Iberspeech 2016
23-25 November 2016, Lisbon, Portugal

Use of the harmonic phase in synthetic speech detection

Jon Sanchez, supervised by Inma Hernáez and Ibon Saratxaga

Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Bilbao
[ion, inma, ibon]@aholab.ehu.eus

Abstract. This PhD dissertation was written by Jon Sanchez and supervised by Inma Hernáez and Ibon Saratxaga. It was defended at the University of the Basque Country the 5th of February 2016. The committee members were Dr. Alfonso Ortega Giménez (UniZar), Dr. Daniel Erro Eslava (UPV/EHU) and Dr. Enric Monte Moreno (UPC). The dissertation was awarded a ente cum laude” qualification.

Keywords: speech processing, harmonic models, harmonic phase, , speaker verification, anti-spoofing, synthetic speech detection

1 Introduction

Nowadays it is critical for some applications to handle the access people have to some places or information. In the last years there has been a growing tendency of using biometric features instead of access-cards, keys or keywords. Biometric characteristics have one main advantage: they cannot be forgotten or stolen. Among all biometric vectors, voice is particularly appealing, as it can be clearly used for identification and users feel largely comfortable about it.

Speaker Verification (SV) systems [1][2] use voice as biometric vector. Impostors could try to deceive the system by impersonating another enrolled user by means of spoofing techniques. The development of voice conversion (VC) [3][4] and text-to-speech (TTS) systems [5][6] has taken them to such a quality level that it is possible to create artificial voices, either converted or synthesized from text, able to fool a biometric speaker verification system.

In this thesis an speaker independent Synthetic Speech Detection (SSD) system is proposed [7] [8]. It can be used to complement a speaker verification system or work independently. The main system is based on a GMM classifier with two different models: a human speech model and a synthetic speech model. The likelihood of the input signal corresponding to each model is calculated and the most probable is selected. To train the acoustic models, some parameters obtained from the harmonic phase of the speech signal are investigated and their performance compared with more traditional parameters obtained from the spectral envelop. The information carried by the phase of the speech signal has been traditionally discarded since [9] established his Acoustic Law for the phase, supporting that the human hearing is able to capture the magnitude of the sounds, but discards the phase. On this basis, many speech technology applications, even nowadays, don't make a great effort to correctly model the phase of the signal, when not directly discard it. This fact can be used as a distinctive element to differentiate natural human voice form that processed in a VC or TTS system. This is why the voice parametrization used to create the acoustic models is based on the harmonic phase of the voice, using the RPS parameterization [10] [11] [12].

The results demonstrate how it is possible to detect synthetic voice based spoofing attacks, using RPS parameter based models trained by means of vocoded speech instead of realistic attacks.

The paper is organized as follows: First, the motivation and objectives of the thesis are described. Then, the main contributions are listed. Finally, the main lines that remain open are detailed.

2 Motivation and objectives

The starting point of this thesis is the work developed in [13] where a new representation of the harmonic phase of the speech signal is proposed, the so-called Relative Phase Shift (RPS) parameters. Additionally, in [14] the harmonic phase of the speech signal is used to protect a ASV system from speaker adapted TTS spoofing. This thesis further explores the use of the harmonic phase and the developed parameterization in anti-spoofing systems.

The system presented in [13] and [14] implements a speaker dependent SSD. For each user, a pair of models is created (natural and artificial), so the decision about the human or synthetic nature of a given input sample is related to the decision about the speaker identity. In the experiments in [13] and [14] the synthetic speech signals were created using speaker adaptation (HTS, HMM Toolkit Speech synthesis [15] [16]). The results showed 100% success on the SSD task.

However, the described system has two major limitations. The first one is that the system is speaker dependent and the decisions about the human or synthetic nature of a given input are only valid for the speakers enrolled in the system. Additionally, in order to train the system models as many adapted TTS systems as speakers are in the system must be developed, which makes the training system very tedious. The second important limitation of the system described in [13] and [14] is that it has been trained/tested only with on type of attack, namely adapted TTS signals from HTS.

In this thesis the development of SSD systems is deeply studied. Keeping the phase information as a decisive parameter, the aim is to get to a more universal synthetic speech detection system, capable of detecting not only adapted voices created using HTS, but a wider set of possible attacks, including different synthesis and voice conversion systems.

2.1 Objectives

The main objective of this thesis is the creation of a universal SSD system, capable of detecting any type of attack separately from the speaker verification system. This involves the following partial objectives:

- Evaluation of the performance of the RPS representation, namely the DCT-mel-RPS parameterization, for the detection system: the detector is designed using DCT-mel-RPS parameters, and the system performance in the detection task is tested and compared with a baseline system based on MFCC (Mel Frequency Cepstral Coefficients) parameters [17].
- Speaker independency: the aim is to create a speaker independent system. With this premise, different models are created using different speakers and different amounts of speakers, and the performance of the related systems are tested.
- Testing and validating the use of attacks created by copy-synthesis, using vocoders to generate the synthetic signals that will train the SSD system, so that the system creation process can be simplified as it is not necessary to make use of real spoofing attacks.
- Vocoder independency: vocoders are selected to train the system and the detector aims to discriminate attacks created with any vocoder available.
- Evaluation of the SSD against real attacks: finally, in order to validate the developed methods and models, the system is faced against attacks from real situations, like synthesized speech or converted speech using unrelated technologies.

3 Main contributions

In this thesis new strategies have been presented to design and implement synthetic speech detection techniques, in the speaker verification area. The independence of the system with the speaker as well as with the vocoder used has been analyzed. A novel training technique has been used to develop the statistical model of the artificial voices, by means of copy-synthesis. This technique can be used to make the training and evaluation process much easier than using real spoofing signals. Finally, the validity of the proposed strategies and models has been thoroughly evaluated using different realistic attacks.

3.1 Usefulness of the RPS parameterization for SSD

A novel SSD system has been developed based on a representation of the Relative Phase Shift: the RPS parameterization for the harmonic phase of the voice.

The RPS is a representation for the harmonic phase information described in **¡Error! No se encuentra el origen de la referencia.**~~[15]~~. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency, as equation (1) shows.

$$h(t)=\sum_{k=1}^{N}A_k\cos\left(\varphi_k(t)\right)\qquad \varphi_k(t)=2\pi k f_o t+\theta_k \tag{1}$$

where N is the number of bands, A_k are the amplitudes, $\varphi_k(t)$ is the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k is the initial phase shift of the k-th sinusoid. The RPS representation consists in calculating the phase shift between every harmonic and the fundamental component (k=1) at a specific point of the fundamental period, namely the point where φ_o=0.

$$\psi_k(t_a)=\varphi_k(t_o)=\varphi_k(t_a)-k\varphi_1(t_a) \tag{2}$$

Equation (2) defines the RPS transformation which allows computing the RPSs (ψ_k) from the instantaneous phases at any point (t_a) of the signal. The RPS values are wrapped to the [-π, π] interval.

The RPS values are not suitable for statistical modelling, so to create and test the models the so-called DCT-mel-RPS parameterization is used instead. These parameters, thoroughly explained in [18], have produced good results in other tasks where statistical modelling is used, such as ASR, Speaker Identification and also Synthetic Speech Detection tasks. To obtain the parameters, the differences of the unwrapped RPS values are filtered with a mel filter bank (48 filters) and a discrete cosine transform (DCT) is applied to the resulting sequence. The DCT is truncated to 20 values and the Δ and ΔΔ values are calculated. The averaged value of the slope of the unwrapped RPS values is also included which leads to a total of 63 phase-based parameters, calculated only for voiced frames, every 10ms.

The system performance has been compared with that got using a more traditional MFCC module parameterization. Comparing the results of both systems, the one based on RPS performs better in most cases, demonstrating the usefulness of RPS parameters in the synthetic speech detection task. Also, RPS parameterization has performed better than other phase parameterizations such as MGD [19] [20].

3.2 Speaker independent SSD

The viability of an SSD system based on speaker independent models has been demonstrated. Even though some previous works had experimented with speaker independent models, they were a part of the SV system, and therefore the results were dependent on the quality of the SV system itself. The speaker independency on a system separated from the Speaker Verification module has been tested for the first time. Speaker independency has been validated with both phase RPS and module MFCC parameters.

3.3 Generating attacks using copy-synthesis

It has been proved that it is possible to work with vocoded copy-synthesized signals instead of creating realistic voices by means of TTS or VC in order to get the synthetic voice model. In most of the studied cases, the error rate of the vocoded voices is similar to these related to TTS voices, even improving in some cases.

The use of vocoders to simulate spoofing attacks brings important practical benefits. On the one hand, it enlarges the signal availability, since it is not necessary to train voice conversion or adapted synthesis algorithms. On the other, it gives a large coverage of spoofing techniques, since most spoofing attacks are vocoder-based.

3.4 Vocoder independent SSD

In this thesis, the robustness of the parameters to the different existing state-of the art vocoders has also been studied. The single-vocoder models performed well when used to detect signals created with the same vocoder, both with RPS and MFCC parameterizations. However, in general, it failed when used to detect signals created with another vocoder, i.e. the system is vocoder-dependent. With RPS parameterization, results were slightly better than those from the MFCC baseline.

To overcome the vocoder dependency problem, the creation of multivocoder models has been proposed. It has been proved that bringing different vocoders together in a single model improves the detection of signals created with vocoders not included in the trained model, i.e. protects the system from unknown attacks. Additionally, the detection error rate for the signals created with vocoders present on the model keeps low. This advantageous effect of bringing vocoders together occurs only for RPS parameterization and does not show up when using MFCC parameterization.

Using this technique of vocoder aggregation a model with information from three different vocoders was created- AHOCODER [21][22], STRAIGHT [23] and MLSA [24]. This multivocoder model covers most of the actual threatens and has been used to protect the systems from unknown attacks, as described below.

3.5 Usefulness against realistic attacks

The SSD system working with the multivocoder RPS models has succeeded in detecting real examples of artificial signals created by unknown statistical synthesizers or voice conversion techniques. The samples were obtained from the Automatic Speaker Verification Spoofing and Countermeasures Challenge Spoofing Challenge (ASVspoof 2015) [25] and the corpus-based TTS Blizzard Challenge, on the 2011 [26] and 2012 [27] editions.

In most experiments, the detection error has been lower than that obtained with the MFCC baseline, or with MGD parameters. The good results demonstrate the generalization capability of the RPS-based models, and represent an advance towards a real universal synthetic speech detector.

4 Future works

During the development of this work some research lines have been identified to improve the proposed system.

First, the vocoders that keep the original phases of the voice are a real threat for the proposed SSD system. Some actions are to be taken to solve this problem, such as training the multivocoder model including signals created with this kind of vocoders, like GlottHMM [28] [29] or AHOCODER-RPS[1].

In a similar way, the proposed system fails to detect signals created with vocoder-less TTS, such as waveform concatenation systems. These synthetic signals were out of the scope of this work, and therefore the necessary improvements have not been tested. Some experiments were performed with the MaryTTS concatenative TTS [30], to evaluate this disability [31].

5 Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competitiveness with FEDER support (RESTORE project,TEC2015-67163-C2-1-R) and the Basque Government (ELKAROLA project, KK-2015/00098).

6 References

1. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE. 85, 1437–1462 (1997).
2. Wang, L., Minami, K., Yamamoto, K., Nakagawa, S.: Speaker identification by combining MFCC and phase information in noisy environments. In: ICASSP. pp. 4502–4505 (2010).
3. Stylianou, Y., Cappe, O., Moulines, E.: Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6, 131–142 (1998).
4. Erro, D., Navas, E., Hernáez, I.: Parametric Voice Conversion Based on Bilinear Frequency Warping Plus Amplitude Scaling. IEEE Trans. Audio. Speech. Lang. Processing. 21, 556–566 (2013).
5. Tokuda, K., Nankaku, Y., Toda, T., Zen, H., Yamagishi, J., Oura, K.: Speech Synthesis Based on Hidden Markov Models. Proc. IEEE. 101, 1234–1252 (2013).
6. Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., Isogai, J.: Analysis of Speaker Adaptation Algorithms for HMM-Based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm. IEEE Trans. Audio. Speech. Lang. Processing. 17, 66–83 (2009).
7. Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D.: A Cross-vocoder Study of Speaker Independent Synthetic Speech Detection using Phase Information. In: Interspeech. pp. 1663–1667. , Singapore (2014).
8. Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D., Raitio, T.: Toward a Universal Synthetic Speech Spoofing Detection using Phase Information. IEEE Trans. Inf. Forensics Secur. PP, 1–1 (2015).
9. Ohm, G.S.: Ueber die Definition des Tones, nebst daran geknüpfter Theorie der Sirene und ähnlicher tonbildender Vorrichtungen. Ann. Phys. 135, 513–565 (1843).

[1] This vocoder is based on AHOCODER but includes information about the original phases, by means of the MRRPS parameterization

10. Saratxaga, I., Hernáez, I., Erro, D., Navas, E., Sanchez, J.: Simple representation of signal phase for harmonic speech models. Electron. Lett. 45, 381 (2009).
11. Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D.: Using Harmonic Phase Information to Improve ASR Rate. In: Interspeech. pp. 1185–1188 (2010).
12. De Leon, P.L., Pucher, M., Yamagishi, J., Hernáez, I., Saratxaga, I.: Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Trans. Audio. Speech. Lang. Processing. 20, 2280–2290 (2012).
13. Saratxaga, I.: La fase en los modelos armónicos de la señal de voz: estrategias de representación, tratamiento y aplicaciones, (2011).
14. De Leon, P.L., Hernáez, I., Saratxaga, I., Pucher, M., Yamagishi, J.: Detection of synthetic speech for the problem of imposture. IEEE (2011).
15. HTS Working Group: HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp/.
16. Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S.: Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. IEEE Trans. Audio. Speech. Lang. Processing. 17, 1208–1230 (2009).
17. Imai, S.: Cepstral analysis synthesis on the mel frequency scale. In: ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 93–96. Institute of Electrical and Electronics Engineers (1983).
18. Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D.: Using harmonic phase information to improve ASR rate. In: Proc. Interspeech 2010. pp. 1185–1188. , Makuhari, Japan (2010).
19. Zhu, D., Paliwal, K.K.: Product of power spectrum and group delay function for speech recognition. In: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. p. I-125-8. IEEE (2004).
20. Hegde, R.M., Murthy, H.A., Gadde, V.R.R.: Significance of the Modified Group Delay Feature in Speech Recognition. IEEE Trans. Audio, Speech Lang. Process. 15, 190–202 (2007).
21. Erro, D., Sainz, I., Navas, E., Hernáez, I.: Improved HNM-Based Vocoder for Statistical Synthesizers. In: Interspeech. pp. 1809–1812. , Florence, Italy (2011).
22. Erro, D., Sainz, I., Navas, E., Hernáez, I.: Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. IEEE J. Sel. Top. Signal Process. 8, 184–194 (2014).
23. Zen, H., Toda, T., Nakamura, N., Tokuda, K.: Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005. IEICE Trans. Inf. Syst. E90–D, 325–333 (2007).
24. Yoshimura, T., Tokuda, K., Kobayashi, T., Masuko, T., Kitamura, T.: Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis. In: Eurospeech. pp. 2347–2350 (1999).
25. Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J.: ASVspoof 2015 : Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan, http://www.spoofingchallenge.org/asvSpoof.pdf.
26. King, S., Karaiskos, V.: The Blizzard Challege 2011. In: Proc. of The Blizzard Challenge 2011. , Torino, Italy (2011).

27. King, S., Karaiskos, V.: The Blizzard Challenge 2012. In: Proc. of The Blizzard Challenge 2012 (2012).

28. Raitio, T., Suni, A., Pulakka, H., Vainio, M., Alku, P.: Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4564–4567. IEEE (2011).

29. Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., Alku, P.: HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering. IEEE Trans. Audio. Speech. Lang. Processing. 19, 153–165 (2011).

30. Schröder, M., Trouvain, J.: The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In: International Journal of Speech Technology. pp. 365–377 (2003).

31. Sanchez, J., Saratxaga, I., Hernáez, I., Navas, E., Erro, D.: The AHOLAB RPS SSD Spoofing Challenge 2015 submission. In: INTERSPEECH 2015, 16 th Annual Conference of the International Speech Communication Associationth Annual Conference of the International Speech Communication Association. , Dresden, Germany (2015).