ANALYSING AND UNDERSTANDING THE DIFFERENCES BETWEEN NATURAL VS. SPOOFED SPEECH.

Sunny Arora[1], Amit Tuteja[2]
[1,2]Guru Kashi University, Talwandi Sabo

ABSTRACT

Voice conversion techniques pose a challenge to speaker verification systems since they convert speech to text. We are investigating ways to automatically discriminate between natural speech and synthetic/converted speech in order to improve the security of speaker verification systems. GCI sites are estimated in this work using the Zero Frequency (ZF) filtering technique, and source features, such as F0, SoE1, and SoE2, are retrieved from GCIs calculated using the Zero Frequency (ZF) filtering technique and the IAIF approach, accordingly. As well as investigating the influence of vocoded speech, we are also attempting to understand the relationship between the F0 and SoE. As a result, we get the F0 contour computed from the speech signal at the GCIs, as well as the locations of those GCIs. For F0, SoE1, and SoE2, we look at the distinctions between natural and spoofed speech to see how they differ. Because of the variances in the parameters of the excitation source, there are differences between natural and spoofed speech.

KEYWORDS: Speech, Natural, Spoofed, Synthetic, Voice, Conversion.

I. INTRODUCTION

Spoofing biometric systems that use automatic speaker identification technology can be accomplished by creating synthetic speech signals using a variety of voice conversion (VC) and speech synthesis (SS) technologies. Over the past several years, a significant amount of research has been committed to the development of various countermeasures to protect speech recognition systems. Countermeasures are divided into two parts: the front-end, which is responsible for parameterizing the speech signal, and the back-end, which is responsible for determining if the speech signal is natural or synthetic. When processing a speech signal, the front-end or feature extraction unit should gather good data from the signal that represents artefacts associated with the conversion or synthesis process. The other section contains a modelling technique that is used to accurately reflect certain speech characteristics. A number of strategies have been developed for both sections in order to increase the performance of the spoofing detection system. For instance, mel-frequency cepstral coefficients (MFCCs), cosine phase, and modified group delay features have all been examined for use in VC-based synthetic speech recognition with a Gaussian mixture model (GMM) as the back-end in some research. A similar technique, relative phase shift (RPS), is utilised in SS-based synthetic speech detection, with higher identification accuracy than MFCCs because of the phase information received through RPS. Despite the fact that converted speech has been found to be capable of confusing a speaker verification system, informal listening experiments have demonstrated that the human ear is capable of distinguishing between real speech and converted speech with ease. Considering that phase spectrum is important for speech perception and that the original/natural phase information is absent from converted speech, we would like to investigate whether or not converted speech can be recognised using phase features.

II. BASIS OF USING F0 AND SoEs

It is possible for humans to modify their vocal fold motions and SoE at the glottis based on the type of utterance and the situation, which can have an impact on the F0 contour as well as the SoE of the speech signal. Consequently, there is a certain amount of correlation or resemblance between the SoE calculated at the glottis and the SoE estimated from the voice. Furthermore, these SoEs are likewise associated in some way to the F0, as previously stated (as shown later in Figure 2). The development of machine-generated speech does not result in a true glottal closure phenomenon (particularly for HTS-based Synthetic Speech (SS) and Voice Converted Speech

(VCS), as is the case with human speech. There are a variety of ways that can be utilised to supply an excitation source in a vocoder during the process of creating speech. In some ways, this is analogous to the mixed excitation model, in which both periodic and aperiodic components are utilised in the generation of speech sounds during the production process. When using excitation information, it is required to change the periodic waveform according to the speaker's F0 range in order for the speaker to sound like the one intended for the application. As an addition to the frequency of occurrence (or F0), the sine of eigenvalue (SoE), or the envelope of the periodic waveform, i.e., the excitation source, will also have an impact on speech quality in various ways. For genuine speech, it has been discovered that the SoE at the glottis calculated by negative peaks of the (t) and the SoE estimated from speech are related to one another in terms of correlation.

III. RESEARCH METHODOLOGY

In order to estimate GCI sites, we employ the Zero Frequency (ZF) filtering approach. With a frame size of 25 ms and a frame shift of 50%, the source features, i.e., F0, SoE1, and SoE2, are extracted at GCIs calculated by ZF and IAIF methods, respectively, using a frame size of 25 ms and a frame shift of 50%. (After discarded the unvoiced regions). When the F0, SoE1, and SoE2 are used, they produce a three-dimensional (3-D) static feature vector, which is denoted by the letter Ds for each GCI position. By taking their first derivative, or velocity, (d1: F0, SoE1, and SoE2) and appending it to the Ds to obtain a 6-D feature vector (D1=Ds+d1), we can additionally account for the dynamics of the F0, SoE1, and SoE2 features. This was performed until the 5th order derivative (i.e., acceleration, jerk, jounce, crackling) was obtained, resulting in D2, D3, D4, and D5, which correspond to 9-D, 12-D, 15-D, and 18-D feature vectors, correspondingly, after which the process was repeated.

IV. DATA ANALYSIS

4.1 Analysis of F0, SoE1 and SoE2 on Spoofed Speech

In this Section, we look at the contrasts between natural and spoofed speech for F0, SoE1, and SoE2 to see what we can learn about them. Using the SAS database, Figure 5.4 depicts the F0 and SoE1 obtained from speech, as well as SoE2 derived from g(t) for a real speech (Panel I) and an HMM-based SS spoof (Panel II) for comparison. Only the negative portion of g(t) is depicted in Figure 1 (d), and the magnitude of g(t)at the GCI is denoted by the symbol SoE2 in Figure 1 (d). As shown by the dotted areas in Figure 1, there are differences in the excitation source properties for natural and SS speech, as well as between the two. When comparing the F0 contour of natural speech to that of SS speech, the former showed more variability (i.e., more dynamic information of the F0 contour of natural speech as compared to the SS speech in Figure 1). Interestingly, these fluctuations were also detected in the SoE computed from speech andg(t). When comparing this particular case of spoofed speech to natural speech, the variances in F0, SoE1 and SoE2 were less than in natural speech. Similar fluctuations in SS and VCS spoof were detected over a number of utterances using vocoder-based SS and VCS spoof, respectively.

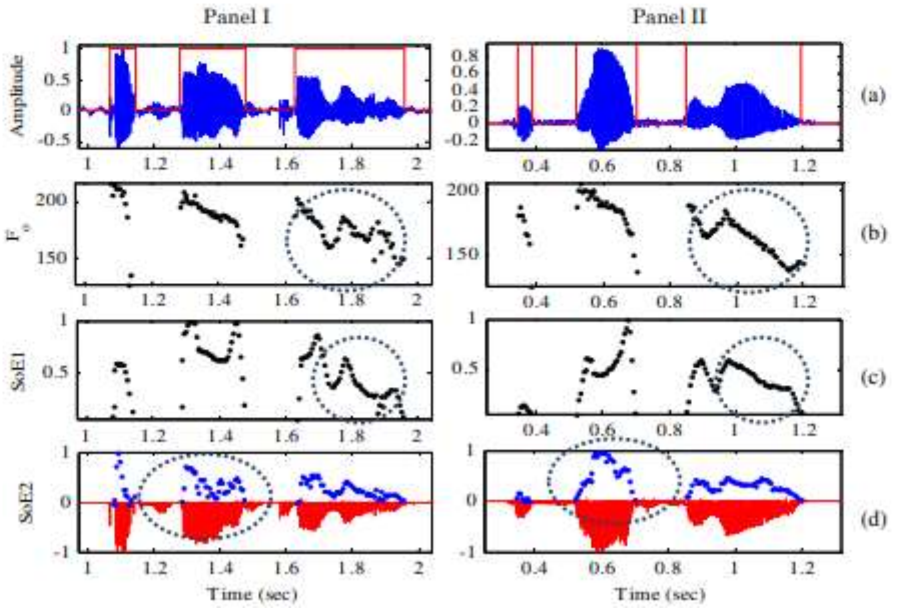
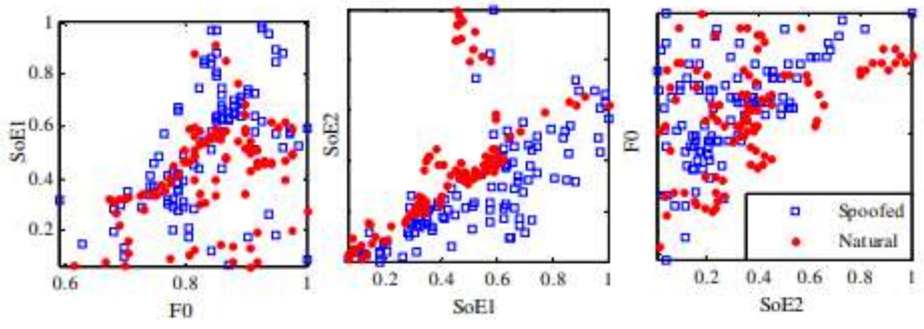


Figure 1: Panel I: Natural speech and Panel II: vocoder-based SS: (a) speech signal \It's nice to hear\, (b) F0 contour assessed by ZF filtering (c) normalized SoE1 at GCIs assessed by ZF filtering and (d) the ġ(t) (red) and normalized SoE2 assessed from ġ(t) at GCIs assessed from ZF filtering (dotted blue).



(a) (b) (c)

Figure 2: Scatterplots for (a) F0 vs. SoE1 (b) SoE1 vs. SoE2 and (c) SoE2 vs. F0 for the natural and vocoder-based SS utterance in Panel I and Panel II, correspondingly (from Figure 1).

Figure 2 depicts the relationship between source-based features for natural and SS spoof in Figure 1 as a scatter plot of F0, SoE1, and SoE2 at GCIs. Figure 1 depicts the relationship between source-based features for natural and SS spoof. When comparing natural speech with SS speech, the correlation are 0.51, 0.73, and 0.51, correspondingly, when comparing F0 with SoE1, SoE1 with SoE2, and SoE2 with F0 (for the speeches depicted in Figure 1). As a result, it is discovered that correlations differ between natural and SS speech. Even though no clear relation can be established between F0, SoE1, and SoE2 for different spoofing methods, there are variances between natural and spoofed speech due to the features of the excitation source used to generate the speech. This will be demonstrated by employing F0, SoE1 and SoE2 as discriminant features for the SSD problem, as well as their dynamics.

4.2 Impact of source features as well as their dynamics:

The source-based features and their dynamics are smaller in scale than the other features. The impact of these source characteristics is therefore investigated by determining the percent EER of the detector for a variety of mixture components in GMM and comparing it to a control (as shown in Figure 3). Testing is carried out on the development set for models that have been developed using the training data.

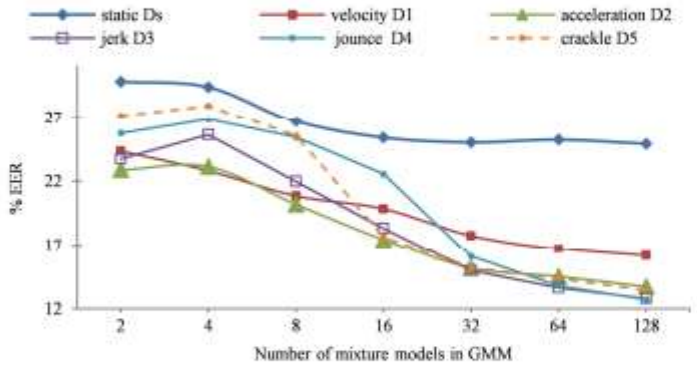


Figure 3: If you take into account the static and varied dynamics of F0, SoE1 and SoE2, such as velocity and acceleration (as well as jerk, jounce, and crackle), you will get a percent EER.

When dynamic information is supplied to the static characteristics in Figure 3, the percentage of EER on the development set drops dramatically. As the number of mixture models in GMM grows, we see a considerable drop in the percent EER. On the basis of 128 combinations, the EER for Ds, D1, D2, D3, D4, and D5 is 24.8%, 16.1%, 13.6%, 12.7%, 12.6%, and 13.4%, respectively. The drop in the jerk (D3) is not considerable, but it also increases significantly when the derivative is larger than the jerk (D3). D3 feature vector with 128 mixture GMM can be evaluated as well.

An estimate of the percent EER using F0, SoE1 and SoE2 up to third order derivative (12-D) was utilised to observe the effects of F0, SoE1 and SoE2. The efficiency of only utilising two attributes simultaneously is also being investigated. Table 1 shows that the EER percentage is relatively high (28 percent) for F0, SoE1 and SoE2 features individually. The percentage EER increased when the D3 feature vector of F0 features was fused with two SoEs one at a time. On the other hand, when the two SoEs and the F0 were combined, the percentage of EER reduced greatly (indicating that the SoEs capture complementary information). As a result, using SoEs in place of F0 results in higher SSD performance. While using all three functions, the EER comes in at a whopping 12.7 percent (as shown in Figure 3). This means that all F0, SoE1, and SoE2 traits are necessary for detecting faked speech.

Table 1: F0, SoE1, and SoE2 feature sets are used individually and when merged with each other using the D3 feature set have the same EER (in percent).

Individual Feature Set	% EER	Feature-level Fusion	% EER
D3: F0	28.10	D3: F0 & SoE1	46.00
D3: SoE1	26.00	D3: F0 & SoE2	44.00
D3: SoE2	27.90	D3: SoE1 & SoE2	**19.00**

V. RESULTS

Table 2 shows the outcomes of the source-based features from Ds to D5, where it can be seen that the percent EER decreases from Ds to D5 for all of the systems from B to K as the source-based features increase.

Table 2: EER (in %) for F0, SoE1 and SoE2 feature set using Ds to D5 feature vectors on training with the ASV spoof data and testing with the Blizzard Challenge 2012 database

Blizzard 2012	Systems	Feature Sets					
		Ds	D1	D2	D3	D4	D5
USS	**B**	39	34	36	32	33	**30**
Hybrid	**C**	55	61	57	54	55	**49**
Hybrid	**D***	61	56	33	16	13	**6**
HMM	**E***	9	10	5	3	3	**1**
USS	**F**	61	59	50	44	40	**38**
USS	**G**	33	24	17	11	10	**8**
HMM	**H**	47	37	26	19	17	**12**
USS	**I**	48	45	38	30	30	**24**
Diphone	**J***	47	40	30	17	16	**10**
HMM	**K***	8	7	3	1	0	**0**

Reduced percent EER is more noticeable for systems with reduced MOS, which indicates that the system is more efficient. Despite the fact that both systems C and D are hybrid systems, their EERs are vastly different. Utilizing the D5 feature vector, the percent EER for USS-based systems B, F, G, and I is 30, 38, 8, and 24, respectively, when using the D5 feature vector. The percent EER for the HMM-based systems E, H, and K is much lower than the percent EER for either the USS-based speech or hybrid systems. The relative enhancement in USS-based systems B, F, G, and I from the Ds to D5 feature vector is 23.07 percent, 37.70 percent, 65.21 percent, and 50.00 percent, respectively, when comparing the Ds to D5 feature vector. These benefits are less significant when compared to the statistically based methods E, H, and K, which have a relative improvement of 88.88 percent, 74.46 percent, and 100 percent, respectively, when compared to the other approaches. Even the diphone-based J system demonstrated a relative improvement in EER, with a 78.72 percent reduction in the EER. As a result, for the Blizzard Challenge 2012 database, the prosody-based feature vector, which was developed from dynamic variation of F0, SoE1, and SoE2 features, produced the highest percent EER when combined with the D5 feature vector. As a result, source-based characteristics can be used to assist in the detection of vocoder-based spoofs, demonstrating that the F0 and SoEs features distinguish between natural and spoof talks in a significant manner.

VI. CONCLUSION

As part of the research described here, we investigated the dynamic variation in the SoE1, and SoE2 features that are produced from speech and from an external excitation source, g(t). It is the movement of the vocal folds that has an effect on the F0 and the loudness of speech in the case of natural human speech production mechanisms. In computer-generated speech, this type of representation is not present (especially, vocoder-based speech). The source-based features do not, by themselves, have the opportunity to greatly degrade the performance of SSD systems.

REFERENCES

1. A Kain and M. Macon, “Spectral voice conversion for text-to-speech synthesis,” in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, vol. 1, May 1998, pp. 285–288 vol.1.

2. Z Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and countermeasures for speaker verification: A survey,” Speech Communication, vol. 66, no. 0, pp. 130 – 153, 2015.

3. Y Stylianou, O. Cappe, and E. Moulines, “Continuous probabilistic transform for voice conversion,” Speech and Audio Processing, IEEE Transactions on, vol. 6, no. 2, pp. 131–142, Mar 1998.

4. Z Wu, et al., "SAS: A speaker verification spoofing database containing diverse attacks," in IEEE Int. Conf. on Acous, Speech, and Sig. Process. (ICASSP), Brisbane, Australia, 2015, pp. 4440-4444.

5. A F. Machado and M. Queiroz, "Voice Conversion: A critical Survey," in Proceedings of Sound and Music Computing (SMC), 2010, pp. 1-8.

6. A W. Black, H. Zen, and K. Tokuda, "Statistical Parametric Speech Synthesis," in Int. Conf. on Acous., Speech and Sig. Process. (ICASSP), Honolulu, Hawaii, USA, 2007, pp. 1229-1232.

7. T Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in IEEE Int. Conf. on Acous., Speech, and Sig. Process. (ICASSP), San Francisco, CA, 1992, pp. 137-140.

8. J Nurminen, H. Silén, V. Popa, E. Helander, and M. Gabbouj , "Voice Conversion," in Speech Enhancement, Modeling and Recognition - Algorithms and Applications, S. Ramakrishnan, Ed. InTech, 2014, ch. 5, pp. 69-94.

9. International Telecommunication Union: A Method for subjective performance Assessment of the quality of speech voice output devices, ITU-T Rec. P.85. https://www.itu.int/rec/T-REC-P.85/en.

10. T Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, “Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech,” in Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, March 2012, pp. 4401–4404.

11. Z Wu, C. E. Siong, and H. Li, “Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition,” in INTERSPEECH, 2012.