University of Zurich[UZH]

CFPA

Can prosody be the key to spot fake voices?
Acoustic and automatic speaker verification analyses on digital and natural voices

E. Pellegrino, T. Kathiresan, C. Roswandovitz, S. Fruholz, V. Dellwo

INTRODUCTION

- **VOICE CONVERSION (VC):** technique used to convert the perceived identity of one source speaker into that of a target speaker while preserving the linguistic information, provided a sufficient amount of speech material is available for the algorithm training.
- Almost all speech synthesizers and voice-conversion tools are MFCC based, therefore, the VC process generates utterances that sound similar to the target voice in terms of its spectral characteristics.
- With the advances of voice conversion tools, nowadays, it is easier to clone someone's identity/voice.
- Although this has numerous advantages, in forensic caseworks, one can raise the question whether the evidential sample is from a fake or a natural voice.
- In forensic voice analysis is then crucial to find recognition/verification algorithms that allows to distinguish the fake voices from their natural counterparts.

THE STUDY

Given that VC tools are based on MFCC, the VC based-voices should contain the prosodic characteristics of the source speaker. The aim of the present project is twofold:

- **AIM 1**: To examine whether it is possible to identify the source speaker from the prosody of synthesized voices. Do the synthesized voices mirror the prosodic characteristics of the source speaker only, or may they contain also traces of the target voice prosody (e.g. f0 range and contour) stemming from the manipulation of its spectral features?

- **AIM 2:** To understand whether it is possible to spot the fake voices more accurately using voice verification algorithms, like x-vectors, that use statistics pooling layers to capture more temporal based information.

VOICE CONVERSION

Step 1: Data Requirements

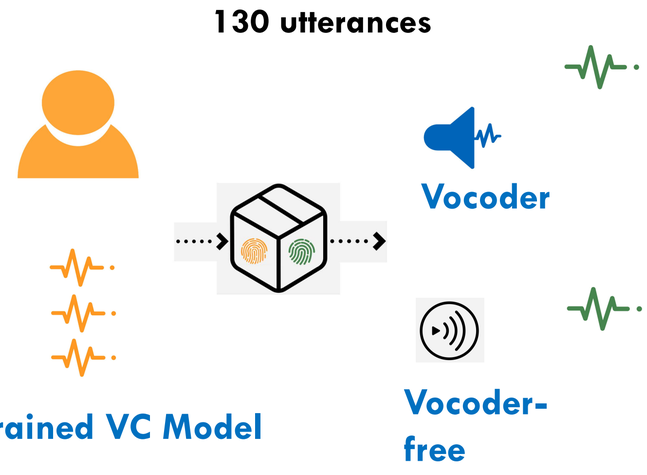
330 utterances

Step 2: Training phase

200 utterances

Step 3: Testing/Synthesis phase

130 utterances

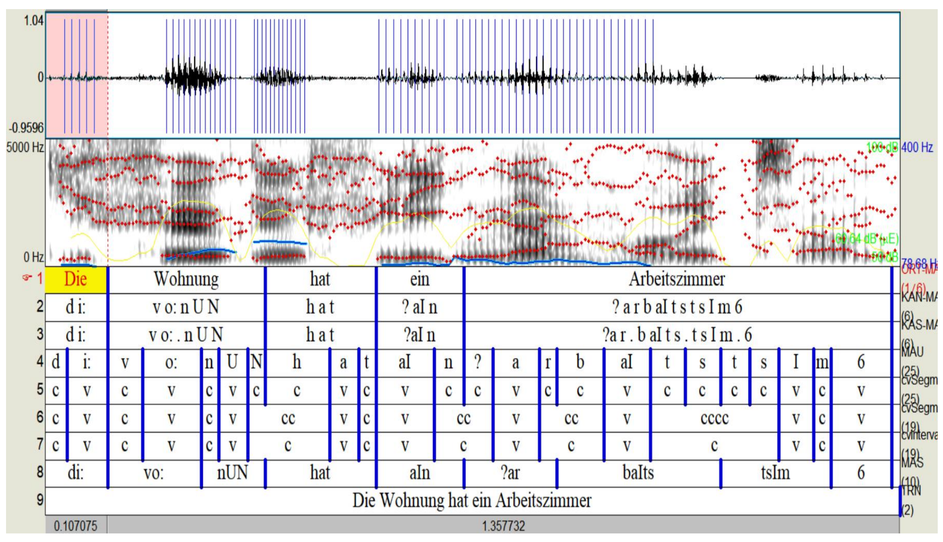


Objective 1: Prosodic characteristics of synthesized voices

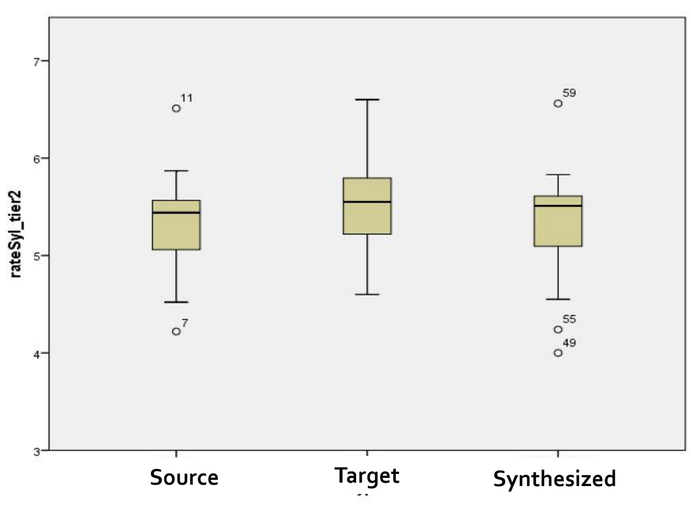
- ***Step 1:*** Automatic annotation of source (S), target (T) and synthesized (Sy) utterances, using Webmaus and CV interval creator.

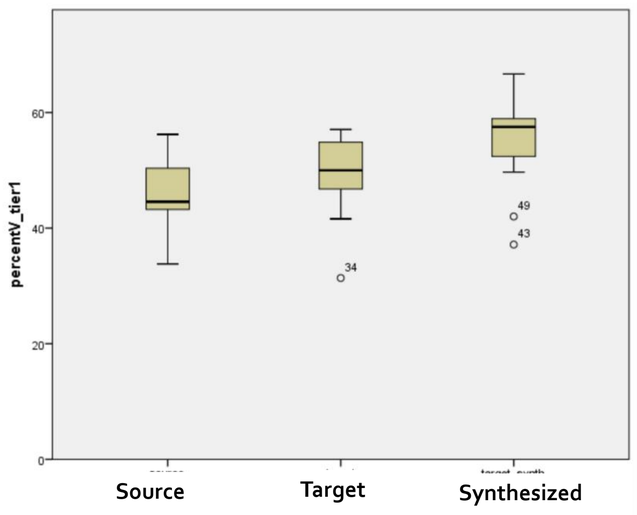
- ***Step 2:*** Acoustic analysis of S, T, and Sy utterances, using Duration Analyzer

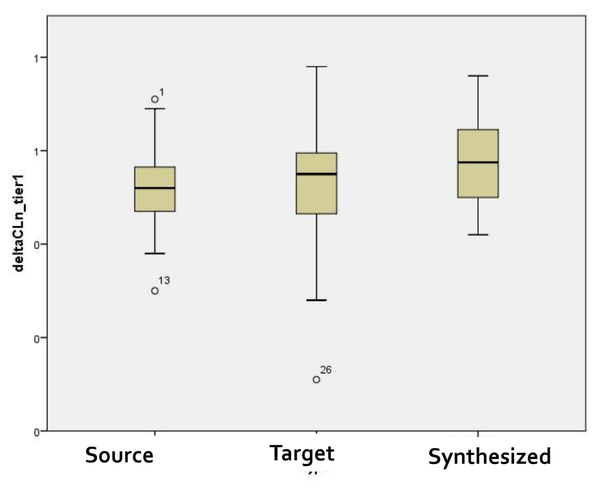
- ***Step 3:*** Rhythmic and prosodic comparisons between S, T and Syn utterances



RESULTS







Objective 2: Comparisons of speaker verification algorithms

Forensic Voice Comparisons of synthesized and the natural voices, using i-vectors and x-vectors voice verification algorithms. If the VC based-voices contain the prosodic characteristics of the source speaker, we expect that the chances of spotting the fake voices increase with x-vectors algorithm as they capture also temporal based information.