Creating a ”spoofed” database

Michael Shell, *Member, IEEE,* John Doe, *Fellow, OSA,* and Jane Doe, *Life Fellow, IEEE*

I. Objectives

The project will consist in creating a speech corpus for researching spoofing and countermeasures in speaker recognition. The prospective database will:

- be a large (¿ 100 speakers), multi-session speech corpus;
- be based on an existing corpus/corpora;
- contain a large variety of spoofed trials, created with different approaches and different algorithms [1];
- contain speaker identities;
- not contain speech versions (genuine speech, speech spoofed by A, B, ...);
- be accompanied with standardized evaluation protocols.

In order to keep the effort and the size of the corpus at reasonable levels, the following measures are proposed:

- focus on most effective attacks (based on the previous research);
- skipping unlikely spoofing scenarios (e.g., impersonation).

II. Attacks considered

A. Replay

B. Speech synthesis

Speech synthesis systems are constantly improving their quality and can be relatively easy created for multiple voices [3], so they are going to pose an increasing risk to be used for spoofing [2] Suggestions:

- state-of-the-art HMM-based speech synthesis employed, as the one providing best speech quality without the need of large speech data [4];
- to consider: F0 contours can be copied from target sentences (as they are usually too smooth in HTS [5]), to anticipate quality of future HTSs;
- shall we include unit selection speech synthesis? This would require a source DB with a large amount of speech per speaker;
- other suggestions?

C. Voice conversion, incl. artificial signals

It is proposed to use methods preserving phase of natural speech (as they seem to be more resistant to countermeasures such as the one described in [6]).

- A state-of-the art voice conversion method should be used, e.g., Gaussian-dependent filtering [7], joint density GMM [8];
  - MFCC-based conversion proposed. *Justification:* most ASV systems use MFCC-parameterisation [9], so the potential attack can be more effective;
  - to consider: conversion of prosodic parameters (F0, duration) to match that of target speaker;
- shall we also consider unit selection speech conversion? This would require a source DB with a large amount of speech per speaker [10].
- artificial signals [11] optimised for i-vectors-based ASV justification: most state-of-the-art ASV systems use i-vectors, so a potential attack with artificial vectors created that way can be more effective;
- additional suggestions?

III. Attacks omitted

- Impersonation. *Justification*: due to very low practicality [2] and difficulties in potential data collection;
- Replay at transmission level. *Justification*: there is no chance (?) to distinguish it from genuine speech.

IV. Potential source databases

Recording new speakers is not foreseen (costly, time-consuming); therefore it is proposed to use any of the existing speech corpora. Below are the potential candidates for text-independent and text-dependent scenarios.

A. Text-independent scenario

B. Text-dependent scenario

V. Protocols

Dependent on the source DB(s) selected. Both text-independent and text-dependent scenarios shall be foreseen.

VI. Metrics

Scalars:

- EER
- FAR, SFAR (spoof FAR)
- HTER
- MinDCF
- else?

Plots:

- DET profiles
- EPSC curves [15]
- else?

VII. Additional issues to consider

- Shall we use one or multiple source databases?
- Shall we use sensor-level spoofing for synthesis and voice conversion (i.e. combined with a playback), or just channel-level simulations?

M. Shell is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.michaelshell.org/contact.html).

J. Doe and J. Doe are with Anonymous University.

Manuscript received April 19, 2005; revised January 11, 2007.

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LATEX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.