Speaker verification – an overview

The basic problem to solve in a speak verification task is to determine who is speaking, in order to take action based on the identity of the speaker. I will start this paper by going through speaker recognition, that is: the techniques used to determine who is speaking rather than what is being said. After that I will describe how speaker recognition can be used in a speaker verification setting, and then finish with a summary.

Speaker recognition

Speaker recognition is the process of determining the speaker given a speech signal. This can be seen as a pattern-matching problem (Campbell 1999), or as a problem of determining who's "voice print" is most like the observed signal (Reynolds 2002). The two views are somewhat compatible, and they both require that there is some enrolment of users that are to be identified.

There are a number of different ways to collect and use these patterns or voice prints.

One way is by matching the speech signal to a template that has been stored (or enrolled, Campbell 1999, Reynolds 2002). This limits the recognition to exactly the enrolled utterance, but a distance between the enrolled and the observed utterance can easily be calculated by using the Dynamic Time Warp algorithm (Campbell 1999, Holmes & Holmes 2001). Reynolds (2002) refers to this method as template matching, whereas Campbell (1999) simply calls it Dynamic Time Warp.

Another way is by vector quantization (Campbell 1999, Holmes & Holmes 2001, Reynolds 2002), where certain features are extracted from the enrolled speech samples to construct a code book that characterizes the speech of the enrolled user. The same features are then extracted from the new signal, and a distance is measured between the new feature vector, and the closest vector in the code book. The distance between the signal and the code book can then be calculated by summing the distance over all feature vectors in the speech signal. The features to extract can be determined automatically. As there is no analysis of how the vectors fit together in vector quantization, this method cannot be used to distinguish two people whose speech only differs in the length of the sounds. Whether this is necessary or not is still an open issue.

Another way to measure the likeliness to a voice print is nearest neighbour (Campbell 1999, Reynolds 2002). This improves the vector quantization by storing all information from the enrolment session, allowing for distinction between temporal shifts as well. As there is little analysis (compared to other methods) during enrolment, a lot of number crunching has to be done during recognition, which means that it is very resource intensive.

Yet another way is to use a stochastic model (hidden Markov model, HMM or Gaussian mixture model, GMM, Campbell 1999, Holmes & Holmes 2001, Reynolds 2002). The likeliness is then calculated as the probability that the model produced the observed signal. This is very similar to those methods used in speech recognition, and has proven to give the most reliable results for this kind of task as well (Reynolds 2002).

As in many other natural language processing fields, the current trend lies in combining different approaches to gain the benefits of all of them.

Speaker Verification

Speaker verification is the process of, given a claimed identity and a speech signal, determining whether or not the speech signal has been produced by the person with the claimed identity. This is called a binary decision, since the system only has to accept or reject a claim. Depending on where such a system is deployed, it can vary significantly.

The reason for trying to determine whether or not a person has claimed his/her true identity is usually because it is desirable to grant or deny access to something (a computer system, building etc.) In this context, speaker verification is a biometrical authorization method. The benefit of biometrical authorization methods over traditional methods (keys, passwords, magnetic cards, pin-codes etc) is that you do not need to carry or remember anything in order to prove your authorization. You might say that you as a person constitute the key. As for different biometrical authorization methods, speaker verification is very unobtrusive and natural compared to e.g. retinal scanning or finger-print reading.

Evaluation

Since the system has to make a binary decision, all possible outcomes of that decision can be summarized as a two by two table like table 1. There is a natural trade-off between false accepts and false rejects, as lowering the number of false accepts will also raise the number of false rejects and vice versa. False accepts means that the security of the system has been breached, whereas false rejects can be an unacceptable hassle, or even lock authorized identities out. As there is a trade-off, there is also a point where the number errors of both types are equally frequent. This point is referred to as the point of equal error, and is usually used for comparing systems.

	System accepts	**System rejects**
Claim is true	Correct	False Reject
Claim is false	False Accept	Correct

Table 1: Possible outcomes of a verification system's responses to a claimed identity.

A generic framework

Depending on what recognition method and under which circumstances the system will be used, the system will vary greatly. There is, however, a generic process that describes any speaker verification system. There are two principal steps to this process: enrolment and verification. During enrolment, a model of a user is created, and during verification, these models are tested against a claim. The process can be illustrated as in figure 1. First a confirmed identification and a speech sample is supplied. The speech sample is then analyzed, and stored in association with the confirmed identity. Later, when someone wants to gain

access to the system, he/she will claim to be enrolled and authorized. He/she does this by claiming an identity and supplying a speech sample. The model for the claimed identity is then fetched, and the system tries to decide how confident it is that the supplied speech matches the model. Based on this confidence, the system makes a decision to either accept or reject the identity claim. The threshold used for accept/reject places the system somewhere in the false-accepts/false-rejects trade-off: a high confidence threshold means a secure system, while a low confidence threshold means a user-friendly system. The variable components in the process are the analysis, the matching and the models. As has been showed in the section about speaker recognition, these three are intertwined, and depend upon each other.

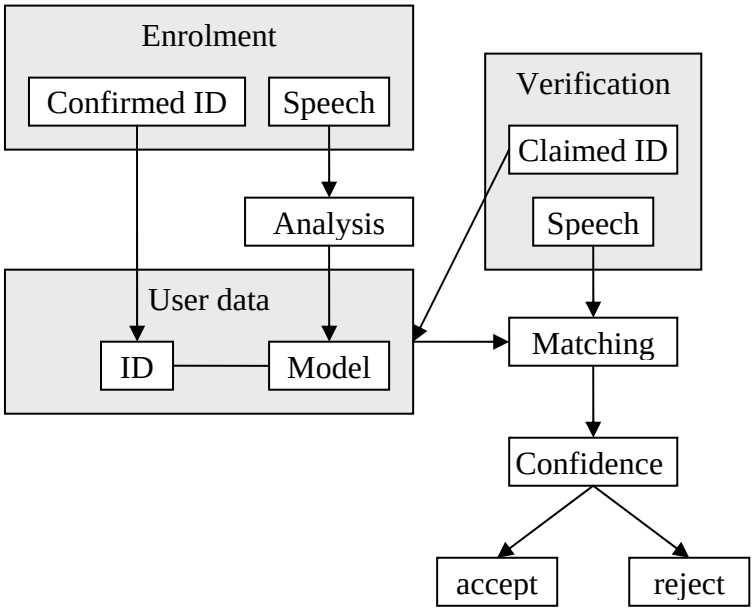

Figure 1: Generic speaker verification process.

Specific methods

The simplest system is to use template matching and store only one pass phrase for each user. This means that the user must repeat the pass phrase exactly as it was enrolled. Among the benefits is the fact that the model is very small, and that the analysis and matching process are very simple. Such a system may fit into e.g. mobile devices where processing power and available memory are limited. Usually some dynamic time-warping is used to allow for some variance between the enrolled utterance and the supplied speech sample. This method also means that the pass phrase has to be memorized, or that a written version is also supplied during enrolment so that it can be prompted.

An easy way to circumvent this system is to record an authorized user uttering his/her pass phrase. The recording can then be played to the system which would grant access. The first step toward a more secure verification system is thus to have variable input.

The natural first step to allow variable input is to use the above system with several pass phrases. The system then prompts the user to utter one (or several) of the predetermined pass

phrases. This means that there is now a trade-off between difficulties and space requirements in enrolment on the one hand, and security on the other, as there needs to be one model for every speaker for every possible prompted word, and the security of the system depends on the number words to choose from when prompting. A solution to this is to model how the authorized users speak rather than how they say specific words. This requires a more advanced model e.g. vector quantization or stochastic modelling. Generally, the stochastic methods work better (Reynolds 2002).

The enrolment procedure still has to make sure that the system has a decent chance to correctly analyze the future input, so the number of promptable words is still constrained. Optimally, the enrolment should be spread out over time, as the qualities of a persons voice (which are modelled) may vary over time. The voice may vary depending on time of day, as well as illnesses (e.g. a cold).

As a stochastic system is limited to ordering the models so that the most probable one is at the top, an intruder could try to claim different identities until he/she found the one that matches his/her speech the best, which would then grant him/her access. This is usually remedied by including an out-of-set speaker model from a large corpus. This means that there is a model that means automatic rejection if it is deemed to match the speech sample the best.

Examples

In this section I will give some examples of systems described in the literature.

Lodi, Toma & Guerrieri (2002) reports on a stochastic system where the memory requirements are kept to a minimum (less than 450 bytes). This means that the voice prints can be stored on virtually any kind of device, making it useful for systems where the user is required to carry a “key” containing their own voice print. They report an equal error rate of 1.2 %, using 3 seconds utterances.

Hsu, Yu & Yang (2003) reports on a GMM-based system called OSCILLO, where there is no need to model out-of-set speaker. The system was tested using three different corpora: TCC-300, TIMIT and NIST (Pryzbocki & Martin 2000). For two of the three, the system lowered both false rejects and false accepts. For the TIMIT task, the baseline system performed better. False accepts are kept under 1 % for all tasks.

Fette et.al. 2000 reports on a system called CiperVOX, which is implemented as a scalable solution used in a server situation (where access to information is granted following authorization.) They use discriminative methods as well as stochastic modelling to achieve an equal error rate 1.42 % in the YOHO corpus.

Ang & Kot (1997) report on a home security system using speaker verification. Only four phonemes are modelled: vowel, glide, stop and fricative, and the equal error rate is 4.5 % (which would not be considered very good today). I have not come across a similar system on the market, and I would think that there is not much of a market for thins kind of system, probably because the traditional authorization systems such as keys are still practical for the relatively small number of people who need access to a house.

Speaker verification, or rather a special case of speaker verification more closely related to speaker identification, is used in the forensic field. In trials, a computational linguist may be called upon to testify that it is indeed likely that some person is actually the one speaking in some recording that is used as evidence. Although that computational linguist is personally giving the testimony, some automatic analysis has most certainly been applied to verify that the testimony is indeed true. The prosecution (usually) claims that it is indeed person X that is herd on the recording (e.g. phone call or bugging), and the system will give a confidence value of the probability that this claim is true. For further information on how a speaker verification system can be used as a forensic tool, see e.g. Gonzalez-Rodriguez, Fierrez-Aguilar & Ortega-Garcia (2003).

Summary

In this paper I have briefly explained the standard techniques involved in speaker verification, as well as given some motivations for the necessary improvements that have been made over the years. I have also given some examples of systems in the scientific literature. As the error-rates go down, I think we can expect to see this technology becoming common-place, especially since it is relatively cheap to deploy and maintain.

The main problem for this new technology is that there are still too many mistakes (1 % may not sound like much, but consider a building with 400 people passing in every day, that means that 4 intruders slipped past), there also appear to be some scepticism towards this kind of technology, and a lot of people would perhaps rather have a system that examines e.g. fingerprints, since it is more hands-on (no pun intended).

Both of these problems are likely to be addressed in the future, as trust in the new technology will grow when it is no longer new. Since there is a very vital scientific community (AVBPA, ICASSP, ICSLP, and Odyssey) dealing with (among other things) this technology, the quest for fewer errors will continue, making this kind of systems more reliable, and thus more trusted.

References

Ang, Kwok K. & Alex C. Kot (1997) “Speaker Verifcation for Home Security System” in *Proceedings of ICASSP 1997*.

Campbell, Joseph P. Jr. (1995) “Testing with the YOHO CD-ROM Voice Verification Corpus” in *Proceedings of ICASSP 1995*.

Campbell, Joseph P. Jr. (1999) “Speaker Recognition” in Jain, Anil K. (ed) *Biometrics: Personal Identification in Network Society*, Kluwer Academic Publishers, Hingham, MA, USA.

Fette, B.A., C. C. Broun, W. M. Campbell & C. Jaskie (2000) “CipherVOX: Scalable Low-Complexity Speaker Verification” in *Proceedings of ICASSP 2000*.

Gonzales-Rodriguez, J., J. Fierrez-Aguilar & J. Ortega-Garcia (2003) “Forensic Identification Reporting using Automatic Speaker Recognition Systems” in *Proceedings of ICASSP 2003*.

Holmes, John & Wendy Holmes (2002) *Speech Synthesis and Recognition*, Taylor & Francis, London, UK.

Hsu, Chun-Nan, Hau-Chung Yu & Bo-Hou Yang (2003) “Speaker Verification without Background Speaker Models” in *Proceedings of ICASSP 2003*.

Lodi, Andrea, Mario Toma & Roberto Guerrieri (2002) “Very Low Complexity Prompted Speaker Verification System based on HMM-modeling” in *Proceedings of ICASSP 2002*.

Przybocki, Martin & Alvin Martin (2002) “2000 NIST Speaker Recognition Evaluation” in *Speech Communication* 31.

Reynolds, Douglas A. (2002) “An Overview of Autmatic Speaker Recognition Technology” in *Proceedings of ICASSP 2002*.