

The AHOLAB RPS SSD Spoofing Challenge 2015 submission

Jon Sanchez¹, Ibon Saratxaga¹, Inma Hernaez¹, Eva Navas¹, Daniel Erro^{1,2}

¹ Aholab Signal Processing Laboratory, University of the Basque Country UPV/EHU, Spain

² Ikerbasque, Basque Foundation for Science, Bilbao, Spain

ion@aholab.ehu.eus, ibon@aholab.ehu.eus, inma@aholab.ehu.eus, eva@aholab.ehu.eus, derro@aholab.ehu.eus

Abstract

This paper introduces the Synthetic Speech Detection system developed by Aholab for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015). The detector is a classifier based on Gaussian Mixture Models that are created using the Relative Phase Shift (RPS) transformation for the phase information. Different strategies have been evaluated: modeling the specific attacks using the information provided by the ASVspoof 2015 organizers, and modeling the vocoders possibly used in the spoofing signals, using data from previous works. The evaluation results show that attack specific models work for known attacks but they do not cope with the unknown attacks correctly. When using vocoder models build with other databases, the results suggest that the followed strategy do not take advantage of the available data and thus model adaptation should be explored.

Index Terms: synthetic speech detection, phase information, anti-spoofing

1. Introduction

In applications like access control, electronic transactions, or, in general, secure environments, biometric authentication has proven to be highly valuable [1]. But as the use of this kind of systems is spreading and becoming more and more common, the possibility of being spoofed by counterfeit biometric samples is also turning into a real concern.

In the particular case of the voice, Speaker Verification (SV) systems [2][3][4][5] have improved their performance, being useful and reliable. Speech signals can be easily and non-intrusively obtained, and they have obvious features that identify the speaker almost unmistakably. But in recent years creating a synthetic voice to deceive a speech driven biometric identification system has become feasible and relatively easy. Consequently, the concern on the security of these systems when attacked with this type of spoofing artificial voices is arising [6][7].

Two strategies have been mainly described in order to gain the ability to detect faked voices. The first one focuses on improving the verification system itself, working on the modeling technique or the parameters used. The aim is to get the synthetic impostors detected and blocked by the SV system, like it does with the human impostors [8][9][10]. The second strategy implies a separate synthetic speech detection (SSD) module, that implements specific parameters and detection techniques focused on the presumed differences of synthetic speech: interframe statistical similarities in some parameters [11][12], pitch variations [13], phase information

[11][14], temporal modulation[15], etc. It can be used before or after the SV system.

This second approach has been successfully used for vocoded speech detection in [16]. Since vocoders are used in most of the state-of-the art voice conversion and speech synthesis systems, and these techniques can be used to create a fake voice to attack a SV system, detection of vocoded signals can be an effective anti-spoofing countermeasure.

While module-based parameters (MFCC) are widely used in SSD, our system, described in [16], uses only relative phase shift (RPS) information to perform the detection. Since most popular vocoders do not use phase information, the phase differences between a natural signal and a counterfeit have proved to be relevant. The capability of this approach to cope with synthetic speech attacks was stated in [16]. The system performed well even when the spoofing signals were generated by unknown TTS systems.

In the Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015), independent SSD modules are evaluated [17]. Participants use specific spoofing detection tools on a provided database that contains different spoofing techniques such as speech synthesis and voice conversion. The performance of the different systems is assessed by the organization using standard metrics. Up to 6 submissions are permitted to the participants, with two different types: the ‘common’ submissions must use only information from the training database to create the system, and the ‘flexible’ submissions can make use of other databases. For each type, one of the submissions is designated as primary, while the others are designated as contrastive. We have submitted 4 different variations of our SSD system using different models.

In the next section of the paper the detection system is described, including the processing applied to the signals provided by the organizers and the models created for the different SSD variations. Then, the results of the different submissions are presented and discussed. Finally, some conclusions close the paper.

2. System Description

2.1. General Architecture

In ASVspoof 2015 the SSD system detailed in [16] is used. The system is a Gaussian mixture model (GMM) based binary classifier, whose purpose is to take a decision about the synthetic nature of the input speech signal. Figure 1 shows the general architecture of the SSD system.

The SSD system uses two GMM models for natural speech (λ_{human}) and synthetic speech (λ_{synth}). These models are

created during the training stage, using vectors of harmonic phase based parameters obtained applying the RPS transformation to the harmonic instantaneous phases.

To perform the synthetic speech detection task, the system tests a candidate vector sequence \mathbf{Y} of length N against both natural speech and synthetic speech models to get the corresponding likelihood values $p(\mathbf{Y}|\lambda_{human})$ and $p(\mathbf{Y}|\lambda_{synth})$.

$$\log p(\mathbf{Y}|\lambda) = \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{y}_n|\lambda) \quad (1)$$

$$\Lambda(\mathbf{Y}) = \log p(\mathbf{Y}|\lambda_{human}) - \log p(\mathbf{Y}|\lambda_{synth}) \quad (2)$$

Then according to (2) the log likelihood ratio Λ is calculated, taking the candidate as human if it exceeds a certain decision threshold θ which will be set to the Equal Error Rate (EER) point in the experiments. For the ASVspoof 2015 Challenge, the Λ ratios have been submitted for every input signal.

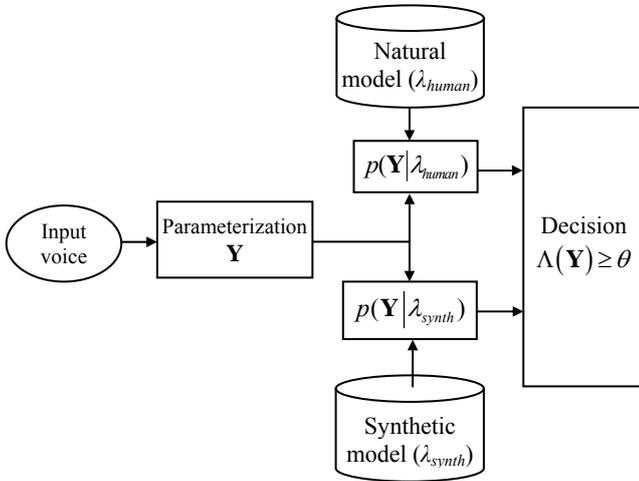


Figure 1: SSD system structure.

2.2. Signal pre-processing

Previous to the parameterization step, the signals are downsampled to 8 kHz, in order to limit the computational load, and their DC component is filtered out. Also, the polarity of the signals is homogenized [18], since the RPS parameterization is highly sensitive to polarity changes.

2.3. RPS Parameterization

The speech signals are parameterized using RPS parameters. The RPS is a representation for the harmonic phase information described in [19]. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency.

$$h(t) = \sum_{k=1}^N A_k \cos(\varphi_k(t)) \quad \varphi_k(t) = 2\pi k f_0 t + \theta_k \quad (3)$$

In (3), N is the number of bands, A_k are the amplitudes, $\varphi_k(t)$ the instantaneous phase, f_0 the pitch or fundamental frequency and θ_k 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 $\varphi_0 = \theta$.

$$\psi_k(t_a) = \varphi_k(t_a) - k\varphi_1(t_a) \quad (4)$$

Equation (4) 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 $[-\pi, \pi]$ 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 [20], have produced good results in other tasks where statistical modelling is used, such as ASR [20], Speaker Recognition [21] and also Synthetic Speech Detection [16] 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 $\Delta\Delta$ values are calculated.

For the experiments, the speech signals are windowed every 10 ms (using hamming windows of a length of 3 pitch periods) and the RPSs are calculated from the Fourier spectrum only for voiced frames. Then the DCT-mel-RPS parameterization is applied to every frame and the averaged value of the slope of the unwrapped RPS values is also included which leads to a total of 63 phase-based parameters.

2.4. Modelling

For this challenge 4 different model sets are tested, using both data from previous works and the database provided by the organization. Each set consists on a natural speech model (λ_{human}) and a spoofing speech model (λ_{synth}).

The first two sets (designated M1 and M2) have been trained using the human and spoofing signals provided by the organization. The signals have been used directly to elaborate the human and synthetic models. Thus, the synthetic model captures the specific features of the known spoofing signals.

The human and spoofing models used for the first submission, designated M1, are generated using only the training part of the database provided by the organization. The amount of signals used is detailed in Table 1. The spoofing methods included are:

- Two implementations of voice conversion using STRAIGHT [22].
- Voice conversion using MLSA [23].
- Two implementations of speech synthesis of adapted voices, using STRAIGHT.

The model set is trained with 1024 Gaussian mixtures. It fulfills the mandatory conditions for the primary common submission.

The model set M2 is created following the same approach as M1, but uses all the available data: both train and develop databases, with a total of 7247 genuine signals and 62500 spoofing, using the same methods that are found in M1. The models are trained with 1024 Gaussian mixtures and they are used for the primary flexible submission.

For the M3 model set our multivocoder model created in [16] is used. Unlike M1 and M2, M3 is aimed to model the vocoded speech instead of the specific attack technique or algorithm that will be unknown in realistic spoofing scenarios.

Multivocoder model set was created using the WSJ database [24]. The human model is created using a subset of the WSJ database containing 8599 natural signals from 283 speakers. The synthetic signals are created by means of copy-synthesis of the natural ones, using three different vocoders: MLSA, STRAIGHT and AHOCODER [25] [26], obtaining 25797 synthetic signals that are used to create the spoofing speech model. The signals provided by the ASVspoofer 2015 organizers are not used at all to train the M3 model set. The models are trained with 512 Gaussian mixtures, and the M3 set is used as first flexible contrastive submission.

The last model set (M4) mixes two different strategies: using the provided spoofing samples to improve detection of known attacks, and using the multivocoder approach to improve the detection of unknown attacks. Consequently, the models are created using the provided training set (used for M1) together with the WSJ signals used in M3, summing up 12349 genuine and 38422 spoofing signals. The model set is trained with 1024 Gaussian mixtures and used as a second contrastive flexible submission.

Table 1. Signals used to train the models classified by vocoder and attack method: Voice conversion (VC), adapted speech synthesis (SS), copy-synthesis (CS)

	M1	M2	M3	M4
Natural	3750	7247	8599	12349
VC STRAIGHT	5050	12500	-	5050
VC MLSA	2525	12500	-	2525
SS STRAIGHT	5050	12500	-	5050
CS STRAIGHT	-	-	8599	8599
CS MLSA	-	-	8599	8599
CS AHOCODER	-	-	8599	8599
Spoofing Total	12625	62500	25797	38422

3. Results

The evaluation dataset provided consists of 9404 genuine speech samples and 184000 spoofing ones, detailed in table 2. Half of the counterfeit signals were created using the same five methods that were known, that is, they were already used for the train and development sets (voice conversion using the vocoders STRAIGHT and MLSA, and speech synthesis of adapted voices, using STRAIGHT). The other half of the evaluation database was created using five unknown attacking techniques: four STRAIGHT based voice conversion algorithms and one new speech synthesizer, Mary TTS [27], which does not use any vocoder.

Table 2. Signal amounts used to test the models.

Subset	Characteristics	Known	Amount of signals
N	Natural	Yes	9404
S1	VC/STRAIGHT	Yes	18400
S2	VC/STRAIGHT	Yes	18400
S3	SS/STRAIGHT	Yes	18400
S4	SS/STRAIGHT	Yes	18400
S5	VC/MLSA	Yes	18400
S6	VC/STRAIGHT	No	18400
S7	VC/STRAIGHT	No	18400
S8	VC/STRAIGHT	No	18400

S9	VC/STRAIGHT	No	18400
S10	Vocoderless	No	18400

The following results were obtained by computing the EER with the submitted likelihoods for each test signal.

Table 3. EER (%) of the 4 Aholab submissions.

Model	Known attacks	Unknown attacks	All attacks
M1 (Common Primary)	0.210	8.883	4.547
M2 (Flexible primary)	0.154	8.918	4.536
M3 (Flexible Contrastive 1)	9.845	17.371	13.608
M4 (Flexible Contrastive 2)	2.042	11.291	6.667

Figures in table 3 show that both M1 and M2 model sets perform reasonably well, with EER values below 0.25%, when coping with signals generated with attack methods that were present in the models. But the error with unknown spoofing attacks rises to figures near 9%. Several factors explain this degradation of the performance:

- Different attacks types can show very different RPS features, thus modeling specific attacks does not assure coverage of unknown ones. In fact, the bad error rate obtained for Mary TTS completely burdens the averaged performance of the system. Mary TTS is based in unit selection where no vocoder is used, and this kind of system was out of the domain of the training material.
- Harmonic analysis necessary to obtain the RPS features requires a certain quality in the signals. Some attacking methods can degrade the quality to a point that makes the RPS parameterization incorrect.
- The simple GMM classifier used has no further mechanism to detect unknown attacks.

The case of the M3 model requires a separate analysis: it was created using a completely different database, not including the provided test or development database signals. Therefore, every signal in the test was unknown, even those designated as 'known attacks'. The performance of the system with this model set is poor, with EER values near 10% and beyond. There are some relevant points that are worth indicating:

- The generalization capability of the multivocoder models that was reviewed in [16] did not work with the provided evaluation signals.
- Internal experiments show that the score of the human signals of the ASVspoofer 2015 database when tested with the M3 model set are unexpectedly low. This suggests an underlying difference in the human signals from both databases. This difference can be due to recording conditions modifying the phase structure and it needs further study.
- Most spoofing attacks in the evaluation set were generated using voice conversion systems. The capability of the system to detect synthetic speech was well established in [16], where the performance of these

same models was remarkable for vocoder-based unknown speech synthesis. But the performance of the system on voice conversion spoofing had not been tested before, and it requires further study.

- Again, the presence of the Mary TTS synthetic speech in the test set is relevant, since the system is not capable of detecting speech synthesis algorithms that do not make use of vocoders.
- In the last few years, several versions of the MLSA and STRAIGHT vocoders have been developed that in practice perform like different vocoders in terms of phase, thus raising the error rate.

The results of the detection performed with the M4 model set are coherent with a design that models both vocoders like in M3 and specific attacks like in M1 and M2. The EER is between the values performed with M3 and those from M1 and M2.

Table 4. Results of every participant in the ASVspoof 2015 challenge.

Team	Known attacks	Unknown attacks	All attacks
A	0.408	2.013	1.211
B	0.008	3.922	1.965
C	0.058	4.998	2.528
D	0.003	5.231	2.617
E	0.041	5.347	2.694
F	0.358	6.078	3.218
G	0.405	6.247	3.326
H	0.67	6.041	3.355
I	0.005	7.447	3.726
J	0.025	8.168	4.097
K	0.21	8.883	4.547
L	0.412	13.026	6.719
M	8.528	20.253	14.391
N	7.874	21.262	14.568
O	17.723	19.929	18.826
P	21.206	21.831	21.518

Table 4 shows the Primary Common Submission result for every participant in ASVspoof 2015, being Aholab the one designated ‘K’. It’s worth pointing that the rank is elaborated in the base of the ‘All attacks’ column, but those really decisive, because of the greater difficulty of the task, or the balance of the known and unknown attacks in the test database, are the unknown ones. In the Aholab case, those have been particularly burdening, and further research is necessary to get a generalization capable system, mostly in the detection of signals created with voice conversion techniques.

4. Conclusions

The Aholab submission to the ASVspoof 2015 Automatic Speaker Verification Spoofing and Countermeasures Challenge is a binary classifier based on human and synthetic GM models of DCT-mel-RPS parameters. Two different strategies have been tested: attack modeling using the training

signals provided by the ASVspoof 2015 organizers, and vocoder modeling using data from previous works.

With models of specific attacks, the proposed architecture got promising results with signals similar to those used to create the models. The unknown attack performance is largely biased by Mary-TTS synthesizer, but it seems to be worse than that from the known attacks. This can be explained based on the basic GMM classification system and the RPS parameters modeling the specific attacks instead of the underlying vocoder.

The performance of the system is modest for models created to detect vocoded speech, in contrast to our results in previous works. We have found that these models perform particularly poorly with human signals from the ASVspoof 2015 Challenge database, probably due to phase treatment differences between the WSJ signals used to create the model and the evaluation signals. Some model adaptation technique should be applied to solve this issue. Without more detailed results, other reasons that probably contribute to this high error rate are the inclusion of a unit selection system among those to be tested, and the presence of voice conversion attacks, which had not been previously evaluated.

5. Acknowledgements

This work has been partially supported by the Basque Government (Ber2Tek Project, IE12-333) and the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03), with FEDER (Fondo Europeo de Desarrollo Regional) support.

6. References

- [1] A. K. Jain, A. Ross, and S. Pankanti, “Biometrics: A Tool for Information Security,” *IEEE Trans. Inf. Forensics Secur.*, vol. 1, no. 2, pp. 125–143, Jun. 2006.
- [2] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoust.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.
- [3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, Jan. 2000.
- [4] J. P. Campbell, “Speaker recognition: a tutorial,” *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.
- [5] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.
- [6] N. W. D. Evans, T. Kinnunen, and J. Yamagishi, “Spoofing and countermeasures for automatic speaker verification,” in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013.
- [7] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, “Spoofing and countermeasures for speaker verification: A survey,” *Speech Commun.*, vol. 66, pp. 130–153, Feb. 2015.
- [8] T. Masuko, K. Tokuda, and T. Kobayashi, “Imposture Using Synthetic Speech Against Speaker Verification Based On Spectrum And Pitch,” in *ICSLP*, 2000, pp. 302–305.
- [9] Z. Kons and H. Aronowitz, “Voice Transformation-Based Spoofing of Text-Dependent Speaker Verification Systems,” in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, 2013, pp. 945–949.
- [10] T. Kinnunen, 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 *2012 IEEE International Conference on*

- Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4401–4404.
- [11] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, “Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech,” *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 20, no. 8, pp. 2280–2290, Oct. 2012.
- [12] F. Alegre, A. Amehraye, and N. Evans, “Spoofing countermeasures to protect automatic speaker verification from voice conversion,” in *ICASSP*, 2013, pp. 3068–3072.
- [13] B. Steward, P. L. De Leon, and J. Yamagishi, “Synthetic speech discrimination using pitch pattern statistics derived from image analysis,” in *Interspeech*, 2012, pp. 370–373.
- [14] Z. Wu, E. S. Chng, and H. Li, “Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition,” *Interspeech*, pp. 2–5, 2012.
- [15] Z. Wu, X. Xiao, E. S. Chng, and H. Li, “Synthetic speech detection using temporal modulation feature,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7234–7238.
- [16] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, “Towards a Universal Synthetic Speech Spoofing Detection using Phase Information,” *IEEE Trans. Inf. Forensics Secur.*, vol. PP, no. 99, pp. 1–1, 2015.
- [17] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, “ASVspoof 2015 : Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan,” 2014. [Online]. Available: <http://www.spoofingchallenge.org/asvSpoof.pdf>.
- [18] I. Saratxaga, D. Erro, I. Hernaez, I. Sainz, and E. Navas, “Use of harmonic phase information for polarity detection in speech signals,” in *Interspeech*, 2009, pp. 1075 – 1078.
- [19] I. Saratxaga, I. Hernaez, D. Erro, E. Navas, and J. Sanchez, “Simple representation of signal phase for harmonic speech models,” *Electron. Lett.*, vol. 45, no. 7, p. 381, 2009.
- [20] I. Saratxaga, I. Hernaez, I. Odriozola, E. Navas, I. Luengo, and D. Erro, “Using harmonic phase information to improve ASR rate,” in *Proc. Interspeech 2010*, 2010, pp. 1185 – 1188.
- [21] I. Hernaez, I. Saratxaga, J. Sanchez, E. Navas, and I. Luengo, “Use of The Harmonic Phase in Speaker Recognition,” in *Interspeech*, 2011, pp. 2757–2760.
- [22] H. Zen, T. Toda, N. Nakamura, and K. Tokuda, “Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005,” *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.
- [23] T. Yoshimura, K. Tokuda, T. Kobayashi, T. Masuko, and T. Kitamura, “Simultaneous Modeling Of Spectrum, Pitch And Duration In HMM-Based Speech Synthesis,” in *Eurospeech*, 1999, pp. 2347–2350.
- [24] D. B. Paul and J. M. Baker, “The design for the wall street journal-based CSR corpus,” in *Proceedings of the workshop on Speech and Natural Language - HLT '91*, 1992, p. 357.
- [25] D. Erro, I. Sainz, E. Navas, and I. Hernaez, “Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis,” *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 2, pp. 184–194, Apr. 2014.
- [26] D. Erro, I. Sainz, E. Navas, and I. Hernaez, “Improved HNM-Based Vocoder for Statistical Synthesizers,” in *Interspeech*, 2011, pp. 1809 – 1812.
- [27] “MaryTTS – Introduction.” [Online]. Available: <http://mary.dfki.de/>. [Accessed: 09-Mar-2015].