\mathbf{O}, k)$ is compared with τ_{low}. If $S_{\mathrm{nor}}(\mathbf{O}, k) > \tau_{low}$ the speaker is accepted, otherwise is rejected.
2. If there is only one error in the sequence of digits, $S_{\mathrm{nor}}(\mathbf{O}, k)$ is compared with τ_{high}. As before, if $S_{\mathrm{nor}}(\mathbf{O}, k) > \tau_{high}$ the speaker is accepted, otherwise is rejected. By one error in the sequence of digits we mean one substitution, one insertion or one deletion.
3. If there is more than one error in the sequence of digits the accessing attempt is rejected.

This can be expressed as follows:

$$\left.\begin{array}{c} N_{errors} = 0 \\ S_{\mathrm{nor}}(\mathbf{O}, k) > \tau_{low} \end{array}\right\} \Rightarrow \textit{Accepted}$$

$$\left.\begin{array}{c} N_{errors} = 1 \\ S_{\mathrm{nor}}(\mathbf{O}, k) > \tau_{high} \end{array}\right\} \Rightarrow \textit{Accepted}$$

$$Otherwise \qquad \Rightarrow \textit{Rejected}$$

The estimation of thresholds τ_{low} and τ_{high} is performed using TelCorreoDB. One ROC (Receiver Operating Characteristic) is built representing the false acceptance rate percentage (%FA) versus the false rejection rate percentage (%FR) when the value of the threshold is varied. We decided to use τ_{low} and τ_{high} as shown in table 1. It can be argued that the used database is rather small and that these results may have low statistical significance. However, these thresholds are a first estimation and will probably have to be modified.

	Value	%FA	%FR
τ_{low}	-1	≈5	≈10
τ_{high}	-0.5	≈3	≈18

Table 1. Estimated values for thresholds τ_{low} and τ_{high}

5 Conclusions and Further Work

TelCorreo is a bilingual e-mail client over the telephone line that integrates state-of-the-art speech technology [1]. One of the main and novel components of TelCorreo is the User Authentication Module which is described in depth.

TelCorreo is still under development. A few days before writing this article, TelCorreo was deployed in real-world with real users. The real-world working of the system will raise new issues that will have to be solved. Examples of such issues are the refinement of the thresholds or the adaptation of the users' models.

References

1. L. Rodríguez-Liñares et al.: TelCorreo: A bilingual e-mail client over the telephone. TSD'2000.
2. In `http://www.gts.tsc.uvigo.es/telcorreo` a description of TelCorreo can be found
3. Q. Li et al.: Verbal Information Verification. EuroSpeech'97, Rhodes, Greece, (1997).
4. D. Reynolds: Comparison of Background Normalization Methods for Text-Independent Speaker Verification. Speech Communication, vol. 2, pp. 963-966, (1997).
5. L. Rodríguez-Liñares: Estudio y Mejora de Sistemas de Reconocimiento de Locutores mediante el Uso de Información Verbal y Acústica en un Nuevo Marco Experimental. Ph. D. Thesis, Universidade de Vigo, Spain, (1999). `http://www.tsc.uvigo.es/~leandro/archivos/Tesis.ps.gz`
6. L. Rodríguez-Liñares and C. García-Mateo: A Novel Technique for the Combination of Utterance and Speaker Verification Systems in a Text-dependent Speaker Verification Task. ICSLP'98, Sydney, Australia, (1998).