

# Speaker verification systems and security considerations

David A. van Leeuwen

TNO Human Factors,  
Soesterberg, The Netherlands.  
vanLeeuwen@tm.tno.nl

## Abstract

In speaker verification technology, the *security* considerations are quite different from performance measures that are usually studied. The security level of a system is generally expressed in the amount of effort it takes to have a successful break-in attempt. This paper discusses potential weaknesses of speaker verification systems and methods of exploiting these weaknesses, and suggests proper experiments for determining the security level of a speaker verification system.

## Introduction

Automatic Speaker Verification (ASV) technology has been maturing in the past decade to the point that it is ready for large scale application. This means that the systems need to be made robust to all kinds of potential attacks. Speaker verification systems are usually evaluated on their performance with respect to a proper speech database, in terms of the Detection Error Trade-off (DET) [1], a form of Receiver Operating Characteristic (ROC). A good example of this characterization of a system is the yearly NIST Speaker Recognition evaluation [2], that is a driving force in improving the performance of systems in a variety of speaker recognition tasks. The DET describes the tradeoff of false acceptance of so-called impostor speakers versus the false rejection of genuine target speakers of a system. This curve is very indicative of the quality of the system, but it is not enough to evaluate the level of security of a system when it is used in a productive set-up. Security of a system involves robustness against all possible forms of ‘attacks’ to break the system. For instance, a system may function extremely well, with virtually no false acceptance at a low false rejection rate, but if the system is defeated by a simple recording and play back of the target user, it is not secure, and has very little use. In this paper we study the question: given a particular level of performance of a speaker verification system, what issues need to be addressed in order to have a *secure* system? We will concentrate on issues that involve speech and speech technology.

The purpose of this paper is not to give a potential fraudulent reader ideas on how to break into a system that has been (partly) secured by a speaker verification system, but rather to discuss the vulnerabilities of systems in general. The idea is that identifying and knowing weaknesses allows counter measures to be taken

and should, in the end, lead to more secure systems. In the field of information security some organizations offer prizes for people that can crack their codes, in order to get a good estimation of the amount of effort needed to do so [3]. This does not completely rule out the possibility of very clever and malicious person that is capable of doing the same job with less effort and/or resources, but it makes it very unlikely.

## Outline

We will start to define some terms, and then concentrate on resources available to a potential intruder. Then we will list a number of potential weaknesses, some of which are found in literature, and finally discuss them.

## Definitions

**System:** an Automatic Speaker Verification system (ASV), integrated in a set-up where it determines the authenticity of a user for using another part of the set-up.

**Target user:** a person that is meant to use the set-up.

**Impostor:** a person who is not the person who he claims to be.

**Intruder:** a person that tries to defeat the speaker verification system.

**Victim:** a target user as whom the intruder has chosen to try to identify.

**Score of a system:** a figure that is the outcome of a verification trial of a speech utterance with a target user model. In this article, we will assume that the score increases if the likelihood increases that the speech is uttered by the target user.

## Resources

An intruder may have a range of different resources available in order to try to defeat a particular system.

**The system.** The intruder may have a private copy of the system, which allows him to inspect all aspects of operation. The level of detail that intruder has may vary from the complete source code and development manuals, via ‘black box’ operation possibilities, to a paid subscription of, e.g., a web-based verification system.

It may seem far-fetched that an intruder has possession of this information, but in a different area, the Smart Card technology, this is quite normal. In a Smart

Card, a chip card with processing power and memory, the algorithm of encryption is known publicly and the implementation of the algorithm is executed on the chip. The security is concentrated around a key stored on the chip, for which it has been made difficult to get access to. There can be millions of these cards around, so the level of security of the hidden key should be very high.

It may be part of the security policy of a system to hide the algorithm and other details of operation. This then puts demands on many other issues in security, such as physical security, trust of (former) employees, etc.

**Target user's speech.** The intruder may have access to speech samples from his victim. This may include enrollment data and utterances recorded during using the system, and general speech data.

The speech can be recorded in various ways, e.g., by 'bug' microphones, explicitly by interviewing the victim, by installing Trojan horse type of software on the victim's desk-top computer if it has a microphone, or obtained from audio/visual material if the victim is a celebrity.

**Human voice resources.** An organized intruder may have access to a large population of people, that have similar characteristics as his victim, e.g., same gender, race, age group and accent. He may also employ professional impersonators.

**Speech technology.** The intruder may have access to state-of-the-art speech processing technology, including analysis/re-synthesis tools, speech synthesis, voice transformation tools, as well as skilled engineers that can operate these tools.

**Time and computing power.** The intruder may have a certain time to prepare his attack, a particular amount of computing resources in preparation and during an attack.

## Human-based weaknesses

In this section we will discuss potential weaknesses of a system that are based on using a human voice in one way or another in order to defeat the system.

**'Lambs' and 'Wolves.'** It is known from literature [2] that some target users in an evaluation database are exceptionally sensitive to impostors, in other words, that there are many speakers in the database that yield high scores when verified against these target users. In so-called 'bovine metaphor' [4] these speakers are called 'lambs.' Similarly, there are some speakers that function very well as an impostor, *i.e.*, their speech fits many target users's models, and these are traditionally called 'wolves.'

An intruder may analyze some speech of potential victims, and find some that have lamb-characteristics. Similarly, he may have analyzed his own circle of friends, and found some with wolf-characteristics. By combining use of a 'wolf' impostor with a 'lamb' target user, the false acceptance rate might increase by a large factor.

**Impersonators.** In an excellent chapter on the evaluation of speaker recognition systems by Bimbot and Chollet [4] an experiment is described where professional impersonators are given the opportunity to try to mimic

target users in an ASV setup. It is commonly believed among speech researchers, however, that impersonators will not be very effective as impostors. The reason often given is that the impersonator's strength lies in mimicking characteristics of a human's voice that are specifically effective to the human's way of hearing, and not to the verification system which uses different speech characteristics. Also, impersonators tend to caricature certain idiosyncrasies of the voice and way of speaking of their (often well known) victims.

Unfortunately, the experiment described is hypothetical, and we are not aware of such an experiment that is carried out. Just like professional chess players can change their style so that they have a better chance at beating a computer, an impersonator may be able to change his strategy in order to fool an automatic system. Especially when the impersonator can be trained with score feedback from a speaker verification system (in the hypothetical experiment this was not a possibility).

**Family relations.** It is well known in literature that the voice characteristics of members of the same family can be very similar. Typically sons and fathers are difficult to separate, as well as brothers. The reasons for this may be apparent: there is both a large similarity in the genes and therefore also in the physical shape and size of the vocal tract between family members, and in the social and cultural environment. In a forensic context, we know of at least one effort where a database is made of speech of members of the same family [5]. It would be very interesting to see what happens to the DET curves of a system when it is evaluated on such a 'family' database. We are not aware of evaluations of ASV system that use this particular database.

Intrusion on a large scale exploiting this fact is less obvious, but on a small scale this may lead to vulnerabilities in the set-up of a service, especially considering the fact that family members are likely to have access to other forms of authentication.

**Corruption of score threshold.** In an ASV the acceptance/rejection decision is usually taken on the score being above threshold or not. This threshold can be dependent on the target user, or scores can be normalized over target users, so that an independent threshold can be chosen. An intruder may try to influence the choice of the threshold in several ways.

For a specific victim, the intruder may try to lower the quality of the microphone recording by adding noise or other distortions, with the goal to have the victim being rejected by the system. The victim may then complain that he cannot get accepted, and the authorities may lower the threshold for that individual target user. This enlarges the probability of a successful attack by the intruder. The intruder would need a way to tamper with the recording quality of the victim.

On a large scale, the intruder may set up many bogus users of the system that use the verification with an inconsistent user population. Non-trained users will get rejected at start, but can then complain massively with the ASV implementation authorities, and they may decide to lower the threshold in order to keep their (bogus)

users happy. This enlarges the probability of a successful attack by the intruder.

## Technology based weaknesses

In this section we will discuss some speech-technology related methods for defeating an ASV system.

**Recordings.** The most trivial attack that an intruder can use is perhaps a plain recording of the victim's voice. This is a method that has been recognized long ago, *e.g.*, see Ref. 6. Generally counter measures are taken by a challenge/response set-up, requesting a user to utter a particular utterance, and to verify what has been said. The possibilities to defeat a prompted verification challenge by recordings of the victim will depend on the vocabulary of the verification system, the modeling technique used and the technical skills of the intruder. If the challenge vocabulary is limited (consisting of, say, the ten digits) it may be technically feasible to concatenate pre-recorded instances of these words in a way that defeats the system.

Coarticulation of the intruding system will be very poor for a human listener, but a verification system must model cross word coarticulation in some way in order to detect this. A system that uses sub-word units to model speech, *e.g.*, an HMM based system, may be able to do this. Also text-independent systems that use a time derivative of features will be better at detecting lack of coarticulation.

When the intruder has possession of many recordings of digit-string verification sessions by his victim, it is not too difficult to generate a full  $10 \times 10$  cross word coarticulation matrix, and use any 'overlap-add' synthesis-by-concatenation technique (*e.g.*, PSOLA [7] or MBROLA).

**Voice conversion techniques.** A relatively new field in speech processing is that of voice conversion or speaker transformation [8]. This line of research attempts to transform the voice characteristics of a source speaker to the speech signal of a target speaker. A crude description of a voice conversion system is that it uses analysis/synthesis of speech, whereby the parameterized speech after the analysis is transformed according to the speech characteristics of the target speaker. Voice transformation is advertised as a technique that may help automatic speech recognition systems to perform better, help development of voices in text-to-speech systems [9], can be used in the film industry to allow the original actor's voice appear in a dubbed localization of the movie, and even help ASV system developers with insights! [8]

Obviously, a voice transformation system is an ideal tool for an intruder of a ASV system. Interestingly, voice transformation systems are usually evaluated in a subjective listener test [8, 9], and not using an ASV system. We are aware of one publication where the conversion technology is evaluated by using an ASV system [10]. Apparently the voice conversion technology is not yet mature enough to be used to evaluate the ASV, rather than being evaluated by using ASV.

**Speech Synthesis.** Closely related to the previous two methods is speech synthesis. There is a range of technolo-

gies varying from synthesis from first principles (formant-based synthesis) to corpus-based synthesis. The former method has a lower perceptual quality than the latter, but it is completely parameterized. An intruder with a ASV test system may experiment with a synthesis system and tune parameters (semi) automatically until scores for a victim model begin to improve. It is likely that an ASV system is much less critical about the quality of synthesis performance than human listeners are.

The corpus-based speech synthesis systems need many hours of recordings (up to 50 hours is not uncommon) of a professional speaker in order to produce enough words and expressions in the corpus to provide a high-quality voice. However, there is a commercial interest to quickly customize a synthesized voice to a different speaker. Demonstrations of a customized synthesis voices on only one hour of audio material are quite convincing [11]. Depending on how much speech of the victim is available, a corpus-based synthesis system may work to mimic a target user very well. Unfortunately, the production of a voice is rather laborious and it is therefore quite costly to set up an experiment for ASV systems using this type of synthesis.

**Artificial Signals.** Most ASV systems use some form of feature extraction as a first step in the analysis of speech, it is therefore not strictly necessary to expose the ASV with speech. With cleverly composed signals an intruder may be able to control the values of the features that are extracted, and thus the statistics of these features that are modeling a target user. In case the intruder has detailed knowledge of the ASV system, and can get hold of the model data, this technique should allow him to obtain high scores for verification attempts, and thus defeat the system.

A problem that the intruder might have to overcome is that the ASV system will probably work with a challenge/response prompt in which case the signals should be engineered to match the prompt text speech. This way, the intruder is developing a speech synthesis system with the purpose of defeating a speaker verification system rather than having a good perceptual quality of speech.

Admittedly, this attack seems only of academic interest, but it shows that in most realistic implementations the security of the system lies partly in the fact that details of the verification algorithm are kept secret.

## Discussion

Speaker verification is generally presented as a biometric that has a low threshold with users and is well suited for telephony applications because it does not involve any additional technical infrastructure. It is also considered as less powerful than other biometrics such as fingerprints or iris-scans, and vendors of ASV systems generally advise that their technology is to be augmented with authenticity verification based on possession (*e.g.*, a key) or knowledge (*e.g.*, a password). Still, we believe that it is important to assess ASV systems on some of the security weaknesses mentioned above, and it would be interesting to set up some experiments.

Firstly, we would like to investigate what the strategies and performance are of professional and ‘casual’ impersonators. In the experimental set-up the subjects’s goal is to defeat a ‘standard’ ASV system, whereby feedback of the scores of an ASV is given. Of course the subjects are allowed to listen to the target user they are supposed to impersonate. Experimental conditions can include length of training speech material of both the ASV and the subjects, and feedback of score or decision. Measures can be percentage of successful attempts, score development and opinions of the subjects on the difficulty of the task.

Secondly, it would be interesting to record or extend speech databases with speakers that are member of the same family, and investigate the influence of family relationships on the position of the DET curve. In recording the SIVA speech database, available through ELDA, interest was expressed in recruiting impostor speakers with family relationships. [12]

As for voice conversion and speech synthesis techniques, we can imagine experiments where a voice conversion system will be trained on the same material that a ‘standard’ ASV system can use for building speaker models. Experimental condition can include training speech length for conversion and verification system, with an ‘extended data’ condition similar to the NIST evaluations [13]. Measures can be the relative shift of DET curve or Decision Cost Function (DCF) operating point.

### Security model

The proposition of using artificial signals leads to a discussion on the security model that we are to pursue for ASV systems. Which specific parts of an ASV system should be kept secret with high efforts? A paradigm where the internals of the ASV system are part of the security model is sometimes referred to as ‘security by obscurity’ and has repercussions to the security policy of personnel, network and physical access. If the security model of computer cryptography is adopted, the algorithms (feature extraction, modeling, decision strategy) should be open and the security could be concentrated around the speaker (and impostor) models, similarly to the cryptographic key. Generation of the speaker models, however, cannot be carried out securely unless the enrollment of a user is included in the security model—an intruder can otherwise re-create the models on his own copy of the ASV system.

### Other speech technology related issues

There are many more aspects of security that need to be addressed in an ASV set-up. We have only discussed a few aspects related to false-acceptance, but there are also speech-related aspects in so-called ‘denial of service’ attacks. For example, the electrical input amplifiers may be saturated by sub/super sonic noise sources located close to the microphone, or the feature extraction code may crash when exposed to digital zeros in the speech signal during log-energy calculations. These attacks may lead to (temporary) removal of the speaker verification set-up, leaving the security burdens to other parts of the system.

### Conclusions

We have brought up the subject of security in the context of speech technology. We have listed a number of potential weaknesses, but this list will by no means be complete, and the purpose of the presentation is to start a discussion on the issue. The problem with security is that the level of security is that of the weakest link in the ‘chain,’ and that therefore all links should—in principle—be checked. We believe that many interesting experiments can be carried out. When more experimental results will become available, the discussion about a proper security model for ASV systems can develop.

### References

- [1] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech '97*, pages 1895–1898, Rhodes, Greece, 1997.
- [2] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, and Douglas A. Reynolds. The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective. *Speech Communication*, 31:225–254, 2000.
- [3] See, e.g., <http://www.rsasecurity.com/rsalabs/challenges/>.
- [4] Dafydd Gibbon, Roger Moore, and Richard Winski, editors. *Handbook of Standards and Resources for Spoken Language Systems*, chapter Assessment of speaker verification systems, pages 408–480. Mouton de Gruyter, 1997.
- [5] Jos Bouten. Personal Communication.
- [6] G. Doddington. Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE*, 71(11):1651, 1985.
- [7] E. Moulines and W. Verhelst. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467, 1990.
- [8] Oytun Türk. *New methods for voice conversion*. PhD thesis, Boaziçi University, 2003.
- [9] H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using PSOLA techniques. *Speech Communication*, 11:175–187, 1992.
- [10] Levent M. Arslan. Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28:211–226, 1999.
- [11] ScanSoft. Demonstration, 2003.
- [12] Mauro Falcone and Alessandro Gallo. The “SIVA” speech database for speaker verification: description and evaluation. In *Proc. ICSLP*, Philadelphia, 1996.
- [13] The NIST year 2003 Speaker Recognition Evaluation Plan. <http://www.nist.gov/speech/tests/spk/2003/index.htm>, 2003.