Text Dependent and Text Independent Speaker Verification Systems. Technology and Applications

Svetoslav Marinov
Institutionen för Språk
Högskolan i Skövde
svetljo@hotmail.com

February 26, 2003

Abstract

This paper discusses the differences in Text Dependent and Text Independent Speaker Verification Systems. It shows the basic principles behind these technologies. Some most common applications are reviewed.

1 Introduction

Multi-modal technologies become more and more wide-spread and "fashionable". Yet, some of the applications they are aimed at, like bank, access to buildings or web-sites, etc., entail a certain (if not stringent) security. After all only the authorized users should be allowed access and no others. Whenever spoken communication is involved, designers could choose between two possibilities for their systems - Text Dependent or Text Independent Speaker Verification.

This paper intends to look briefly at these two types of Speaker Verification. I begin by looking at the Biometric Technology, which uses our physical and behavioral characteristics (among which voice/speech) in security applications. I go on to show what is the difference between Text Dependent and Text Independent Speaker Verification. I review the subparts of systems that employ these. Projects and applications that make use of these two types of verification are also presented.

1.1 Biometrics and Biometric technology

The dictionary entry for *biometrics* says - "the statistical[1] study of biological phenomena". While this might seem too broad a notion, it takes a slightly more concrete form in the field of *Security Technology*. Companies looking for more secure authentication for user access to different applications, usually choose among the following three types:

1. something you know - a password, PIN, or piece of personal information (say, the nickname of your godfather);
2. something you have - a card key, smart card, or token;
3. something you are - a biometric.

Of all these, a biometric is the most secure one. It cannot be easily stolen, forgotten, borrowed or forged[2]. Our physical or behavioral characteristics are the actual biometrics. Fingerprints, hand or palm geometry, retina, iris, facial characteristics are our physical, while signature, voice, keystroke pattern and gait are the behavioral characteristics which biometrics measure in order to authenticate our identity.

Speaker verification is among the widely used biometrics when it comes to our behavioral characteristics. Before I discuss some of existing technologies and applications, I will explain two kinds of speaker verification, here - text-dependent and text-independent.

2 Speaker Verification

Quite general, Speaker Verification (SV), is the process of verifying the claimed identity of a registered speaker by using their voice characteristics. SV could be classified as a subpart of the wider field of Speaker Recognition (SR) (Gold and Nelson, 1999), and further subdivided into text-dependent and text-independent SV.

The process is usually initialized by the user, who identifies himself/herself (eg by typing a PIN or entering a secret code) and this claimed identity is further verified by their voice characteristics, thus arriving at the outcome of a binary decision "Accept/Reject" (Blomberg, 2002). The whole process can be summarized in the figure below.

[1] I assume the authors mean that biological phenomena (f.ex. voice characteristics, eye movements, etc.) are captured in terms of mathematical formulas and explained by the laws of physics.

[2] Although it has to be noted that a biometric changes (f.ex. as we grow older, after a car crash, not to mention deliberate plastic operations, etc)

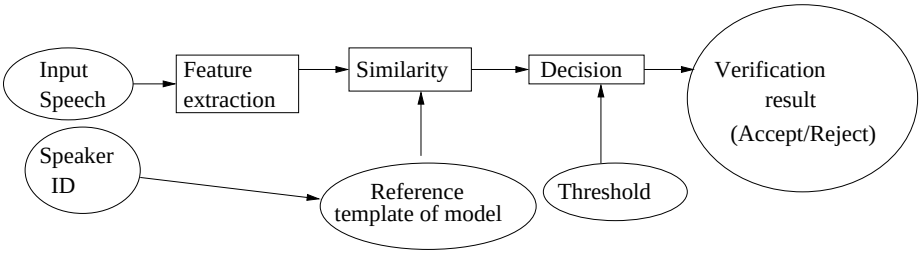

Figure 1: A Speaker Verification System

There are two steps in the process of SV (see Fig. 2). The first is the enrollment (or the training) period, when one creates a model of a new user. The second is the actual verification, when one also computes a model of the voice one hears but then one compares it to the already stored models in order to decide whether to accept the speakers or reject them. The first phase (ie Registering on Fig. 2) is not much different than training a model for Speech Recognition. It employs the same techniques and its outcome serves as the pattern one will eventually compare a new utterance with. In the second phase, the role of the threshold is of an importance. The threshold has been determined after substantial testing of the system and is important for the final rejection or acceptance of a user (see Section 5). Further explanation of the two phases is given in the following sections.

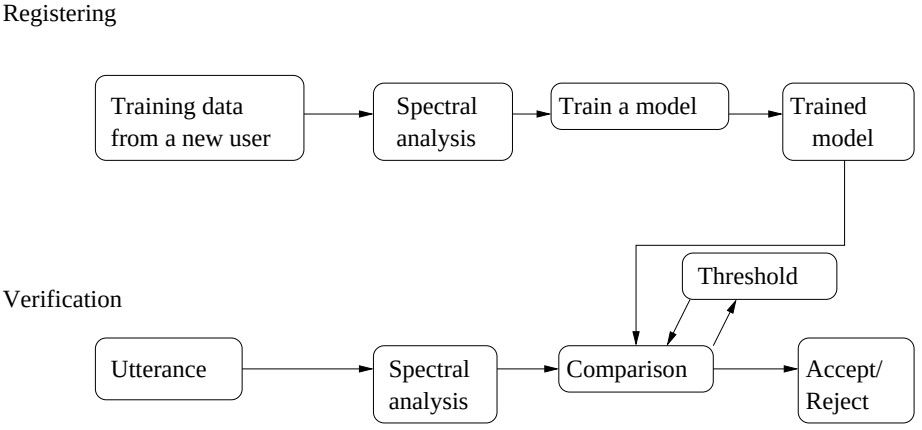

Figure 2: Two Phases in a Verification System

2.1 Feature Parameters

Making a spectral analysis of the speech is the first step in both the enrollment and verification processes. To extract the necessary feature parameters, short-

term (of about 20ms) spectral measurements are widely preferred, such as the Linear Predicative Coding (LPC)-derived cepstral coefficients and their regression coefficients. It is said that a truncated set of cepstral coefficients provides a more stable representation of a speaker's utterance from one repetition to another (Furui, 1996). The regression coefficients are the first- and second- order coefficients extracted from the frame periods in order to represent the spectral dynamics. They are also called delta- and delta-delta-cepstral coefficients.

Another type of stochastic model used for speaker verification is the Gaussian Mixture Model. Such a model is a set of Gaussian distributions over the space of the data, where each Gaussian is characterised by a mean, a covariance matrix, and a prior probability. To use Gaussian Mixture Models, one first breaks the speech sample into successive, short (e.g. 20ms) chunks of speech, and then represents each chunk by a vector of features, such as cepstrum coefficients. The likelihood of the speech is taken to be the product of the likelihoods of each feature vector. These likelihoods are determined from the Gaussian Mixture Model.

2.2 Normalization Techniques

Another important point is that an utterance can never be repeated twice precisely the same way even by the same speaker during the same session. Several factors influence this, such as the natural voice change over time, illness or disturbing background noises. Therefore, the so called normalization techniques, are employed in order to overcome or disregard these variations that are not extremely relevant for the verification process.

When it comes to the effects of the environment, eg background noises, different recordings, transmission conditions, Murthy et al. (1997) suggest that an optimization of the front end processing[3] of the speech signal could significantly improve the recognition process.

2.2.1 User aspects

In an article, from The Ottawa Telephony Group Inc. (OTG)[4], the environmental issues are dealt with improvement on the following fronts:

- Make the user more cooperative. He/She should be aware that the system is not like humans and that it needs a different approach. One has to be prompted to do things clearly, slower, etc.
- One has to be aware of the Biometric System immediately (ie it has to be made overt and not be disguised). Desperately trying to make the impression that we are dealing with a human will only make the results poorer.

[3] ie how and by what means the signal is captured and processed
[4] available on Internet at www.otg.ca

- Let users become habituated to a new Biometric System. If possible already form the first steps of designing it.
- It is recommended that access to the system is supervised in the beginning. In case of confusion a human operation has to be able to intervene.
- Non-standard sampling environments are better left in the research labs before one is sure that they will show similar results to standard ones (ie don't introduce voice-opening-parachutes before you are sure that at 2000 meters hight and heading straight to the earth with more than 300 km/h a user will speak with calm and composure to his parachute.)

The other major issue, the intra and inter speaker variations (ie the changes of the voice of a single person and the distinction of the voices of several people) Rydin (2001) gives a list of factors, adopted from (Doddington, 1998), which might have an influence on the recognition:

1. General changes in a person's voice are seen over time, from session to session
2. Physical and psychical health
3. Educational level and intelligence
4. Speech effort level and speaking rate
5. Experience with the verification system

As Furui (1996) points out, it is important for the SR systems to take into consideration all these variations in the speech samples.

2.2.2 Techniques

Two main normalization techniques are in general being employed - the blind equalization method (in the parameter domain) and the probability method (in the distance/similarity domain).

The first one is effective for the text-dependent SV. It reduces the linear channel effects and long-term spectral variations. It works well if we have a longish utterance, where the cepstral coefficients are averaged over its duration and then these values are subtracted from the cepstral coefficients of each frame.

In the second, one calculates a likelihood ratio dependent on two probabilities - the likelihood of the acoustic data given the claimed identity of the speaker and the probability given the speaker is an imposter.

I turn now to some of the main issues in this survey - the distinction between text dependent and text independent SV.

3 Text Dependent Speaker Verification

Text Dependent Speaker Verification (TDSV) is usually connected with the fact that a predefined utterance is used for training the system and for testing/using it. Several sources point out, however, that there is not a clear distinction between TDSV systems and the Text independent (TISV) ones, (Blomberg, 2002; Rydin, 2001) among others. Here is a scale (Blomberg, 2002) that captures the major system types from mostly text-dependent to mostly text independent.

- Text dependent with a predefined password
- Text dependent with a specific password for each customer
- Vocabulary dependent
- "Action" dependent (f.ex. look at certain phonemes in the text)
- Text independent (the system chooses the texts)
- Text independent (the user chooses a text)

The first two points above can be summarized as a Fixed Phrase Verification, where a predefined phrase is used both during the training and the verification periods. For these cases the Dynamic Time Wrapping (DTW) approach is mostly used. According to Gold and Nelson (1999) "... the password of each user is simply represented as a small number of acoustic sequence templates corresponding to pronunciation of the password. ... the score associated with a new utterance of the password is computed by means of dynamic programming ... against the reference model(s).". The authors discuss the use of Hidden Markov Models (HMM) for the same purpose. There are certainly other methods, see f.ex. (Mathew et al., 1999).

The vocabulary dependent SV systems are being designed with the idea that a possible imposter should not be able to record utterances from a customer and then play them back. Although it seems contradictory that since the vocabulary is known, it is more difficult for an eventual fraud - intuitively I think that if I know what vocabulary the system uses/expects to hear, I might steal some user's saying these words and pretend to be him. The use of digits is the most common fixed vocabulary and if they are generated randomly, and prompted to the user at the time of testing/verification, it is claimed that it becomes harder for anyone to break in instead of you. Another possibility is to use a very large lexicon. But still, this method gives the user the chance, if rejected, to try himself/herself once again on a new prompted utterance. A new test sentence/digit will not be correlated to the previous one and thus the two acoustic vector sequences will not be very similar. To train such a model, phonetic HMMs using Gaussian or multi-Gaussian distributions are typically employed (Gold and Nelson, 1999). Since there might not be enough training data for the HMMs, single-Gaussian single-state phonetic models are the preferred solution.

4 Text Independent Speaker Verification

As the name hints the users here are not restricted to any fixed or prompted phrases. They have the freedom to say whatever they want. To account for the expected freedom of utterances different methods have been proposed among which the following:

1. Long-term statistics and multidimensional autoregressive
2. Vector quantization
3. Fully connected (ergodic) HMMs
4. Artificial Neural Networks

(Gold and Nelson, 1999)

5. Gaussian Mixture Models (GMMs)

As for (1.) above, these methods include calculation of the mean and variance on a sufficiently long acoustic sequence. Another approach is to use statistics of dynamic variables in the cepstral domain (for example) and model them by Multidimensional Autoregressive. In (2.) vector quantization of spectral or cepstral vectors is used to replace the standard vector with an index to a codebook entry. Thus, "... the spectral characteristics of each speaker can be modeled by one or more codebook entries that are representative of that speaker. The score associated with an utterance is then defined as the sum of the distances between each acoustic vector in the sequence and its closest prototype vector from the codebook associated with the putative speaker." (Gold and Nelson, 1999). Good references as for (3.) and (4.) are (Gold and Nelson, 1999) and (Oglesby and Mason, 1990), respectively, as well as the proceedings of the 1990 IEEE Int. Conf. Acoust. Speech Signal Process.

The GMMs are the basis in most of the Speaker Identification systems. The distribution of feature vector extracted from a person's speech is modelled as a Gaussian mixture density. Thus, for a group of speakers, represented by GMMs, the objective of recognition is to find the model which has the maximum a posteriori probability for a given observation sequence. Compared to HMM systems, the GMM systems use one large model and allow the sharing of training data between different mixtures, disregarding phonetic specific information. This leads to a better trained mixture parameters.

Still, how dependent of a certain language model the TISV systems are, is reviewed in Auckenthaler et al. (2000). I could not find other literature on this topic, although from this short study it turned out that TISV systems are really dependent on a certain language model. Vietnamese and Mandarin f.ex. behave differently than English and raise the false alarm percentage. This is due to training the world model for these languages in a similar fashion as the world model for English. Such issues deserve to be tested further.

5 Some Residual Issues

A quite-too-often-cited, though undoubtedly necessary-to-mention, point, is the one about the mistakes which the TD and TI SV systems make. The following is taken from (Blomberg, 2002):

Decision	Speaker Identity	
	True	False
Accept	OK	FA
Reject	FR	OK

The table illustrates the fact that there are cases when the genuine user is falsely rejected (FR) and the imposter is falsely accepted (FA). Currently there are no systems that are 100% reliable. One has to sacrifice something. It is here that the role of the Threshold (see Fig. 1 and 2) plays an important role. In most cases this is the line we draw by saying "we are willing to falsely accept such a percentage of speakers and falsely reject such a percentage of speakers." Such a decision threshold could be matched against a calculation of dividing the probability of the client speaking in a certain manner by the probability of someone else talking this way. Rydin (2001), gives an example of a decision threshold that has a False Reject Rate of 3.9% and a False Accept Rate of 0.3%. It is all a matter of how impenetrable one wants to design the system, so that at the end one might achieve 0.00% FA but only at the expense of having, say 5-8% FR[5].

This FA/FR-problem of an SV system is actually an example of the so called Hypothesis Test in Statistics. Suppose we have a problem to solve that can be split into two competing hypotheses. There are two cases that are therefore relevant. If one of the hypothesis is "simpler", it is given a priority until evidence against it is found. It is then rejected and the competing hypothesis is accepted. It might be that we want to reject a hypothesis, then still he give it a priority/support until it is been proved otherwise. Whatever the method the result is either Accept or Reject.

6 Applications

6.1 Secure Access via telephone

The most straightforward way to employ SV is in the cases when one has to gain access to some secure place via telephone. Voice is completely compatible with the existing transmission protocols via telephone channels, therefore no special adaptations of the system (besides the installment of a SV system) are necessary.

One such example is the so called "calling card services" where SV exchanges

[5] According to the article by the Ottawa Telephony Group Inc.

the use of a PIN. These are the situations where someone is a registered user to a service that allows making phone calls in a foreign country but where the charge is booked directly to the phone bill of the customer. The standard case is that one types the number of a calling card and then his/her identity is verified by a PIN. The non-standard case is simply to speak out the calling card number to a speech recognition engine. The latter tries to verify this claimed identity and if the match is close enough (ie not lower than certain Decision Threshold), the customer is allowed to use the service, ie make a phone call. This method is favoured because of two things - no additional bother for the customer to remember PIN and an extra security against theft of a card and PIN.

Home banking is another application where SV can be applied. For the time being such a service is restricted to operations within the accounts maintained by a single individual. One can f.ex. check the status of their account, transfer money between ones own saving accounts, etc. The security is pretty low in these cases, the users are verified only by saying their PIN and FR almost never occurs (after all who wants to play "robber" with his own savings!). Still, however, it is being researched how secure it could be to use SV for transactions including a second and third party (ie the so called high-risk bank transactions). It is always noted that the security measures should be proportional to the value that could be obtained by this service.

Home shopping (see f.ex. http://www.hsn.com) is the service that is most uninteresting to an imposter. SV is here being employed, though backed up by a human operator. In this service people ring to order products that are later on shipped to their home addresses. In cases when all lines are busy, a customer can always choose to use the automatic service. They just have to speak their telephone number and if their identity is successfully verified, they can start ordering products. If they are rejected, they are redirected to a human operator. But even if their identity is mistaken for someone else and some products are send to another customer, there is no harm because these products cannot go to an unauthorized party (ie a criminal).

6.2 Other application

Detection of speakers in forensic cases boils down, in most situations, to deciding whether a given recording is really from a suspect or not. This is exactly the case discussed at the end of Section 5 - a Hypothesis Test. Leaving aside legal issues (as mentioned in Rydin (2001)), SV can help police discover how many different individuals are involved in a conversation on a tape.

Although I could not find more concrete information on which commercial systems and applications employ SV, I found information on what research projects test this possibility. Such projects tested and worked on at the Center for Speech Technology, the Department of Speech, Music and Hearing, KTH, as given by Blomberg (2002) are:

- TViT - Speaker Verification in the Telephone Network (1995 - 1998)
- CAVE, PICASSO (1995 - 1997; 1998 - 2000) - CAVE was a two-year European Caller Verification Project which was later on continued by its follow-up PICASSO. The projects aimed at testing SV systems in applications with calling cards and banking. Reports available at http://www.kpn-telecom.nl/cave/
- PER - The Prototype Entrance Receptionist. The system is intended to verify, greet and guide employees and visitors at the Department of Speech, Music and Hearing.
- CTT-Bank - Still this is just a virtual bank that employs SV over telephone to verify the users.

Finally, I would like to give a short comparison of the different biometrics and how secure they are.

Characteristics	Fingerprints	Hand Geometry	Retina	Iris	Face	Signature	Voice
Ease of Use	High	High	Low	Medium	Medium	High	High
Error incidence	Dryness, dirt, age	Hand injury, age	Glasses	Poor lighting	Lighting, glasses, age, hair	Changing signature	Noises, colds, weather
Accuracy	High	High	Very High	Very High	High	High	High
User Acceptance	Medium	Medium	Medium	Medium	Medium	Medium	High
Required Security Level	High	Medium	High	Very High	Medium	Medium	Medium
Long-term stability	High	Medium	High	High	Medium	Medium	Medium

7 Conclusion

In this paper I touched briefly the topic of SV. I describe the two fields - TD and TI SV, and I showed in what applications these systems can be used. While most of the applications I reviewed dealt with SV over telephone, there are other "live" applications (similar to PER) that allow users to gain access to buildings. Certainly, surrounding noises might prove crucial for the final adoption of such a system. In general, I see a lot more potential in the use of SV, but at this stage some more "education" on how these kinds of systems should be used is necessary for the future users. This might considerably improve the results and gain confidence in people's minds.

References

Auckenthaler, R., M. J. Carey, and J. S. D. Mason. 2000. Language dependency in text-independent speaker verification. URL `citeseer.nj.nec.com/446216.html`.

Blomberg, Mats. 2002. Speaker Verification. Slides from Introductory Lectures.

Doddington, G. 1998. Speaker recognition evaluation methodology - an overview and perspective. In *Proceedings for RLA2C*.

Furui, Sadaoki. 1996. 1.7: Speaker recognition. In *Survey of the State of the Art in Human Language Technology*. http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.

Gold, Ben, and Morgan Nelson. 1999. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. WileyEurope. Chapter 36.

Mathew, M., B. Yegnanarayana, and R. Sundar. 1999. A neural network-based text-dependent speaker verification system using suprasegmental features. url= citeseer.nj.nec.com/404364.html.

Melin, Håkan. 1996. Speaker Verification in Telecomunication. Seminar in Speech Technology, KTH.

Murthy, Hema A., Françoise Beaufays, Larry P. Heck, and Michel Weintraub. 1997. Robust Text-Independent Speaker Identification over Telephone Channels. In *IEEE Trans. Speech and Audio Proc.*. url = citeseer.nj.nec.com/murthy97robust.html.

Oglesby, J., and J. S. Mason. 1990. Optimization of neural models for speaker identification. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 261–264. Albuquerque, N.M.

Rydin, Sara. 2001. Text dependent and text independent speaker verification systems. technology and applications. Term paper in Speech Technology.