An assessment of automatic speaker verification vulnerabilities to replay spoofing attacks

Artur Janicki, Federico Alegre and Nicholas Evans

Review rebuttal

First, we wish to thank the reviewers for their comments. We have done everything reasonably possible to address their concerns. As outlined in detail below, this has resulted in substantial modifications to the manuscript. We feel that these changes have helped to improve the quality and hope that they meet with the reviewers' approval.

We do, however, wish to highlight that two of the four reviewers have criticised the authors for their having not cited work that was neither published nor in the open domain at the time of submission. We also note that this criticism forms a major component of one reviewer's feedback. Our article was submitted in July 2015 whereas most of the recommended citations are from as late as September 2015. While we appreciate the need for an IEEE T-IFS paper to be up to date, we could obviously not have foreseen work published subsequent to submission; the manuscript was up to date at the time of submission.

The authors have consulted with the Associate Editor on this issue. Once again, while the authors agree that published work must be up to date, this is to be judged by the research status at the time of submission. While the authors can, and have added the suggested references, more major criticism made in light of work published subsequent to submission is deemed to be unfair. This includes the suggestion to repeat the work with new databases published long after submission.

We thank the reviewers kindly in advance for their attention to this position when considering the revision.

Except for the above, we have done all we can reasonable do to address the reviewers' comments and suggestions. We provide below our detailed responses to the reviewers' remarks and highlight the resulting modifications.

Artur Janicki, Federico Alegre and Nicholas Evans

Reviewer: 4

The authors report a very thorough assessment of automatic speaker verification vulnerabilities to replay spoofing attacks. Th replay attack is contrasted to other popular types of attacks: voice transformation and speech synthesis.

The paper is well written, reviews well previous works and evaluates thoroughly the impact of replay attacks.

The results indicate that replay attacks should be taken more seriously and support the importance of authentication conditions such as prompted pass-phrases.

We thank the reviewer for their kind appreciation. We are pleased that the intended contribution, focus and scope of our article is reflected so precisely in this review.

Reviewer: 3

This reviewer requires a number of relatively minor issues to be fixed before publication. The authors fully agree with all but one these issues and have accordingly made a number of edits to the paper in order to address them.

While one that we entirely understand, the remaining issue regarding evaluation is something which we feel is outside of the scope of this contribution; we feel that the current evaluation metric is adequate given the intended contributions of this work, even if we fully acknowledge that future work is needed to address such issues in the future.

I find some issues that should be fixed before publication:

- In eq. (6) mic and spk should be in \mathrm, otherwise it looks like mic(t)=m x i x c(t).

Thank you - now corrected.

- The authors should include some qualitative explanation of why LBP should help to detect replay attack better than the original MFCC.

Thank you for this helpful comment which has caused us to reflect further on specifically why the LBP feature has potential, especially for replay detection. We acknowledge that such an explanation was both missing and an essential addition to the paper.

The initial capture of replay attacks entails the convolution of a speech signal with the impulse response of an acoustic environment and a microphone. Replaying then entails convolution with the impulse response of a loudspeaker and then another acoustic environment. Newly added Figure 4 illustrates acoustic environment impulse responses. Compared to Figure 3, which

illustrates the same for loudspeaker responses, the significant distortion lasts for durations in the order of 100s of milliseconds. As a result, the convolutive effects of combined recording and replaying impact upon the speech signal over an interval of duration in excess of the window period used typically in MFCC extraction. It is also in excess of that implied by the application of conventional dynamic features. It is thus reasonable to assume that any feature which aims to capture these artefacts, must be extracted over a similar, extended duration. This explains the reduction in the time resolution.

If we now consider the frequency response instead of the time response, then we see from Figures 3 and 4 that disturbances in frequency are also significant and different for microphone and acoustic environment responses. The frequency resolution of LBP analysis is significantly greater than that of a typical MFCC filterbank. This supports the capture of replay characteristics and hence explains the frequency resolution.

Certainly the LBP feature is not the only candidate, but one which we have used previously and found to work reasonably well in this case - at least it outperforms the previously reported approach to replay detection. We have added new Figure 4 and substantial new text to Section IV-D in order to include this justification and thank the reviewer for the suggestion.

- Also a more detailed explanation of how de LBP is computed, in figure 2, there is a uniform operator, at first I thought that they were sampling from the uniform distribution but then after reading reference [29], I saw that it is something different. Include a short explanation of what is this uniform operator and why it is needed.

We have added a suitably short explanation in Section III.C.

- In the experiment section, why do you use 8conv-1conv if you enroll with only 1 random segments, why don't you use the standard 1conv-1conv?

This choice is for consistency with our previous work, results from which are included in the paper. The use of data from the 8conv-1conv condition stems from the need for speaker training data used to implement speech synthesis and voice conversion. We have introduced modest changes and additional text in Section IV.C to explain this setup. We stress that this choice is extremely unlikely to have any meaningfully significant impact on the findings presented.

- In Table II you have EERS of 80, 90%, but EER cannot be larger than 50%, if you have EER=90% you just change the sign of your scores and obtain EER=10%, you should check what happened there.

We fully understand this query. Of course it is initially counter-intuitive to report EERs in excess of 50%. One may readily reduce the EER by reversing the sign of the scores as the reviewer suggests. However, this would render the system entirely vulnerable to naive, zero-effort impostor attacks, hence why the reporting of EERs in excess of 50% is common in the literature.

This is a recurring issue with all the current work in spoofing and countermeasures for speaker verification which considers especially effective spoofing attacks. Since the use of countermeasures essentially brings the EER well under 50%, we did not investigate two-threshold solutions which would provide one solution to what is really an evaluation issue.

While similarly high values are widely reported in the literature, we have added a note to the caption of Table II in Section V.A to explain the high EERs reported in the paper.

- In Section V.C you don't mention that the Spoofing detector increase the number of false rejections of the Speaker verification system. When the spoofing detector says that a true target is a spoof the number of misses of the SV system increases. So to measure the performance of the combination of Spoofing detector+SV, I think that we should measure, the decrement of FA and the increment of FR for a given SV threshold that we setup using only clean trials, for example the EER threshold.
However, in table III you are measuring decrements of FA in the points where FR= original EER but that means that for each case the FA is measured at different thresholds, so they are not completely comparable.

The reviewer raises another important issue related to evaluation and one which is again recurrent in the literature. While we do not disagree with their position, there are currently so many different approaches to assess the impact of spoofing and the performance of countermeasures that everyone has a different opinion. An unfortunate but hopefully understandable consequence is that we cannot satisfy everybody.

Since we want to understand the relative threat of replay attacks compared to those of voice conversion and speech synthesis, we are satisfied in the choice of metric, namely the EER and SFAR, both of which are common in the literature and, together, will hopefully satisfy most readers. We have nonetheless added a reference to the SFAR in Section V.C. This was missing in the original manuscript.

We do fully agree with the reviewer, however, that the metrics used in this type of research need more thought. Given the above arguments, we hope that the reviewer will agree with our position that this work is outside the scope of the present paper.

Reviewer: 2

Reviewer 2 is principally concerned with countermeasure performance and generalisation. The authors feel that the two issues, which concentrate on a small component of the contribution, are now fully and satisfactorily addressed, either through modifications introduced in the revision, or through the presented argumentation. Two of the reviewer's suggested references were already included in the original manuscript. Three more, published after the submission of our manuscript have since been added to the revision. We respond to each comment below.

I feel that their current representation is unsatisfactory for a number of reasons:
1) The countermeasures give moderate improvements for two replay environments out of three, while for the last one (anechoic chamber) the effects are almost negligible.

As illustrated in Table III, the proposed LBP countermeasure is effective in reducing the EER from 30.3% to 9.6% for an office environment, and from 24.5% to 7.0% for a corridor environment. Even if these error rates remain considerably higher than the baseline (no spoofing) EER of 3% for the IV-PLDA system, these are significant improvements. They furthermore outperform the baseline FDD countermeasure which delivers EERs of 13.6% and 11.3%.

As the reviewer highlights, however, the countermeasures is ineffective for the anechoic chamber environment. This being a seemingly negative result, it is not without value. The result shows that the loudspeaker effects are not sufficient on their own in order to detect replay and that the countermeasures are essentially relying upon the detection of distortion introduced by the acoustic environment. This is a useful finding since it highlights the limitations of the new

countermeasure while also identifying opportunities to improve countermeasure performance through further work.

We accept that this argument was entirely missing in the initial submission and have adjusted the presentation of results in Section V-C and also VI accordingly. We thank the reviewer for this remark which has certainly helped us to improve the paper.

2) The countermeasure with the best performance is based on Local Binary Patterns (LBP). While the authors have a number of papers, where LBP was successfully applied for various anti-spoofing tasks, LBP generalization ability to new domains raises some questions. For example, during a large-scale speaker anti-spoofing challenge held earlier this year [4] only one team out of more than ten used LBP features (derived from the same LFCC coefficient as used here) and LBP performance was more than ten times worse compared to the top submissions [5]. It is not unique to LBP-based classifiers. There are other systems [6] that also failed to generalize, so it is not unique to LBP classifiers, but rather it is a common problem of the field.

Generalisation is an important issue and one which the authors have raised repeatedly in several publications, even proposing the first truly generalised, one-class classifier in [Alegre, BTAS 2013]. While the emphasis of this work was on the classifier, rather than features, those used in [Alegre, BTAS 2013] were LBP features, which were shown to work well, even when applied to entirely different spoofing attacks, examples of which were not used for optimisation.

We note that the three papers to which the reviewer refers were published five weeks after the submission of our article. With the review result coming in the middle of November, we trust the reviewer will appreciate that we have not had sufficient time to investigate precisely why LBP seems not to have worked well on the ASVspoof database.

Suffice to say, however, we are sufficiently confident in our own past work that our research with LBP is continuing, even in the face of other researchers‘ less convincing results. These, we believe, are not sufficient evidence to show that the generalisation of LBP features is especially poorer than other, alternative features.

For now, then, we can only speculate as to why LBP was shown to perform relatively poorly on a different evaluation with different data and with different systems. This could be anything related to lack of optimisation, lack of data, differences in data duration, other systems and database nuances, etc. We suspect that, as a more complicated feature, LBP might be more data-hungry than other features, be more noisy and thus require more careful optimisation and adaptation. If this is the case, then LBP may still offer potential for generalisation.

To help appease the reviewer on this issue, we have expanded the presentation of LBP and treatment of generalisation in Sections III.C, IV.D and VI. This new addition references the work in [Sahidullah, Kinnunen and Hanilci, 2015] which was published five weeks after our original submission. We sincerely hope that these modifications meet with the reviewer’s approval.

Unfortunately, we can do no more to address this concern here. Since the reviewer agrees that generalisation is a common problem, not one specific to the use of LBP, since our own work shows that LBP generalises reasonably well, since we show significant improvements in replay detection performance, and since the new countermeasure is only one component of this contribution, we feel that the issue is appropriately addressed in the paper..

3) In the practical scenario, we do not know a type of spoofing attack, so it is vital for a countermeasure to have a good generalization abilities across all of them. Yet, the authors

present the results of their countermeasures for only replay attacks and completely ignore speech synthesis and voice conversion attacks.

This issue is essentially the same as the reviewer's second issue, above. Aside from the edits made to Sections III.C, IV.D and VI in response to this issue, we are sincerely sorry to have to respond negatively to this further request, one which we feel falls outside the scope of our intended contribution.

The paper aims to assess the threat of replay attacks, to demonstrate the importance of their consideration in the future and to show the potential of, but difficulty to develop countermeasures for replay spoofing attacks. While the authors fully appreciate the reviewer's rightful concerns regards generalisation, a contribution in this direction was not intended.

We stress that generalisation should not be considered a strictly necessary characteristic. A countermeasure which protects against one broad form of spoofing attacks is not necessarily useless if it does not protect well against others. It is of course desirable, but not strictly necessary. Even if one can never know the nature of spoofing attack a priori, it is difficult to imagine a single classifier ever being sufficient on its own to detect the full spectrum of spoofing attacks, whatever that might be. Ultimately, a practical spoofing countermeasure may well involve a bank of fused, weak classifiers, each offering potential to detect different forms of attack. Our new countermeasure would then be one suitable candidate.

We also stress that our past work, which is cited in the paper, has evaluated a variant of the LBP countermeasure for the detection of voice conversion and speech synthesis. Even if it is not reported directly in this paper, we are confident that the countermeasures offers potential for generalisation.

In summary, given the contributions and focus of this paper, we do not consider the issue of generalisation as a necessity, nor a weakness in this contribution. It is simply not what the paper is about.

Even given all of this, robust detection of replay attacks is still an important and open question and this paper, after modifications, has a potential to make a valuable contribution to the field. I would like the authors to extend their experiments to evaluate generalization abilities and reasonable performance of their countermeasures across different domains and to present the detection performance for other spoofing types. Personally, I'm aware of at least one open-access corpus that can help them [3], but I do not restrict them by any means to try other ideas.

The reviewer is asking us to fundamentally change the paper, whereas we do not entirely agree with their concerns which are contested above. We thus consider the request to extend the work, in directions which would fundamentally change the intended contributions, as not being justified.

We note that the work in [Ergunay, BTAS 2015] was published five weeks after the submission of our manuscript and, consequently, we consider the request to repeat all of our experimental work and to use it for further, additional experimentation to assess generalisation capabilities, to be unreasonable and unfair. We have nonetheless added a reference to the new database and brief discussion of it in Section VI.

Reviewer 1

A major component of this reviewer's criticism is unfair in relating to work published subsequent to submission. Other comments relate to fundamental misunderstandings. Unfortunately, others still are without foundation or sufficient justification or explanation which would allow us to address them. In general we take objection to the nature of the review which we find to be unbalanced, unjustified and hostile.

The authors were at a loss as to how to prepare a suitable response, let alone sensible modifications to the manuscript. Given the nature of the comments, we suspect that the reviewer will not change their opinion on the basis of our response. At least the authors are satisfied that all three major concerns are either irrelevant, stem from the reviewer's misunderstanding, or have been more than adequately addressed in the revision.

Before comments, the review would like to list several latest publications, which related to the contributions of the paper, and the database used in the work:

[1] Kucur Ergunay, Serife, Elie Khoury, Alexandros Lazaridis, and Sébastien Marcel. "On the Vulnerability of Speaker Verification to Realistic Voice Spoofing." In Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems 2015.
[2] Md Sahidullah, Tomi Kinnunen, Cemal Hanilçi, "A Comparison of Features for Synthetic Speech Detection", In Proc. Interspeech 2015
[3] Da Luo, Haojun Wu, and Jiwu Huang. "Audio recapture detection using deep learning." In Proc. IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), 2015
[4] Shiota, Sayaka, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification." In Proc. Interspeech 2015.
[5] Jakub Gałka, Marcin Grzywacz, and Rafał Samborski. "Playback attack detection for text-dependent speaker verification over telephone channels." Speech Communication 67 (2015): 143-153.

We note that three of these papers were published after submission. We thank the reviewer for their suggestions, in particular for the one paper published in March 2015 and one more published shortly before our submission in mid-July. We note, however, that these papers do not concern specifically the assessment of ASV systems in the face of replay attacks and so we do not agree that they are major omissions. Notwithstanding our general response as regards criticism related to new work published after the submission of our paper, citations to all of the above references have been added to the revised version of our paper.

Major concerns:

1. Novelty: the authors claim two contributions: 1) comparing the threat of replay attacks to those of speech synthesis and voice conversion; and 2) proposing two new countermeasures. The reviewer believes that both of them are not novel.
Regarding the first claimed contribution, in (Ergunay et al. 2015), an excellent comparison of replay, speech synthesis and voice conversion was presented using more realistic setups, and a database containing those attacks was also released. In contrast, this paper still uses an unrealistic way to generate replay attacks, and uses out-of-date techniques to generate speech synthesis and voice conversion attacks. There is no novelty comparing with (Ergunay et al. 2015)? Why not use the database described in (Ergunay et al. 2015) directly?

As per our arguments above, the basis of this criticism is considered wholly unfair. The work in Ergunay et al. 2015 was published in September 2015, five weeks after the submission of our paper. It is the authors' position that the work is novel, meeting entirely the requirements for inclusion in TIFS (which we have checked again). The paper assesses the threat of replay attacks, demonstrates the importance of their consideration in the future and shows the potential of, but difficulty to develop effective replay countermeasures. We consider these contributions to be novel, especially given that the recent ASVspoof evaluation entirely omitted replay attacks.

Unfortunately, since the reviewer has not explained why the generation of replay attacks is 'unrealistic', the authors can do little to respond to this criticism. The reviewer may be referring to the fact that replay attacks are emulated, as opposed to having been genuinely recorded. If this is the case, we do not accept the criticism and note that the reviewer is the only one to have such concerns. In the case that the impulse responses are representative, as they are, then there is no difference between emulation / convolution and genuine re-recordings. This is basic signal processing. Emulation also permits a far broader assessment, including multiple replay scenarios, than would be practicable otherwise. This approach is therefore considered as an advantage, not a weakness. We have added some text to Section IV.D in case this is the reviewer's concern. If this is not the reviewer's concern, we do not know what it is.

We also do not accept the criticism that the techniques to generate speech synthesis and voice conversion are out-of-date. We stress that the approach to voice conversion is one of the best-performing approaches to speaker verification spoofing that we know of. This is supported by the literature, including independent work by the original authors of this spoofing attack, affiliated with a highly respected speaker verification laboratory (LIA). We note that this work has been cited 54 times (according to Google Scholar), by the likes of T. Kinnunen, H. Li, P. De Leon, M. Pucher, J. Yamagishi, Q. Jin, A. Toth, T. Schultz, A. Sizov, E. Khoury, Z. Wu and many others.

The reviewer also criticises us for having not used an open source toolkit for speech synthesis, citing a few examples - including the one we actually used, namely HTS. We note that the very same algorithm was used in the work of Ergunay et al. 2015 referred to by the reviewer. Our use of HTS is stated very clearly in the paper so we are at a loss as to why the reviewer has misunderstood this. We refer again to this criticism below.

Regarding the second claimed contribution, the first countermeasure, the far-field channel detection countermeasure, was originally proposed in [28], and the LBP countermeasure was originally proposed in [30]. Further more, (Sahidullah et al. 2015) conducted a systematic analysis of LBP-based countermeasures, and showed that LBP-based countermeasures did not well on all voice conversion and speech synthesis attacks in comparison with other feature-based countermeasures. Why do the authors still believe it will work to detect voice conversion, speech synthesis and replay attacks? The authors should compare with the countermeasures described in (Sahidullah et al. 2015) and justify.

We are frustrated with the unobjectivity of this remark. The authors believe that the LBP-based countermeasure has potential since the results in the paper show that it does. Other published work has also demonstrated the potential. This work is cited in the paper and cited by other authors. We hence do not understand why the reviewer questions why we think it will work.

We cannot explain other authors' work, nor why they do not observe the same results as us. Possibly it is due to their implementation, differences in their optimisation, or differences in the database. We object to the inordinate emphasis placed by this reviewer on a single paper presented by other authors and note that the work in Sahidullah, Kinnunen and Hanilci, 2015 was published in September 2015, five weeks after submission. As per our arguments above, this criticism is considered unfair and is unjustified.

All we can do to respond to this criticism is to better justify why LBP works, as it is proven to do in this very paper. This explanation has been added to the paper in Section IV.D.

2. Spoofing materials: Firstly, the comparison is meaningless. The speech synthesis trials are generated using arbitrary texts, voice conversion modifies the impostor speech, and replay uses the target speech. Does it mean the language contents of three attacks are totally different?? If so, how can the manuscript give meaningful conclusions when the language contents are different??

Since the trials are long-duration, the phone content in each trial is effectively normalised and thus will have only a negligible effect on the results and findings. We have added a note to the paper in Section IV.C to explain this.

We stress, however, that replay attacks were repeated in each environment with the same phone content as that of the original, target speech - this is obvious given the study of replay. Accordingly, there is no issue with phone variation.

Secondly, the toolkits to generate both speech synthesis and voice conversion spoofing materials are not state of the art. It will give miss-leading conclusions. The reviewer suggests to use state-of-the-art unit-selection-based (e.g. MaryTTS), parametric-based synthesis systems, and state-of-the-art voice conversion systems (e.g. Festvox). The so-called Gaussian-dependent filtering technique is never cited or compared by voice conversion community. The authors should justify it is a valid technique to do voice conversion. Thirdly, the replay attacks should be generated by using more realistic setups, not just simply doing a simulation.

These criticisms have been partly addressed above.

We do not accept the criticism and feel that the reviewer has fundamentally misunderstood the work. The fact that the voice conversion algorithm is not referenced by the voice conversion community is irrelevant. The metrics used by the voice conversion community are not at all the same as those used in the study of speaker verification spoofing. The goals are different. While the approach to voice conversion may well be inferior as far as the conventional voice conversion community is concerned, it is at the state of the art as far as the study of ASV spoofing is concerned. This position is described in new text added to Section II-B.

The criticism as regards the approach to speech synthesis could possibly prove to have some merit given the results of the recent ASVspoof evaluation. The results of this, however, were published after our work was performed and submitted. The reviewer should nonetheless know that there is no evidence that the MaryTTS system is more effective as a spoofing attack than other systems. The reviewer may not know that the only speech synthesis training material in the ASVspoof database was generated by statistical parametric speech synthesis - there was no training data for unit selection speech synthesis. So, while the attacks generated by a unit selection algorithm proved to be the difficult to detect in the evaluation set, these results are not sufficient on their own to suggest that unit selection is a greater threat. Had training data for a unit selection speech synthesis algorithm been provided, then perhaps unit selection and statistical parametric speech synthesis spoofing attacks would have been detected with the same ease. The aspect of the reviewer's criticism is thus without foundation. Our position as regards the use of HTS is described in new test added to Section II-A.

Even if we fundamentally object to the reviewer's criticism, their concerns are not related to the main contributions of this article. This is not to make a strict comparison of each form of attack, but more to demonstrate the need for greater attention to replay attacks. This position is stated clearly throughout the paper.

3. Baseline countermeasures: from the presentation of the related work and the experimental setups, it looks like the authors are unaware of the progress in this field. Recently, several techniques have been proposed to detect playback. The authors only mentioned literatures published several years ago. This year (2015), there are at least three published papers addressing the playback problem (see references listed above). Why not acknowledge these work and compare with them?

The contribution was up to date at the time of submission. The authors cannot foresee work published subsequently and requests for additional comparative experiments to work published subsequent to submission are unreasonable. We therefore do not accept this criticism.

The two works published prior to submission do not concern the reliability of ASV to replay spoofing and thus they are not major omissions.

Nonetheless, all are referenced in the revised version of the paper. Hopefully the reviewer will be satisfied with these additions.

Specific comments:

1. English writing: The authors should choose either British or American English, but please do not mix them.

We have corrected the one instance of American English we found.

2. Please expand the introduction to discuss the related work as listed above.

Notwithstanding our outline remarks above, we have added citations to all of the suggested reference.

3. [17] is a speaker adaptation paper, not a representative work for statistical parametric synthesis.

We have changed the reference.

4. The authors claim 'unit-selection approaches generally require large amounts of speaker-specific data to learn the mapping function', but this is not true. Smaller units (e.g. frame, state, half phone) can be defined to generate synthetic speech with a small amount of training data. Please check literatures regarding the units for unit-selection synthesis.

We have adjusted the text in Section II.A accordingly.

5. In Section II.A, STRAIGHT is a vocoder. The reviewer believes Mel-cepstrum coefficients cannot be extracted by STRAIGHT itself. Please make it clear how exactly these features are extracted. What is the dimensionality for Mel-cepstrum coefficients? The output of STRAIGHT has three parts, spectral envelope, F0 and aperiodicities. It looks like the aperiodicities have been discarded. If so, how to reconstruct waveforms without aperiodicities? Please clarify.

We have adjusted the text. We regret that there is not sufficient space to explain fully the speech synthesis algorithm and have included a reference to [Yamagishi et al., IEEE TASLP, 2009] as a

reference for further information. Given the focus of this paper, we do not believe that additional detail is warranted.

6. What is the frame size to extract the vocoder parameters? What is the window size? Please give a citation to MLSA, STRAIGHT, mixed excitation.

The requested details and citations were added to Section II.A.

7. In Section II.B, the authors should justify that the Gaussian-dependent filtering technique is state of the art, and is widely used by the voice conversion community. A valid technique to generate spoofing materials is important to the conclusion.

This comment is addressed above. In summary, the original paper on Gaussian-dependent filtering has been referenced over 50 times by the speaker verification community. The fact that it has not attracted attention from the voice conversion community is irrelevant.

8. Eq. (4), what is the meaning of the formula? How many Gaussian components?

The equation demonstrates how the voice conversion filter parameters are derived from the tied asr and filter speaker models. As is the case for speech synthesis, the detail given to the description of this algorithm is deliberately brief. This is for space limitations and why we include 3 citations to the original and derived work. Given the focus of this paper, we do not agree that additional detail is warranted. To satisfy the reviewer's curiosity, we used 1024 Gaussian components and have stated this in the paper in Section IV.D.

9. Table I, why put a question mark? The effectiveness of replay is well studied in literatures. Please check the latest papers listed above and the survey paper [2].

This was deliberate to help illustrate the focus of the paper and the contribution. To satisfy the reviewer, we have accordingly updated Table I to illustrate the difference in the findings of past work. We have also made a modest update to the text of Section II-D.

10. Section II.D, before the comparison, the authors should assume an application scenario. In a call-centre application, how can replay maintain a conversation?

We deliberately avoided a restriction to a specific scenario in the original submission. The limitations of this work and the arguments for the approach taken have been addressed through our edits in response to several of this reviewer's remarks and are also acknowledged and discussed in the conclusions.

11. Section II.D, please also take (Ergunay et al. 2015) into account in the discussion.

(Ergunay et al. 2015) is now referenced in Sections II-A, V-B and discussed in Section VI. Since the work in (Ergunay et al. 2015) was performed in parallel, and published after the submission of our article, not before submission, this is the appropriate place for such discussion.

12. The review of past work is not complete. Please also discuss above papers in Section III.A.

References and appropriate discussion has been added to the paper in Sections I, III.A and IV. The reviewer must accept that 3 of their 5 suggestions were published after submission. Two of them do not relate to the study of ASV vulnerabilities and countermeasures and so the request to discuss all in Section III-A is not appropriate. All are nonetheless referenced in the revision.

13. The resolution of Fig. 2 is not good, and the font type is different.

We have improved the resolution of Fig. 2 and updated the font type.

14. Section III.C, the claim of the effectiveness of LBP-based countermeasure is not true. Please take (Sahidullah et al. 2015) into account in the discussion.

This is a duplicate issue. We have responded to it above.

15. Section IV.B, details for all the ASV systems are missing, and most of them do not have a reference!

a) how to train the GMM-UBM system? which criterion? how many Gaussians?

b) a reference to GSL system. Which SVM toolkit is used? What are the parameters and how to set them?

c) which database is used to train the covariance matrix in NAP??

d) a reference to the factor analysis model

d) which database is used to train the subspace in factor analysis?? The dimension of the subspaces?

e) dimension of subspaces in IV-PLDA?

f) why use different normalisation techniques for different systems??

References to all systems have been added to the revision. The coverage of a selection of speaker verification technologies was included so as to give some broadness to the study, namely to show that observations are broadly similar, whatever the speaker verification technology. The paper is not intended to analyse system specifics in detail. We have nonetheless expanded the detail marginally and explained why different normalisation strategies are used (they gave better results) but, given the focus of the paper, we do not agree that all the detail requested by the reviewer is warranted. The systems are standard, and well known in the community. Further details are not necessary and not critical to the contribution.

16. Details of the databases: how many speakers? sampling rate? telephone channel?

We stress that the NIST databases are the defacto databases for speaker verification research. They are very well known by the community and this detail is generally not included in the literature. We have nonetheless added brief details to Section IV.C. We do not agree that any more would be helpful.

17. Why only use male speakers in the study?? Speaker verification research usually uses both male and female. The authors need to justify this. For a complete study, both male and female speakers should be used!

This is not true and needs no justification. We refer the reviewer to one of their own suggested references in which only female speakers were included:

Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. "Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification." In Proc. Interspeech 2015.

There are plenty of others:

Patrick Kenny, “Bayesian speaker verification with heavy-tailed priors,” in Proc. Odyssey, 2010, p. 14-24.

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. A novel scheme for speaker recognition using a phonetically-aware deep neural network. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014), 2014, pp. 1695-1699.

Achintya Kumar Sarkar, Driss Matrouf, Pierre-Michel Bousquet, Jean-Francois Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification." Proc. Interspeech 2012.

18. The authors use NIST06 data as evaluation data, but also use it in the background data??? Do you mean the systems already know information about the evaluation data? It is quit confusing! Why not use Fisher, Switchboard corpora, etc?

The reviewer has misunderstood. We refer them to Section IV-C where this is explained clearly and unambiguously:

“Experiments are performed on the male subsets of the 2004, 2005, 2006 and 2008 NIST SRE datasets, which are from here on referred to as NIST’0x. The NIST’04 and NIST’08 datasets are used for UBM training. The NIST’05 dataset is used for development, whereas the NIST’06 dataset is used for evaluation; only results for the latter are reported here.”

This is a relatively typical setup and the one used in much of our previous work, e.g.:

F. Alegre, R. Vipperla, N. Evans, and B. Fauve, “On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals,” in Proc. 12th EUSIPCO, 2012.
F. Alegre, R. Vipperla, and N. Evans, “Spoofing countermeasures for the protection of automatic speaker recognition from attacks with artificial signals,” in Proc. 13th Interspeech, 2012.
F. Alegre, A. Asmaa, and N. Evans. "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns." IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS 2013). 2013.

The reviewer has not explained why we should have used the Fisher or Switchboard corpora. In any case, the protocol used is standard and acceptable within the community.

19. It is not clear how to divide development and evaluation sets.

We find the above paragraph, copied from the original submission, to be abundantly clear and unambiguous. Unfortunately the reviewer has not mentioned what aspects are unclear and so we can do nothing to address this.

20. What is the sampling rate and accent for EMIME corpus? How many speakers (male and female)? How to train the SI model using the EMIME corpus? Please spell out EMIME.

We have spelled out EMIME in the revision. The sampling rate is 22,050 Hz and with American English accents. Suitable references to further details have been added. More are not needed.

21. The authors should generate replay attacks using more realistic ways, not just simulation by claiming replay spoofing is easy to implement.

We do not agree with this remark and have addressed it in our response to this reviewer's Major Concern 1, which is a duplicate. We feel that the reviewer has misunderstood something here. The ease with which replay spoofing attacks can be implemented is specific to the attack itself and nothing to do with the argument for using emulated spoofing attacks instead of genuine re-recordings.

The reviewer has not explained why the current setup is unrealistic so we can do no more to respond to this issue.

22. It is quit confusing that speech synthesis used arbitrary text, while voice conversion used impostor trials. It implies the language contents are different, and how can the study get meaningful conclusions!

This issue has no impact on the work or findings. We have addressed it in our response to this reviewer's Major Concern 2.

23. In speech synthesis adaptation, how many transforms? how to estimate the transforms? which criterion?

Answers to these questions are provided in Section II-A and other cited work. For the same reasons presented several times above, given the focus of the paper and the intended contributions, we do not agree that this additional detail is warranted.

24. What are the pros of the database built here in comparison with the IDIAP AVspoof database?

The AVspoof database, published five weeks after the submission of our paper, is now referenced througout. Comments related to work published previously are unfair.

25. Why use a commercial voice cloning toolkit? There are many open-source toolkits implementing state-of-the-art techniques, such as MaryTTS, Festival, Festvox, HTS. Open-source toolkits also allow reproducible research.

We have used one of the very toolkits suggested by the reviewer, namely HTS. This is stated very clearly in Section II-A and now also in Section IV-D. This misunderstanding may account for some of the issues listed above.

26. Section IV.E, the description of the far-field channel detection countermeasure is not clear to the reviewer. What is the modulation? What is the signal envelope and how to compute it? How to downsample to 60 Hz and why?

This countermeasure was originally proposed by other authors in [Villalba and Lleida, ICCST, 2011]. The answers to all of these questions are provided in this original paper which is cited. Since they are not the contribution of the authors, since the original work is clearly cited and since there is not sufficient space to include further details in the present paper, we do not agree that they are warranted.

27. In LBP, why use LFCC? As shown in (Sahidullah et al, 2015), LFCC is not the best choice.

We have addressed this issue in our response to this reviewer's Major Concern 1 and Specific Comment 14. We do not accept the criticism.

28. Please describe the bayesian network classifier in detail for reproducible research. Which toolkit is used?

This is now identified in Section IV-E - we used the Weka toolkit.

29. Please describe the AdaBoost M1 meta classifier in detail. Which toolkit is used?

This is now identified and described in appropriate detail in Section IV-E. Again, this was implemented using the Weka toolkit.

30. Why two countermeasures use two different classifiers?

This is explained clearly in Section IV-E. These classifiers returned the best results in each case for an area-under-the-ROC metric. This is a classical procedure and why these two, different classifiers were selected.

31. The protocol is quit confusing. The ASV protocol has 10000 impostor trials, but the countermeasure protocol has 8112 replay trials. With different number of trials, how to integrate ASV and countermeasures?? The two experiments look like totally separate.

We accept that this was not sufficiently clear and is actually a mistake. We have adjusted the text in Sections IV.C and Section IV.E to better explain the protocol and find the new text to be greatly improved. The number 8,112 came from the 1,352 genuine recordings which were replayed in 6 different conditions in order to give 6 x 1,352 = 8,112 spoofing attacks. That number is not 6, but actually 9 including the anechoic condition giving 9 x 1,352 = 12,168 spoofing attacks. The text has been updated to reflect this correction. We thank the reviewer for this comment.

32. The reviewer does not think the results and conclusions are meaningful if above problems regarding the databases cannot be addressed.

We do not accept this criticism. The paper aims to assess the threat of replay attacks, to demonstrate the importance of their consideration in the future and to show the potential of, but difficulty to develop replay countermeasures. The database is entirely fit for this purpose. With one exception, the reviewer's concerns presented above have been fully addressed, either through sufficient argumentation and/or appropriate edits to the revision.

Unfortunately, the one remaining major concern regarding the 'unrealistic' means of generating replay attacks cannot be addressed since the reviewer does not explain why they think the emulation is unrealistic. As argued above, we maintain that this approach is a potential advantage, not a weakness, and defend its use for the experimental work presented in this paper.

Newly added citations

We present here a list of citations added to the revision as a result of the first-round review.

S. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, “On the vulnerability of speaker verification to realistic voice spoofing,” in Proc. IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2015), September 2015.

D. Luo, H. Wu, and J. Huang, “Audio recapture detection using deep learning,” in Proc. IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP 2015), July 2015, pp. 478–482.

S. Shiota, F. Villavicencio, J. Yamagishi, N. Ono, I. Echizen, and T. Matsui, “Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification,” in Proc. Interspeech 2015, September 2015, pp. 239–243.

J. Gałka, M. Grzywacz, and R. Samborski, “Playback attack detection for text-dependent speaker verification over telephone channels,” Speech Communication, vol. 67, pp. 143–153, 2015.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An update,” SIGKDD Explor. Newsl., vol. 11, no. 1, pp. 10–18, Nov. 2009.

K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), vol. 3, Istanbul, Turkey, June 2000, pp. 1315–1318.

H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT.” in Proc. MAVEBA, Firenze, Italy, 2001, pp. 59–64.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Incorporating a mixed excitation model and postfilter into HMM-based text-to-speech synthesis,” Systems and Computers in Japan, vol. 36, no. 12, pp. 43–50, 2005.

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” Digital Signal Processing, vol. 10, no. 1-3, pp. 19–41, 2000.

W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Processing Letters, vol. 13, pp. 308–311, 2006.

T. Ahonen, A. Hadid, and M. Pietikainen, “Face description with local binary patterns: Application to face recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037–2041, 2006.

H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013). IEEE, 2013, pp. 7962–7966.

M. Sahidullah, T. Kinnunen, and C. Hanilci, “A comparison of features for synthetic speech detection,” in Proc. Interspeech 2015, Dresden, Germany, September 2015.

J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, “Analysis of Speaker Adaptation Algorithms for HMM-based Speech Synthesis and a Constrained SMAPLR Adaptation Algorithm,” IEEE Transactions on Audio, Speech & Language Processing, vol. 17, no. 1, pp. 66–83, 2009.

S. Imai, K. Sumita, and C. Furuichi, “Mel log spectrum approximation (MLSA) filter for speech synthesis,” Electronics and Communications in Japan (Part I: Communications), vol. 66, no. 2, pp. 10–18, 1983.

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 Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2280–2290, 2012.

F. Alegre, A. Amehraye, and N. Evans, “Spoofing countermeasures to protect automatic speaker verification from voice conversion,” in Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2013.

F. Alegre, A. Amehraye, and N. Evans, “A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns,” in Proc. Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS), Washington DC, USA, 2013.