# Anti-spoofing: voice databases

Federico Alegre[*], Nicholas Evans[†]
EURECOM, Biot, France
Tomi Kinnunen[‡]
University of Eastern Finland (UEF), Joensuu, Finland
Zhizheng Wu[§]
Nanyang Technological University (NTU), Singapore
Junichi Yamagishi[¶]
National Institute of Informatics, Japan
University of Edinburgh, United Kingdom

## Synonyms

Speaker recognition spoofing; speaker recognition corpora

## Definition

As with any task involving statistical pattern recognition, the assessment of spoofing and anti-spoofing approaches for voice recognition calls for significant-scale databases of spoofed speech signals. Depending on the application, these signals should normally reflect spoofing attacks performed prior to acquisition at the sensor or microphone. Since the collection of large quantities of any biometric data is always extremely time consuming and cost prohibitive, and consistent with some telephony applications, almost all of the existing work to assess spoofing and anti-spoofing in voice recognition is performed with databases of speech signals subjected to post-sensor spoofing attacks. While such a setup can be justified, the lack of suitable sensor-level databases is a weakness in the research field. Additionally, whereas spoofing relates to authentication, and thus predominantly text-dependent scenarios, the majority of the related work involves the use of text-independent databases better suited to surveillance applications than to authentication and spoofing.

[*]alegre@eurecom.fr
[†]evans@eurecom.fr
[‡]tkinnu@cs.uef.fi
[§]wuzz@ntu.edu.sg
[¶]jyamagis@nii.ac.jp

# Main Body Text

## Introduction

A plethora of different databases have been used in a growing number of studies to assess the impact of spoofing on speaker recognition and the performance of anti-spoofing countermeasures. For the most part, the choice of database depends on the form of spoofing under study. Thus far, the community has concentrated on 4 predominant forms [1]: impersonation, replay, speech synthesis and voice conversion.

Whatever the form of attack, however, there are no standard databases which are adequate in their original form for research in spoofing and anti-spoofing. As a result, most studies involve either standard speech databases which are modified according to some particular non-standard spoofing algorithm, or often-small, purpose-collected databases. In neither case are results produced by one study, meaningfully comparable to those produced by another.

Accordingly, this chapter provides only a brief overview of the most significant databases used in prior work and the approaches used to adapt them for spoofing and anti-spoofing studies. The latter part discusses some of the shortcomings in the current research methodology and future needs.

## Existing databases

It is now accepted that automatic speaker verification systems can be vulnerable to a wide variety of spoofing attacks. Impersonation [2] and replay attacks [3] are the least sophisticated and therefore the most accessible [4]. Nonetheless, research to develop anti-spoofing systems capable of detecting impersonation and replay require purpose-made databases; there are no standard databases of impersonated or replayed speech. In addition, it is not possible, or at least extremely troublesome to adapt existing, standard databases for such research and thus impersonation or replay attacks are not considered any further.

Speech synthesis [5] and voice conversion [6] attacks have attracted a great deal of attention. Even if they are the least accessible [4] (they involve sophisticated technology), there is evidence that both forms of attack can provoke significant degradation in ASV performance [1]. In addition, research involving speech synthesis and voice conversion attacks can be performed using adapted, standard databases and thus they are the focus here.

The following provides a brief overview of the most significant databases used in prior work in ASV spoofing involving both text-independent databases and recent efforts using text-dependent databases. First, the general approach is described, with particular focus on how the standard databases are adapted for the study of spoofing and anti-spoofing.

### General approach

A general approach to assess spoofing vulnerabilities and the performance of anti-spoofing countermeasures is illustrated in Figure 1. First, as illustrated in

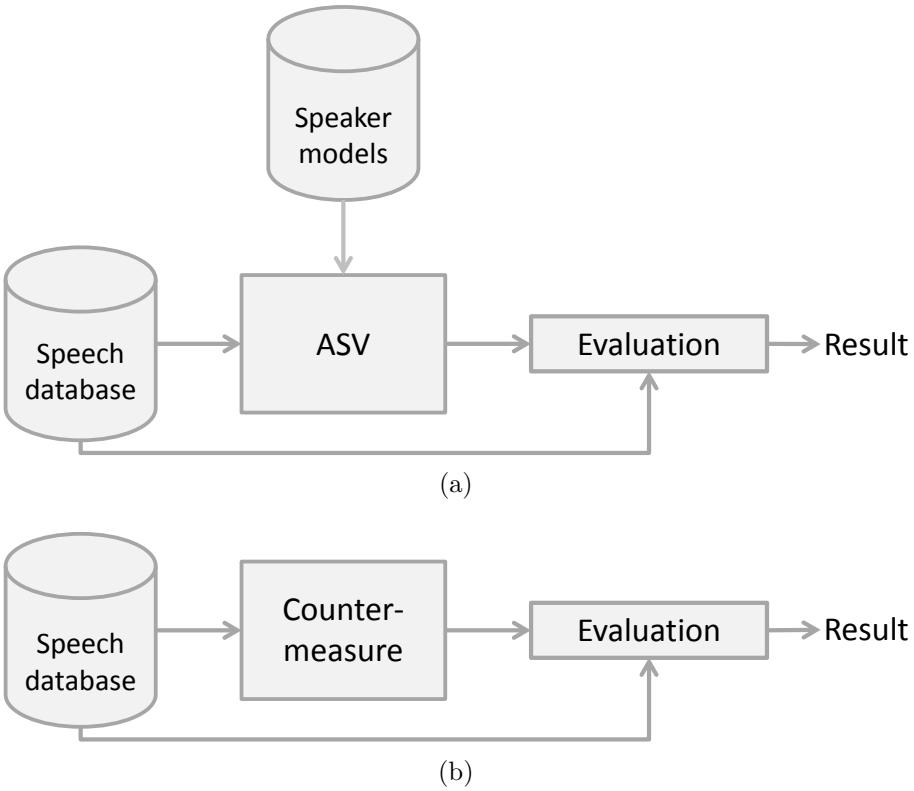
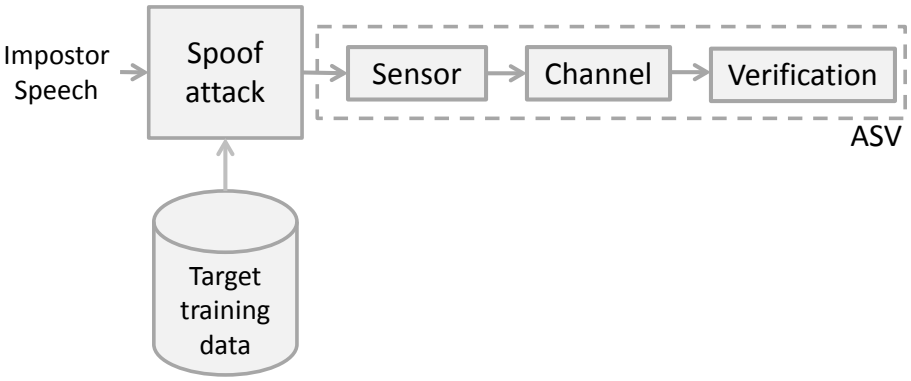



Figure 1: Assessment of (a) automatic speaker verification and (b) anti-spoofing countermeasures. The effect of spoofing on ASV performance is assessed by replacing the naïve impostor trials in (a) by spoofed trials.

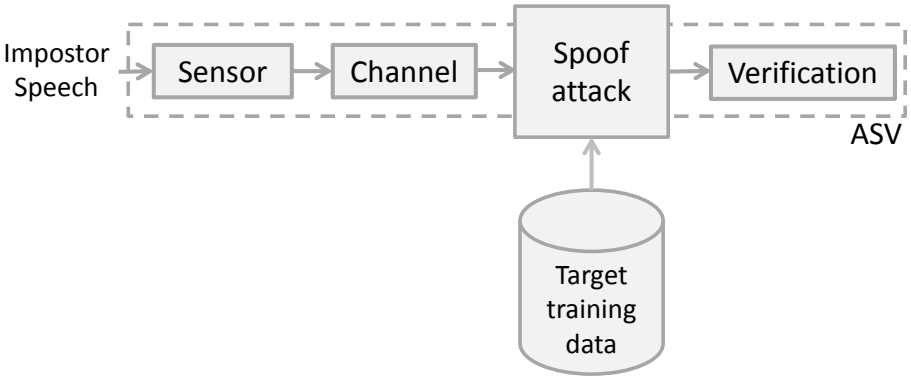
Figure 1(a), a speech database is used to evaluate baseline ASV performance. These experiments assess both genuine client and naïve impostor trials. Second, the naïve impostor trials are replaced with spoofed trials and the experiment is repeated. The aim is then to evaluate the degradation in performance, perhaps in terms of the equal error rate (EER) or false acceptance rate (FAR), usually derived from detection error trade-off (DET) profiles [1, 4].

The performance of anti-spoofing countermeasures is typically assessed in isolation from ASV, as illustrated in Figure 1(b), using the same speech database of genuine and spoofed trials used to assess vulnerabilities in (a). Performance can again be assessed in terms of the EER or FAR. Some researchers have also investigated the resulting effect of countermeasures on ASV performance, e.g. [7].

While this is the common approach, such a setup is not reflective of the traditional consideration of spoofing at the sensor level. Figure 2 illustrates the difference. As illustrated in Figure 2(a), an attacker will normally obtain




(a)




(b)

Figure 2: A comparison of sensor-level spoofing attacks and the general approach to *simulate* spoofing attacks using standard databases.

examples of the target's speech in order to adjust or optimise a spoofing attack at the sensor level. Speech signals are then subjected to acquisition and channel or coding effects before verification. This process differs from the practical setup illustrated in Figure 2(b). Here the spoofing attack is performed post-sensor, immediately before verification.

Such a setup is not necessarily problematic. If the sensor, channel and spoofing attack are assumed to be linear transforms, then the order in which they occur is of no consequence; the two setups illustrated in Figure 2 are equivalent. The issue of sensor-level attacks and the validity of these assumptions is discussed further later on.

The following provides a brief overview of the most significant databases used in prior work. Both text-independent and text-dependent ASV studies are included. Text-dependent ASV systems are arguably the best suited to

authentication applications. This is generally because text-independent systems need comparatively less speech than text-dependent systems to deliver the same recognition performance. Even so, until recently, most research in ASV spoofing was conducted with text-independent databases. This is most likely due to the availability of large, standard text-independent databases and, until recently, the absence of viable text-dependent alternatives.

**Text-independent databases**

The large number of text-independent ASV studies reported in the literature have been performed on a number of different databases. While the earlier studies used relatively small databases, the more recent work has typically used large-scale, publicly available databases.

The Wall Street Journal (WSJ) corpora [8] contain many thousands of high-quality recordings of read Wall Street Journal news texts and have been used in a number of studies involving speech synthesis spoofing attacks. The work of De Leon et al. [10] used a subset of almost 300 speakers which was partitioned into three different subsets for the training and testing of ASV, speech synthesis and countermeasure systems. The databases were used largely as illustrated in Figure 1 to evaluate both vulnerabilities and countermeasure performance.

Most work involving voice conversion spoofing attacks has been performed on the Speaker Recognition Evaluation (SRE) databases [9] developed by the Linguistic Data Consortium (LDC) and the National Institute of Standards and Technology (NIST). They contain different subsets with recordings of varying duration, from short utterances of 10s to multiple conversations, mostly recorded over fixed and mobile telephone conversations. The large number of different databases allows for the use of independent data for the learning of different system elements, for instance background models and countermeasure systems. The work of Matrouf et al. [11] was among the first to investigate voice conversion spoofing attacks with speech data from in the order of 500 speakers. Other work reported by Alegre et al. [12] used an almost identical setup to investigate countermeasures. Both studies were performed according to the general approach outlined above.

While the use of standard databases such as the WSJ and NIST SRE databases would seem to support the comparative analysis of different results, and while there are standard protocols for ASV assessment, the setups used to assess spoofing vulnerabilities and countermeasures are not standardised. Accordingly, results from different studies are typically not meaningfully comparable. These issues are discussed in further detail below.

**Text-dependent databases**

There has been much less focus on text-dependent ASV systems, perhaps due to the lack of databases of a similar scale to those described above. Some of the first work using a publicly available database is reported in Wu et al. [13].

The Robust Speaker Recognition (RSR2015) database [14] contains speech data collected from 300 persons. There are three different subsets containing unique-pass phrases, short commands and connected digits. The work in [13] investigated the effect of text-constraints using matched-transcript, pass-phrase trials.

Once again, while standard ASV protocols might allow for the meaningful comparison of different studies, the RSR2015 database can only be used for spoofing research thorough the use of additional processing with non-standard spoofing algorithms. This is characteristic of all the past work in spoofing, both text-dependent and text-independent.

## Future needs

The recent work has thus utilised existing databases to demonstrate the threat of spoofing and for early studies to develop anti-spoofing countermeasures. The remainder of this chapter reflects on the past work and describes aspects of the existing databases and research methodology which need greater attention in the future. These include the collection of new, properly designed databases and protocols for spoofing and anti-spoofing research which should support reproducibility and provide for comparable results.

### Sensor-level spoofing

With the exception of some small-scale studies involving purpose collected databases, studies in ASV spoofing rarely reflect sensor-level spoofing attacks. Instead, as illustrated in Figure 2 and as already discussed above, attacks are simulated through post-sensor spoofing.

This setup can be acceptable in the case of telephony applications, e.g. [15] or if the sensor, channel and spoofing attack are all linear transforms but, in reality, this is unlikely. The setup is also unrealistic in the case of access/logical control scenarios where the microphone is fixed; the SRE data, for example, contains varying microphone and channel effects.

Accordingly the application of spoofing at the post-sensor level may not be reflective of some practical use-cases. Furthermore, the majority of past work was also conducted under matched conditions, i.e. the data used to learn target models and that used to effect spoofing were collected in the same or similar acoustic environment and over the same or similar channel, whereas this might not be realistic. In order to reduce the bias in results generated according to such setups, future work should study the practical impact of the differences between the two experimental setups illustrated in Figure 2. Alternatively and preferably, future work should include the collection of new databases which more faithfully represent practical scenarios.

### Prior knowledge and generalised countermeasures

Even if they stem from the adaptation of standard databases, all of the past work has been performed on non-standard databases of spoofed speech signals. This has usually entailed the development of a single, or small number of specific spoofing algorithms in order to generated spoofed trials.

Such an approach might be acceptable in the absence of ready-suited, standard databases – at least there is currently no alternative. Even so, countermeasures developed with the current methodology will surely lack generality to new spoofing algorithms or entirely new forms of attack which will likely emerge in the future. In practice, neither the form of the spoofing attack nor the specific algorithm can ever be known. Countermeasure assessments which assume such a priori knowledge are therefore biased towards the specific attacks considered and are likely to over-estimate robustness in the face of varying attacks.

In order to address the inappropriate use of prior knowledge in future work, it will be necessary to collect and make available standard databases of both genuine speech and spoofed speech. Both the form of spoofing and the algorithms used to generates spoofed trials should include as much variation as possible in order to avoid bias and over-fitting. Standard databases will then encourage the development of generalised countermeasures [7] capable of detecting different, varying and perhaps previously unknown spoofing attacks which facilitate the meaningful comparison of different anti-spoofing countermeasures.

## Summary

This chapter describes how existing standard databases have been used for research in spoofing and anti-spoofing for automatic speaker verification. While the use of standard databases and protocols would seem to support the comparison of different research results, none of the existing databases is suited to spoofing research in their original form. The necessary use of non-standard algorithms to simulate spoofing attacks is therefore a limitation in current research. In the future it will necessary to collect new, specifically tailored databases. They should support the meaningful comparison of different results and more faithfully reflect genuine use case scenarios. The inclusion of varying and a priori unknown spoofing attacks will also encourage the development of generalised countermeasures.

# Related Entries

Anti-Spoofing: Voice Conversion
Liveness Assurance in Voice Authentication
NIST SREs (Speaker Recognition Evaluations)
Liveness Detection
Liveness: Voice
Voice Authentication

# References

[1] Evans, N., Kinnunen, T., Yamagishi, J.: Spoofing and countermeasures for automatic speaker verification, INTERSPEECH, Proceedings of (2013)

[2] Lau, Y. W., Wagner, M., Tran, D.: Vulnerability of speaker verification to voice mimicking, in Intelligent Multimedia, Video and Speech Processing, Proceedings of IEEE International Symposium on, pp. 145–148 (2004)

[3] Lindberg, J., Blomberg, M.: Vulnerability in speaker verification-a study of technical impostor techniques, European Conference on Speech Communication and Technology, Proceedings of, vol. 3, pp. 1211–1214 (1999)

[4] Evans, N., Kinnunen, T., Yamagishi, J., Wu, Z., Alegre, F., De Leon, P.: Voice anti-spoofing, in 'Handbook of biometric anti-spoofing', S. Marcel, S. Z. Li and M. Nixon, Eds., Springer (2014)

[5] Masuko, T., Hitotsumatsu, T., Tokuda, K., Kobayashi, T.: On the security of HMM-based speaker verification systems against imposture using synthetic speech, EUROSPEECH, Proceedings of (1999)

[6] Pellom, B. L., Hansen, J. H.: An experimental study of speaker verification sensitivity to computer voice-altered imposters, Acoustics, Speech, and Signal Processing, IEEE Proceedings of the International Conference on, vol. 2, pp. 837–840 (1999)

[7] Alegre, F., Amehraye, A., Evans, N.: A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns, Biometrics: Theory, Applications and Systems, in Proceedings of International Conference on (2013)

[8] Paul, D. B., Baker, J. M.: The design for the Wall Street Journal-based CSR corpus, Speech and Natural Language, Proceedings of the Workshop on, Association for Computational Linguistics, p.p. 357–362 (1992)

[9] Martin, A.: Speaker Databases and Evaluation, Encylopedia of Biometrics, S. Z. Li and A. K. Jain Eds. (2009)

[10] De Leon, P.L., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I.: Evaluation of speaker verification security and detection of HMM-based synthetic speech, Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, pp. 22802290 (2012)

[11] Matrouf, D., Bonastre, J.-F., Fredouille, C.: Effect of speech transformation on impostor acceptance, Acoustics, Speech and Signal Processing, Proceedings of IEEE International Conference on, vol. 1 (2006)

[12] Alegre, F., Amehraye, A., Evans, N.: Spoofing countermeasures to protect automatic speaker verification from voice conversion, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), in Proceedings of (2013)

[13] Wu, Z., Larcher, A., Lee, K. A., Chng, E. S., Kinnunen, T., Li, H.: Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints, INTERSPEECH, in Proceedings of (2013)

[14] Larcher, A., Lee, K. A., Ma, B., Li, H.; The RSR2015: Database for text-dependent speaker verification using multiple pass-phrases, INTERSPEECH, in Proceedings of (2012)

[15] Kinnunen, T., Wu, Z.-Z., Lee, K. A., Sedlak, F., Chng, E. S., Li, H.: Vulnerability of Speaker Verification Systems Against Voice Conversion Spoofing Attacks: the Case of Telephone Speech, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), in Proceedings of, pp. 4401–4404 (2012)