Towards the Detection of Speech Deepfakes for Scam Prevention

Anthony White[1], *Catherine Watson*[1]

[1]The University of Auckland

awhi467@aucklanduni.ac.nz; c.watson@aucklanduni.ac.nz

Index Terms: deepfakes, synthetic speech, machine learning, inverse filtering, bispectral analysis

1. Introduction

Deepfakes are digitally synthesised identities of people, created using artificial intelligence and machine learning. The generation of synthetic speech has improved vastly over the past 30 years, from early concatenative synthesis techniques [1] to modern machine learning models [2]. This improvement has made speech deepfakes far easier to generate and more difficult to detect, with previous literature suggesting that listeners can correctly identify modern deepfakes only around 73% of the time [3]. It is therefore becoming easier to digitally assume another person's identity, which can lead to gross invasions of privacy and violations of data sovereignty. Furthermore, this technology has the potential to amplify the effectiveness of telephone scams, which can lead to huge financial losses. In order to prevent scams and other forms of illegal impersonation, it is necessary to develop robust detection methods that can discriminate between natural and synthetic speech. Most existing methods utilise a machine learning approach, with inputs based on cepstral analysis or spectrogram features. This study investigates detection methods that expose the nonlinearities present in synthetic speech, with a focus on the source-filter model [4] and bispectral analysis [5].

2. Approach

In order to conduct research in the detection of synthetic speech, three synthetic voices have been created using the Fastspeech 2 model [2] at a sampling frequency of 22.05 kHz. Two of these voices are in New Zealand English built with the Mansfield corpus [6], with one male speaker and one female. The third is a male voice in te reo Maori, built with the Nga Mahi corpus [7].

The source filter model states that the production of natural speech can be well approximated as a series of linear filters, which model the behaviour of the speaker's articulators [4]. Synthetic speech, on the other hand, originates from machine learning techniques, which are usually nonlinear processes. Preliminary analysis using linear prediction and inverse filtering has shown that high frequencies are not modelled accurately by synthetic speech models, which can be observed in the derived vocal tract frequency response and volume velocity waveforms.

The bispectrum is a higher-order statistics tool which can be used for nonlinearity detection [5]. For a speech signal with Fourier transform $S(f)$, the bispectrum is

$$B(f_1, f_2) = E[S(f_1)S(f_2)S(f_1 + f_2)], \qquad (1)$$

which can detect correlations between frequencies f_1, f_2 and $f_1 + f_2$. Such correlations are common in quadratically nonlinear systems, which can be used to approximate an arbitrarily nonlinear system function using a Taylor series [5]. It is common to normalise the bispectrum magnitude to a $[0, 1]$ interval to produce the bicoherence. The bicoherence for the utterance 'Sunday is the best part of the week' is shown in Figure 1. The magnitude is shown on a grayscale colourmap, with black representing high magnitudes. Here, it is clear that the bicoherence of the synthetic speech has more regions of higher magnitude, indicating that a higher degree of nonlinearity is present.

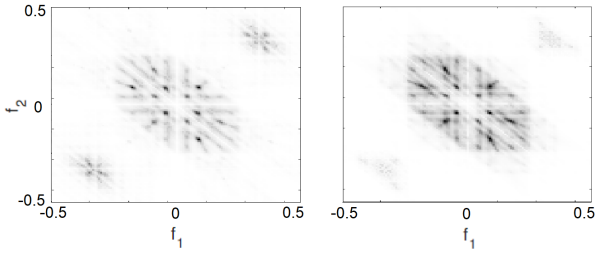

Figure 1: *Bicoherence of natural speech (left) and synthetic speech (right).*

Preliminary testing on 20 speech samples have shown to be promising, with clear differences between natural and synthetic speech appearing in the higher frequencies of the vocal tract responses and in the bicoherence distribution. A larger study is to be conducted, involving a total of 2000 speech samples, which will compare features based on source filter modelling and bispectral analysis. These features will be inputs to a machine learning classifier, which will use standard machine learning techniques (e.g. support vector machines) to discriminate between natural and synthetic speech.

3. References

[1] Moulines E., Charpentier F., Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech communication 9, 5-6 (1990), 453–467.

[2] Ren Y., *et al*, FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, 2020, arXiv preprint arXiv:230301037

[3] Mai K. T., Bray S.D., Davies T., Griffin L.D., Warning: Humans Cannot Reliably Detect Speech Deepfakes. PLOS ONE 18, 8 (August 2023), e0285333.

[4] Fant, G., Acoustic Theory of Speech Production, Mouton, 1960.

[5] Farid H., Detecting digital forgeries using bispectral analysis. Technical report AIM-1657. AI Lab, Massachusetts Institute of Technology; 1999.

[6] Watson C.I., Marchi A., Resources created for building New Zealand English voices. In: Australasian int. conf. of speech science and technology, New Zealand, pages 92–95.

[7] James J., *et al*, Developing Resources for Te Reo Māori Text To Speech Synthesis System. In: P. Sojka, I. Kopeček, K. Pala, and A. Horák, editors, Text, Speech, and Dialogue, pages 294–302.