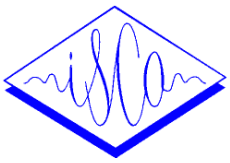

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

APPLICATION OF SPEAKER AUTHENTICATION TECHNOLOGY TO A TELEPHONE DIALOGUE SYSTEM[1]

Leandro Rodríguez Liñares (1)
Carmen García Mateo (2)

(1) E.T.S.E. Informática
Campus As Lagoas
32004 - Ourense
SPAIN

(2) E.T.S.E. Telecomunicación
Campus Universitario de Vigo
36200 – Vigo (Pontevedra)
SPAIN

{leandro,carmen}@gts.tsc.uvigo.es

ABSTRACT

TelCorreo is a dialogue system that allows users to manage their e-mail accounts using the telephone. Regarding speech technology, TelCorreo is a highly demanding application, since it needs of state-of-art text-to-speech system, speech recognizer and speaker authentifier. The main features of TelCorreo are described in [1] and [2]. In this paper, we will concentrate on the speaker authentication module.

Presently, there are very few speaker recognition systems working in real-world applications. The reason is that the performance achieved by most of the techniques proposed in the literature dramatically drops for out-of-lab conditions. Real-world conditions raise new issues to be considered and a compromise between performance and feasibility must be achieved.

We studied the problem of how to increase the performance of a speaker authentifier system by combining the outputs of an utterance verifier and a speaker verifier (the former recognizes the actual verbal content of the speech, while the latter uses speaker voice by itself as a way of verifying the identity).

We have developed a speaker authentifier module suitable to be integrated in a real-world dialogue system. In this paper, we describe the main relevant aspects of this module. The rest of this paper is structured as follows. Section 1 introduces TelCorreo and its speaker authentication module. 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 and guidelines for further work are included.

1. INTRODUCTION

TelCorreo 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.

Speaker authentication is a process by which an hypothesis is verified to determine if the input speech does belong to the claimed speaker or not [3]. Speaker Verification and Utterance Verification are examples of techniques that can be used for Speaker Authentication purposes.

Utterance Verification [4] systems make use of a set of speaker-independent speech models to recognize a certain utterance and decide whether a speaker has uttered it or not. If the utterances consist of passwords, this can be used for identity verification purposes.

Speaker Verification [5] consists of accepting or rejecting the claimed identity of a speaker by processing samples of his/her voice. Usually, these systems are based on HMM models that try to represent the characteristics of the talkers' vocal tracts.

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 order to get a performance adequate to deploy the system in real-world. In [6] and [7]

[1] This project has been partially supported by Spanish CICYT under the Project: 1FD97-0077.

several ways of combining such systems are explored by 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 implementation details of the speaker authentication module are explained in the following.

2. SPEAKER AUTHENTICATION MODULE DESIGN

In this section, 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:

- 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.

- 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

It can be thought that the speaker and utterance verification sub-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 gives good performance both in a uttereance and a speaker recognition system. It consist of the energy and 12 mel-cepstra coefficients with their first and second derivatives. The parametrization details can be seen in table 1.

Parametrization	MFCC's + Energy + Δ + ΔΔ
No. of frames for Δ and ΔΔ	5
No. of vector coefficients	(12+1)*3 = 39
Window type and length	Hamming 25 msec.
Frame period	10 msec.
No. of mel filters	24
Preemphasis	k=0.97
Liftering	22
Passband	0-4000 Hz

Table 1: Parametrization details for the front-end block.

4.2. Utterance Verification Sub-Module

It is well known that the use of utterance verification techniques requires of a set of previously trained speaker independent models that represent the linguistic units or sub-words that can be present in the utterances. In this case, we used a set of 25 phoneme-like models plus noise and silence models trained using a telephone database. The models are left-to-right 3 state HMM's with 16 mixtures per state.

The users' passwords consist of sequences of five digits. As can be seen in figure 1, the grammar used to recognize these sequences of digits are built by concatenation of digits with silence and/or noises between them. As TelCorreo is a fully bilingual galician/spanish system, each digit is present in both languages in the grammar.

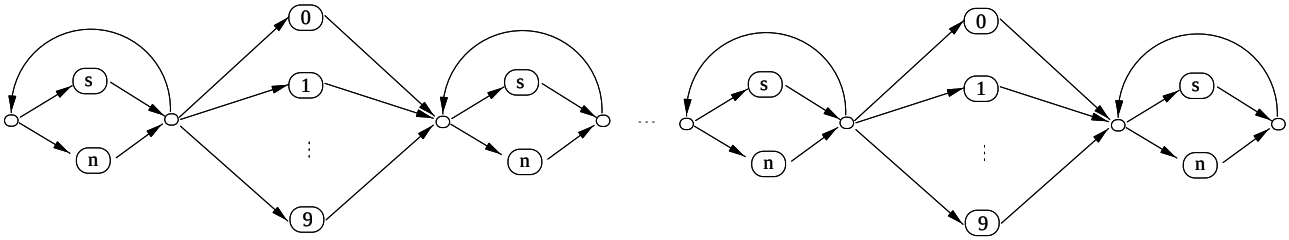

Figure 1: Grammar used by the utterance verifier.

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].

A Voice Activity Detector (VAD) was applied at the input of the Speaker Verification Sub-module to eliminate silence and voice segments. This VAD is based on energy measures over the voice frames.

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 $S_{unnor}(\mathbf{O},k)$ is above a certain threshold τ_k:

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

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 $\log(P(\mathbf{O}/\Lambda_k))$ as a normalization factor to calculate the normalized score:

$$S_{nor}(\mathbf{O},k) = \log(P(\mathbf{O}/\lambda_k))/\log(P(\mathbf{O}/\Lambda_k)) > \tau_k$$

Λ_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:

- If the sequence of digits is totally correct, $S_{nor}(\mathbf{O},k)$ is compared with τ_{low}. If $S_{nor}(\mathbf{O},k) > \tau_{low}$ the speaker is accepted, otherwise is rejected.

- If there is only one error in the sequence of digits, $S_{nor}(\mathbf{O},k)$ is compared with τ_{high}. As

before, if $S_{nor}(\mathbf{O},k) > \tau_{high}$ the speaker is accepted, otherwise is rejected. By one error in the sequence of digits we mean one substitution, one insertion or one deletion.

- If there is more than one error in the sequence of digits the accessing attempt is rejected.

The estimation of the thresholds 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 2. 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 2: 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. 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 weeks before writing this article, TelCorreo was deployed in real-world with real users. The first results are promising, but the real-world working of the system raised new issues that will have to be solved. Examples of such issues are:

- Refinement of the thresholds. As explained, the values for the thresholds were estimated using TelCorreoDB, which was recorded in the very same conditions that TelCorreo will have to cope with. However, the size of this database is rather small, and an adaptation of the thresholds must be made.

- Adaptation of the users' models. Some procedure must be implemented in order to include the voice natural variations into the speakers' models.

- Training evaluation. An open issue is how to test the speaker GMM's in order to decide if they have been correctly trained or not.

6. REFERENCES

1. L. Rodríguez-Liñares et al.: "TelCorreo: A bilingual e-mail client over the telephone". Accepted for publication in TSD'2000 Proceedings, Brno, Chezck Republic (2000).

2. In http://www.gts.tsc.uvigo.es/telcorreo a description of the system TelCorreo can be found.

3. Campbell, J. P.: "Speaker Recognition: A tutorial", *Proceedings of the IEEE*, pp. 1437-1462 (1997)

4. Q. Li et al.: "Verbal Information Verification". *Proceedings of the EuroSpeech'97*, vol. 2, pp. 839-842, (1997).

5. D. Reynolds: "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," *Speech Communication*, vol. 2, pp. 963-966, (1997).

6. 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).
www.gts.tsc.uvigo.es/~leandro/archivos/Tesis.ps.gz

7. 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).