Text dependent and text independent speaker verification systems. Technology and applications

December 21, 2001

Sara Rydin
rydin@speech.kth.se
Centre for Speech Technology, KTH, Stockholm

Abstract

The interest in speaker verification and speaker recognition in general has been increasing as applications driven by speech has become more and more popular and useful. This paper presents a general overview of the subject of speaker verification. It also describes the differences between text dependent and text independent speaker verification, gives an overview of techniques for, and focus somewhat on suggested applications based on speaker verification.

1 Speaker recognition, introduction

As this review will show, recognition of speaker identity from the acoustic signal can be useful in quite a few applications, by it-self or in combination with other techniques for identification of individuals. This section will give an overview of the subjects of speaker recognition and speaker verification. It will report on what characteristic features that are usually extracted from a person's voice in speaker verification systems, on how the claimant's voice is compared to the database of user models, and on problems found when extracting features and building a speaker model. Section two will focus on different types of verification systems; text dependent and text independent systems. The third section will briefly discuss the subject of evaluation and the fourth section deals in more length with possible application for which speaker verification can be useful.

1.1 Recognition, identification and verification

The task of speaker recognition can be divided into the sub-tasks speaker identification and speaker verification. Speaker identification signifies the determination of the identity of the speaker, while speaker verification stands for the process of accepting or rejecting a speaker as the claimed identity [Furui, 1997, Gold and Morgan, 2000]. Other terms used when speaking about speaker recognition are speaker spotting and speaker detection [Doddington, 1998].

Speaker verification is naturally divided into two parts, the training period, when a model of the user's voice is built up, and the actual verification. A system is thus first trained for a new user voice (this can be done in several sessions), which means that an spectral analysis is performed from which features are extracted to form a speaker model. Secondly, verification of a user voice can be performed by comparing the claimant's voice against the database of trained user models. On the basis of this comparison, the system makes a decision on weather the claimant's identity is one modeled by the training material or not. For consistency, a person that is to be verified against the system will in this review be named the *claimant*.

1.2 Building a speaker model, and verifying a claimant against the model

Speaker verification builds on the assumption that there is something characteristic in every person's speech, that can be used in order to verify his or her identity. The characteristic features found in a speaker's voice are used, both to train a user model and to build up a reference representation for the claimant, that in turn can be used for verification against the user model.

The first stage of extracting information from speech is called feature extraction [Gish and Schmidt, 1994]. [Gish and Schmidt, 1994] says that speech information necessary for many speech processing activities (such as speaker verification) is captured in short-term spectra, which, according to the same authors, is spectral information captured during a period of about 20 ms. [Furui, 1997], on the other hand, writes that effective means to discriminate between users in a system are fundamental frequency and different signal measurements, such as short-term and long-term spectra and overall energy. What characteristic features in a person's voice to look for is naturally dependent on the method for verification used; i.e. text dependent or independent. And, as for speech recognition, it must be so that the *correct* representation of the speaker's voice in form of features is dependent on the classification technique that is used [Gold and Morgan, 2000].

Regardless of system type, i.e. whether the verification system is text dependent, text independent or of some other kind, comparing a claimant with a user model always requires the calculation of a score that tells the system how similar/dissimilar the utterance of the claimant is compared to the model. There are two basic ways of performing the comparison [Gold and Morgan, 2000]. Either an explicit comparison is performed between the features extracted for the claimant and the features that the speaker model is built on; e.g. the spectra are compared. Or, the two can be compared with statistical measures, reporting on the (un)certainty of the fit between the claimant's voice and the model. [Gold and Morgan, 2000] reports that the acoustic similarity measure isn't today as often used as the statistical similarity measure.

1.3 Variability in speaker and communication channel

Two of the factors that influences the success of the verification system are differences in speaker's voices, such as intra speaker variability (i.e. changes in one person's voice over time) [Gold and Morgan, 2000, Doddington, 1998] and environmental effects on the speech signal [Furui, 1997, Gish and Schmidt, 1994, Doddington, 1998].

Typical environmental effects on the speech signal comes from, for example acoustic noise [Doddington, 1998], different recordings and transmission conditions for the speech [Furui, 1997, Doddington, 1998] and other interfering voices or sounds in the environment [Gish and Schmidt, 1994].

When it come to differences in speakers' voice (intra speaker or between different speakers) that have an effect on the quality of the verification system, [Doddington, 1998] gives a list of factors that can have an influence, and points out that one explanation to speaker voice variability could be that the use of speech is a result of what a person does:

1. General changes in a person's voice are seen over time, from session to session (all people's voices are changing as he/she is getting older)
2. Physical and psychical health
3. Educational level and intelligence (taboo to talk about though!)
4. Speech effort level and speaking rate
5. Experience with the verification system

The first item is clearly an intra speaker variability. The third item, I think, must be seen as a variability between speakers. All of the other items (I guess) can be seen as both intra speaker variabilities and variabilities between different speakers. Verification of a specific speaker can be dependent on the amount of time since training the system, on his/her health, on his/her speech effort level and on his/her gradual increase in experience with the system. Between speakers, the result for verification is dependent on the state of health, educational level and intelligence, speech effort level and speaking rate, and experience with the system.

2 Text dependent and text independent systems

Verification systems are usually said to be text dependent or text independent (see for example [Doddington, 1998, Gold and Morgan, 2000, Furui, 1997]). A text dependent system is tied to a pre-defined text that is used both for training and verification, while a text independent system should be able to use for any text. Though, as [Melin, 1996] notes, the two-cut division between text dependence and text independence is a bit simplifying, and the boundary between the two types of systems is usually less clear. The following list of systems with a descending levels of text dependency is found in [Melin, 1996] [1]:

1. Fixed password system
2. User-specific text-dependent system
3. Vocabulary-dependent system
4. Speech-event-dependent system
5. Machine-driven text-independent system

[1] The original list came from: Bimbot, F., Chollet G., Paoloni A. 1994, Assessment Methodology for Speaker Identification and Verification systems, Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, pp 75-82.

6. User-driven text-independent system

A third type of automatic speaker verification approach is mentioned by for example [Boves and den Os, 1998]; text-prompted speaker verification. This kind of system addresses and solves the problem that verification systems could be tricked by a taped record of the users password. Usually this problem is solved, in a text-prompted approach, by letting the claimant repeat random digits that are given by the system.

2.1 Text dependent

In text dependent systems, a reference model for the speaker, uttering the expected password, is build up. As the password is thus known by the system in advance, text dependent speaker models can be based both on characteristics for the speaker and for the words/phones in the password [Gold and Morgan, 2000].

[Gold and Morgan, 2000, Furui, 1997] reports that the two methods used for text independent speaker verification (which are similar to methods used for speech recognition) are DTW (dynamic time warping) based methods and HMM (hidden Markov model) based methods. In DTW based methods, acoustic features of the uttered password is during verification scored against the reference model. The score is computed with dynamic programming and dynamic time warping [Gold and Morgan, 2000]. In the case of HMM based methods, the score against the reference model is computed with the viterbi algorithm or with forward (alfa) recurrence [Gold and Morgan, 2000]. HMM based approaches are to prefer, as these has been found to be more accurate than DTW based approaches [Gold and Morgan, 2000, Furui, 1997].

2.2 Text independent

A text independent speaker verification system has no prior knowledge, what so ever, about what the claimant will say during the verification session. It will therefore have to grant the claimant access (or reject him/her access) solely on the characteristics of his/her voice. Due to the fact that the system cannot use lexical knowledge in it's assessment, the performances tends to be worse than for text-dependent systems [Doddington, 1998, Boves and den Os, 1998]

In some types of speaker verification applications the requirements are such that only text independent systems can be used. This is typically systems for forensics or for surveillance, where the vocabulary used cannot be known in advance [Furui, 1997].

[Gold and Morgan, 2000] and [Furui, 1997] describe a few different types of text independent speaker verification systems. The following three types where described by both authors:

1. Long-term statistics based systems: [Gold and Morgan, 2000, Furui, 1997] report that calculating mean and variance for long acoustic sequences (which is the technique used in long-term statistic systems) does not generally give a good statistic model. It rather gives a summary of all acoustic sequences for the user. Instead a MAR (multivariate auto-regression) model has been used more recently, and results for these are similar to those for standard HMM methods.

2. Vector Quantization methods: when training this model, several short-term feature vectors for a speaker are compressed using vector quantization (VQ) techniques [Furui, 1997]. The resulting cluster model, called a code book is then used to characterize a user identity. Normally, when it comes to models of speakers, several code book entries are used to represent one user [Gold and Morgan, 2000].

3. Ergodic HMM-based methods: this resembles the VQ method, but instead of a code book, an ergodic HMM (a HMM where each state has transitions to every other state) is used while training the speaker model [Gold and Morgan, 2000, Furui, 1997].

[Gold and Morgan, 2000] describes a fourth option, a method based on an artificial neural network (ANN) which is trained with positive and negative examples (for user and possible impostors respectively). Also, [Furui, 1997] outlines a forth possibility, a speech-recognition-based method. In this method a model based on the speaker's phonemes or phoneme classes are trained, to which each claimant is compared for confirmation/rejection.

3 Evaluation and threshold

Speaker verification is a 2-class problem, where the system either can accept a claimant as being the targeted speaker, or reject the claimant as being an impostor. The performance of a well balanced verification system should therefore not change with the number of users [Furui, 1997].

Two important measurements used for verification systems are: false acceptance (FA) and false rejections (FR). False acceptance is when an impostor is accepted by the system, and false rejection is when a genuine user is rejected by the system. One way of using these measurements to compare one system with another is to use the equal-error rate (EER). The EER corresponds to the threshold where the FR and the FA levels are equal [Furui, 1997, Melin, 1996]. See figure 1. for an illustration of FA and FR.

decision/user	genuine user	impostor
accept	OK	FA
reject	FR	OK

Figure 1: Figure on decision alternatives in speaker verification, figure 1 from [Melin, 1996]

[Melin, 1996] stresses that there is, of course a trade of between the FR and FA levels in an actual system application. If the FA is too high, impostors might be granted access to information/objects that the system should protect. And if the FR is too high, the user could be annoyed, which could lead to decreased confidence/acceptance for the system. Also, [Furui, 1997] concludes that, in a real situation, the threshold must be set according to the importance of the FR and FA levels in the specific application.

In a speaker verification test for the WWW, [Boves and Koolwaaij, 1998] reports that the threshold was set so that the False Reject Rate was 3,9% and the False Accept Rate was 0,3%. [Melin et al., 2001] reports on a False Reject rate on

4,7% in the CTT-bank test system (no false accept rates where given as no impostor tests where included).

4 Applications

There are potentially many services and applications in which speaker verification could be used. Some of these are general tasks such as banking over telephone network, telephone shopping, database access services [Gold and Morgan, 2000], voice mail [Furui, 1997], and judicial tasks such as surveillance of suspects and forensics [Champod and Meuwly, 1998]. [Doddington, 1998] makes a useful division of speaker verification applications according to who they serve; there are tasks that clearly are for the benefit of the user him- or herself, and there are task, like the judicial tasks, which mainly serves somebody else than the user.

One general problem with speaker verification systems, as used in applications is that, used by it's own, it can never be really safe as the user models builds on statistics [Boves and den Os, 1998]. I.e. there will always be FAs and FRs, however well the model is tuned and this will give an insecurity level to the system. For this reason, speaker verification is usually used in combination with some other technique (like a PIN code) to ensure a more reasonable security level.

4.1 Forensics and surveillance

[Champod and Meuwly, 1998] investigated if speaker recognition methods (identification and verification) could be used in order to help the police to interpret recorded evidence. This usage of speaker recognition has been suggested by, for example [Doddington, 1998, Gish and Schmidt, 1994]. [Champod and Meuwly, 1998] does conclude, though, that speaker verification (or identification) is inadequate for forensic purposes. The method is inadequate since there are assessments done in/by the system about the users identity, that could/should only be done by the court.

4.2 Site access for the WWW

In [Boves and Koolwaaij, 1998] a test is reported on the use of speaker verification for access to a site on the WWW. During enrollment, in which each user is identified by a specific code, each new user has to repeat a fourteen digit sequence eight times. From these digits, one speaker models are built for each digit and user. As the reported system was set up for test purposes, the verification is constructed to enable attempts from impostors. On accessing the site, the claimant chooses a name to log in as, from a pre-defined list (it is possible to try to log in as somebody else). After this, the access number is displayed on the screen and the claimant are prompted to speak the number. Unfortunately, [Boves and Koolwaaij, 1998] does not directly report how claimants was judged as impostors (instead of just being judged as false rejects).

4.3 Telematics transaction services

[Boves and den Os, 1998] reports on three different areas related to telecommunications where speaker verification can be used. [Boves and den Os, 1998] does not,

in any of these areas, report on actual tests or test results.

The first of these application areas is *Home Banking*. In a home banking service the customer can typically get information on his/her account balance via telephone (or via Internet). [Boves and den Os, 1998] reports on a promising approach, that would increase security for home banking, where the telephone log in is done in two parts, first the user gives his/her account number for identification and next he/she is asked for 'secret' information (information that only the bank and the genuine user knows about). Speech recognition is used at least for the second part, and speaker verification can be used to confirm the user identity.

The second area for speaker verification reported by [Boves and den Os, 1998] is *Calling Card Service*, where a commercially used service called 'The Sprint Foncard service' is discussed. During enrollment, the customer uses a PIN code to identify her/him-self after which he/she trains the system by vocally repeating some identification number three times. The identification number is then used both for identification and verification when using the calling card service. Thus, speech recognition and speaker verification are used to enable this service.

The third area that was reported on is *Directory Assistance* for disabled people. KNP Telecom (the Netherlands) did a field trial where people during enrollment where asked to enter their telephone number both orally and by the buttons on the telephone. Then, when the claimant wants to use the system, identification and verification will be based on the claimant uttering his/her telephone number.

[Melin et al., 2001] writes about tests for another Home Banking system (for Swedish); CTT-bank. The first time that someone uses the system, during enrollment, the user has to speak a user unique seven digit number, as to ensure his or her identity before the system builds a user model. To train the user model the user then has to read sequences of ten digits five times. When actually using the system, the claimant calls the system and gives his or her name and a short sequence of digit to identify her- or him-self. Speaker verification is then performed, with a prompted four-digit code, for authentication of the speaker identity. The system reported on in [Melin et al., 2001] has only been used for test purposes, and a user study has been performed with 24 subjects. The results give a FR rate of 4.7% but no report on FA rate is given since no impostor trails where performed.

4.4 Summary

Recognition of speaker identity from the acoustic signal can be useful in quite a few applications such as Home Banking and in accessing pages on the WWW. For increased security, the speaker verification are usually used in combination with other techniques for identification of individuals. The research area is very active and there has been a lot of progress during the last years; still, I think this overview has shown that there is room for improvements, e.g. in evaluation of application usability.

References

[Boves and den Os, 1998] Boves, L. and den Os, E. (1998). Speaker recognition in telecom applications. In *Proceedings IEEE IVTTA-98*, pages 203–208, Torino.

[Boves and Koolwaaij, 1998] Boves, L. and Koolwaaij, J. (1998). Speaker verification in www applications information. In *Proceedings RLA2C*, pages 178–181, Avignon.

[Champod and Meuwly, 1998] Champod, C. and Meuwly, D. (1998). The inference of identity in forensic speaker recognition. In *Proceeding for RLA2C*.

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

[Furui, 1997] Furui, S. (1997). Recent advances in speaker recognition processing. In *Proceedings for AVBPA*, pages 237–252.

[Gish and Schmidt, 1994] Gish, H. and Schmidt, M. (1994). Text-independent speaker identification. In *IEEE Signal Processing Magazine*, pages 18–32.

[Gold and Morgan, 2000] Gold, B. and Morgan, N. (2000). *Speech and Audio Signal Processing*, chapter 5 and 36, pages 521–530. John Wiley & sons, Inc.

[Melin, 1996] Melin, H. (1996). Speaker verification in telecommunication. Bidrag till Talteknologidagen, KTH.

[Melin et al., 2001] Melin, H., Sandell, A., and Ihse, M. (2001). Ctt-bank: A speech controlled telephone banking system - an initial evaluation. Technical Report 1:1-27, TMH, KTH, Stockholm.