Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers

Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, Massimiliano Todisco

▶ **To cite this version:**

Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, Massimiliano Todisco. Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers. Interspeech 2020, Oct 2020, Shanghai, China. pp.1106-1110, 10.21437/Interspeech.2020-1844 . hal-03555611

HAL Id: hal-03555611
https://hal.science/hal-03555611

Submitted on 3 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spoofing Attack Detection using the Non-linear Fusion of Sub-band Classifiers

Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans and Massimiliano Todisco

EURECOM, Sophia Antipolis, France

lastname@eurecom.fr

Abstract

The threat of spoofing can pose a risk to the reliability of automatic speaker verification. Results from the bi-annual ASVspoof evaluations show that effective countermeasures demand front-ends designed specifically for the detection of spoofing artefacts. Given the diversity in spoofing attacks, ensemble methods are particularly effective. The work in this paper shows that a bank of very simple classifiers, each with a front-end tuned to the detection of different spoofing attacks and combined at the score level through non-linear fusion, can deliver superior performance than more sophisticated ensemble solutions that rely upon complex neural network architectures. Our comparatively simple approach outperforms all but 2 of the 48 systems submitted to the logical access condition of the most recent ASVspoof 2019 challenge.

Index Terms: spoofing; sub-band countermeasures; presentation attack detection; ASVspoof; speaker verification.

1. Introduction

A great deal of research in ASV anti-spoofing has focused on the design of specific front-ends tuned to capture artefacts that characterise manipulated or synthetic speech. The results of the ASVspoof 2019 challenge also show that reliable performance usually demands the fusion of scores derived from an ensemble of different front-ends. This observation suggests that no single front-end can detect reliably the full range of artefacts produced by different spoofing attacks.

There is evidence that spoofing artefacts lie at the sub-band level [1, 2, 3, 4, 5, 6, 7] and that these can only be detected reliably using front-ends that have high spectral resolutions in the same bands [8, 9]. This means that conventional cepstral processing may be detrimental to anti-spoofing performance in the sense that cepstral analysis averages information across the full spectrum, rather than emphasising information at the sub-band level. This in turn may explain why reliable performance is obtained only through the fusion of several systems, with similar performance not being achieved with single systems. Results from the ASVspoof 2019 challenge support this hypothesis. Although there are likely to be additional contributing factors, the top-performing fused-system submission achieved an equal error rate (EER) of 0.22 % whereas the same team's single-system submission achieved an EER of 11.40 %, some 52 times higher.

The observation that different front-ends are required to detect artefacts located within different sub-bands may mean that the usual approaches to fusion will be sub-optimal. This is principally because any single spoofing attack may only be detected reliably by a single countermeasure within an ensemble. In this case, linear approaches to score fusion may not exploit the complementarity of each countermeasure to their full potential; linear combinations of mostly non-informative scores may serve to dilute informative scores. Non-linear approaches to fusion may hence be better suited to such scenarios.

The work reported in this paper was designed to test these hypotheses, namely that: (i) an ensemble of relatively simple countermeasures, each tuned to the detection of artefacts in different sub-bands, may help to improve spoofing detection performance beyond what can be achieved through the fusion of different countermeasures operating at the full-band level; (ii) non-linear fusion may better exploit complementarity beyond what can be achieved with linear approaches. To the best of our knowledge, while some work has already demonstrated the benefit of ensemble methods, e.g. [10, 11, 12], none of the past work has investigated the reasons *why* they are beneficial, and neither have they explored ensemble methods in the context of fused, attack-optimised, sub-band front-ends.

2. Research hypotheses

To help illustrate the ideas explored in this work, we consider the hypothetical anti-spoofing example illustrated in Fig. 1. Plotted on each axis are the scores produced by two different spoofing countermeasures: CM_1 and CM_2. Countermeasure CM_1 is tuned to detect artefacts present within a lower sub-band, at 0-4 kHz for example. Countermeasure CM_2 is tuned to detect artefacts within a higher sub-band, at 4-8 kHz for example. Each point in the plot signifies the scores produced by each countermeasure for a set of utterances. Scores for bona fide (genuine / not spoofed) utterances are illustrated by green points (top-right). Also shown are scores for three different types of spoofing attacks: attack A_1, characterised by artefacts predominantly at low frequencies (blue points, top-left); attack A_2, characterised by artefacts at high frequencies (red points, bottom right); attack A_3, which exhibits artefacts at both low and high frequencies (orange points, bottom left). The first hypothesis under investigation in this paper is that different spoofing attacks are characterised by artefacts within different sub-bands and that an ensemble of different front-ends are needed in order to detect such artefacts reliably.

Both countermeasures produce predominantly high scores for bona fide utterances; as per standard ASVspoof practice, high countermeasure scores reflect bona fide trials, whereas low scores reflect spoof trials. Since CM_1 and CM_2 and their respective thresholds θ_1 and θ_2 are tuned for the detection of spoofing attacks A_1 and A_2 respectively, spoofing attack A_1 provokes mostly low scores for CM_1 and mostly high scores for CM_2, and vice versa for attack A_2. Attack A_3 provokes low scores for both countermeasures. Considering *multiple diverse attacks and countermeasures*, a notional decision boundary that best separates bona fide from spoofing utterances might correspond to a non-linear function. Linear score fusion operators may not perform well in this case, leading to poor reliability. The second hypothesis under investigation in this paper is that a non-linear approach to score fusion or system combination is needed in order to best exploit the complementarity of an ensemble of countermeasures tuned for the detection of specific spoofing attacks.

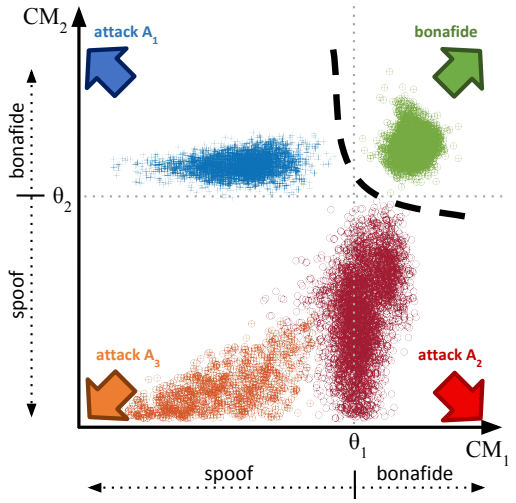

Figure 1: *A scatter plot of scores for countermeasures CM_1 and CM_2. Clusters correspond to bona fide utterances (green) and three spoofing attacks (A1-blue, A2-red, A3-orange). The dashed black line indicates a non-linear decision boundary that best separates bona fide from spoofed utterances.*

3. Experimental setup

Experiments were performed with the logical access (LA) partition of the ASVspoof 2019 database and standard protocols [13, 14]. Sub-band analysis and fusion experiments are furthermore based upon one of the two standard baselines. The database, baseline and assessment metric are all described here.

3.1. Database and metrics

The ASVspoof 2019 LA database consists of three independent partitions: train; development; evaluation. Spoofed speech in each dataset is generated using a set of different speech synthesis and voice conversion algorithms [14]. There is a total of 19 different spoofing attacks. Attacks in the training and development set were created with a set of 6 different algorithms (A01-A06), whereas those in the evaluation set were created with a set of 13 algorithms (A07-A19). The three partitions are used according to standard ASVspoof practice [15].

The primary metric used in this work is the minimum normalised tandem detection cost function (t-DCF) metric [16, 17]. The t-DCF reflects the impact of spoofing and countermeasures (CMs) upon the reliability of an automatic speaker verification (ASV) system. To give a more intuitive impression of countermeasure performance, results are also reported in terms of the pooled equal error rate (EER) computed using [18].

3.2. Baselines

The ASVspoof 2019 baseline systems use either constant Q cepstral coefficient (CQCC) [19, 20] or linear frequency cepstral coefficient (LFCC) [1] front-ends and a common Gaussian mixture model (GMM) back-end. Both LFCC-GMM and CQCC-GMM baseline systems are described in full in [13, 15]. While results for both baselines are reported in this paper, and while most of our group's recent work is based upon constant Q transform (CQT) representations [21], for reasons discussed later, experimental work reported here is all based upon modifications to the LFCC front-end. The baseline LFCC front-end configuration includes 20ms frame-blocking with 10ms shift, a filterbank with 20 linearly-scaled filters and 20 static, velocity (Δ) and acceleration ($\Delta\Delta$) coefficients, thereby giving 60-dimensional feature vectors.

The GMM back-end classifier has two 512-component Gaussian models. The first is a model of bona fide speech whereas the second is a model of spoofed speech, with both being learned using bona fide and spoofed speech data from the ASVspoof 2019 LA training partition. Scores are log-likelihood ratios (LLRs) computed in the usual way.

4. Sub-band front-ends

Our recent work [9] showed that spoofing artefacts reside at the sub-band level and that these are best detected with front-ends that exhibit a high spectral resolution within the same sub-band. That work used two different CQCC front-ends that were tuned to increase the spectral resolution at either low or high frequencies. The current work extends this idea by considering an ensemble of classifiers with each one being tuned for the detection of a set of specific spoofing attacks and the associated artefacts *no matter where they are in the spectrum.* Since the CQT has a non-linear spectral resolution [21, 22] which is difficult to tune to specific sub-bands, the work was performed by adapting the baseline LFCC front-end described in Section 3.2.

Described here is the strategy of spectral resolution and front-end tuning at the sub-band level. Also presented are results for each front-end when used with a GMM back-end and tested against each spoofing attack in the ASVspoof 2019 LA database.

4.1. Spectral resolution

Use of a spectral resolution that is *too* high will result in noisy features. Hence, before sub-band optimisation, we set out first to optimise the spectral resolution at the full-band level. This work was performed using the the full ASVspoof 2019 LA training and development subsets.

While other techniques could also have been applied, e.g. zero padding, we simply modified the baseline LFCC front-end (Sec. 3.2) to use a 30 ms window with a 15 ms shift and used a 1024-point Fourier transform. The resolution was then decreased using a filterbank in the usual fashion with a number of filters N [23]. For any one experiment, training and development data were processed with the given front-end before the GMM back-end was re-learned and used to process the development data in otherwise identical fashion to the baseline.

Results depicted in Table 1 show CM performance in terms of the min t-DCF and EER against the number N of filterbank filters (first 3 columns). For $N > 30$ filters, both the min t-DCF and the EER are zero. An alternative approach to optimisation is hence necessary. We elected arbitrarily to use the Bhattacharyya distance [24] between the CM score distributions for bona fide and spoofed trials given by:

$$D_B(b,s) = \frac{1}{4} ln \left(\frac{1}{4} \left(\frac{\sigma_b^2}{\sigma_s^2} + \frac{\sigma_s^2}{\sigma_b^2} + 2 \right) \right) + \frac{1}{4} \left(\frac{(\mu_b - \mu_s)^2}{\sigma_b^2 + \sigma_s^2} \right),$$

where subscripts b and s indicate parameters for bona fide and spoofed score distributions and where μ and σ refer to the means and standard deviations respectively. Results in the last column of Table 1 show that the distance between score distributions increases for $N > 30$ filters, but with little gain beyond $N = 70$ filters, which is the configuration used for all further experiments reported in this paper.

Table 1: *min t-DCF, EER and Bhattacharyya distance between bona fide and spoofed score distributions for different numbers of subband filters N. Baseline configuration illustrated in bold; selected configuration in italics.*

Filters (N)	min t-DCF	EER (%)	D_B
20	**0.2110**	**2.71**	**0.1338**
30	0.0000	0.79	0.1706
40	0.0000	0.00	0.1770
50	0.0000	0.00	0.1785
60	0.0000	0.00	0.1793
70	*0.0000*	*0.00*	*0.1826*
80	0.0000	0.00	0.1788
90	0.0000	0.00	0.1823
100	0.0000	0.00	0.1830
120	0.0000	0.00	0.1820

4.2. Centre of Mass Function

Attack-specific, sub-band front-ends are designed using heat-map visualisations [9] which show CM performance at the sub-band level. An example for the A04 attack is illustrated in Fig. 2. The heat-map colour signifies CM performance in terms of t-DCF for a front-end with cut-in and cut-off frequencies of f_{min} (x-axis) and f_{max} (y-axis) respectively.

We used the centre-of-mass (CoM) approach [25] to identify a single point in the heat-map and hence to define a specific sub-band for the detection of each attack in the development subset (A01–A06). The CoM is a crude means of coping with a noisy surface containing multiple minima. The CoM of a distribution of mass in space is the unique point where the weighted relative position of the distributed mass sums to zero. We consider the 2D heat-map as a system of particles P_i where $i = 1, \ldots, n$. Each particle has coordinates $r_i = [f^i_{\text{min}}, f^i_{\text{max}}]$ and mass $m_i = (\text{min t-DCF}_i)^{-1}$. The coordinates $R = [f^{CoM}_{\text{min}}, f^{CoM}_{\text{max}}]$ of the CoM satisfy the condition $\sum_{i=1}^{n} m_i(r_i - R) = 0$. Solving for R yields:

$$R = \frac{1}{M} \sum_{i=1}^{n} m_i r_i \qquad (1)$$

where M is the sum of the masses of all the particles in the full heat-map. We obtain a different R for each attack and hence define six attack-optimised, sub-band CMs. For attack A04 the CoM point illustrated by the white cross in Fig. 2 signifies a sub-band of 3209 to 8000 Hz. The CoM-defined sub-bands for each attack A01-06 are listed in the first column of Table 2.

4.3. Sub-band CM results

Results presented in Table 2 show performance in terms of t-DCF for the six sub-band and one full-band CM. The bandwidth of each system is illustrated in the first column. Results for the development set (columns A01-06) show that sub-band CMs all yield zero t-DCFs for the attacks for which they are designed (results in boldface), as they also do for some other attacks. This is not surprising since there is some considerable spectrum overlap in the set of sub-band CMs. Interestingly, the full-band CM is the only one to achieve zero t-DCF for all six attacks in the development set. Results for the evaluation set (columns A07-19) show that the full-band CM gives similar or substantially lower t-DCFs than sub-band CMs. These observations are confirmed by pooled t-DCFs (columns P1) for both the development and evaluation subsets. The questions now

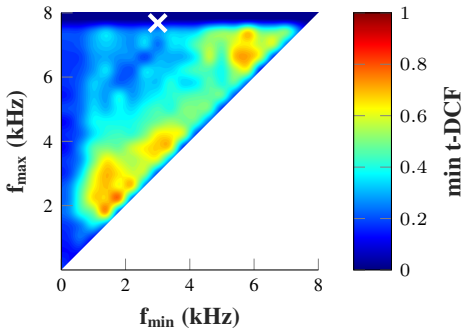

Figure 2: *A 2-D heatmap visualisation (see [9]) illustrating sub-band level CM performance for attack A04 of the ASVspoof 2019 LA database. The cut-in frequencies f_{min} and cut-off frequencies f_{max} are indicated on horizontal and vertical axes respectively. Those of the CoM-defined sub-band is indicated by the white cross.*

are: (i) whether or not the fusion of attack-specific, sub-band CM scores can give better performance even when their individual performance is poor relative to the full-band CM; (ii) what should be the fusion mechanism.

5. Fusion

Fusion experiments aim to assess the second research hypothesis in this work, namely that a non-linear approach to score fusion or system combination is needed in order to best exploit the complementarity of sub-band CMs. We used four different fusion methods to obtain a single score from the set of seven scores: six sub-band CMs and one full-band CM.

Approaches to fusion include: a support vector machine (SVM) [26] with a seventh order polynomial kernel[1]; multinomial logistic regression [27]; traditional linear fusion [28]. Also tested was a GMM-based approach to fusion for which 64-component models are learned from the set of scores for bona fide and spoofed classes. This approach was used previously for the fusion of ASV and CM scores [29]. Both the GMM and SVM are non-linear approaches to fusion; the others are linear. All but the SVM approach produce log-likelihood ratio outputs.

5.1. Fusion results

Fusion results for the four systems are shown in boldface in Table 3. With a t-DCF of 0.0740, the non-linear GMM approach gives the best performance. The next best system is the non-linear SVM approach with a t-DCF of 0.0748. The performance of the two linear approaches yield t-DCFs of 0.0911 and 0.1182. These findings would seem to confirm the hypothesis that a non-linear approach is better suited to the fusion of sub-band CM scores. This is because spoofing artefacts that are localised in the spectrum may be detected only by sub-band CMs whose focus is directed towards the same parts of the spectrum and hence be detected reliably by a sub-set of CMs only (or even only a single CM). In this case, full-band CMs may dilute relevant information by smoothing across the spectrum and linear approaches to fusion may not identify the best decision boundary between bona fide and spoofed speech; such an optimal decision boundary might be non-linear.

[1]Linear and residual basis function kernels were also tested and yielded inferior results. These experiments are not reported here.

Table 2: *Results in terms of min t-DCF for development (A01-A06) and evaluation (A07-A19) partitions and respective pooled min t-DCF (P1) and pooled EER (P2). Results in boldface signify the attack for which each sub-band is optimised, e.g. the CM designed for attack A01 operates within a sub-band of 2011 to 6403 Hz.*

Freq-bands	A01	A02	A03	A04	A05	A06	P1	P2	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	A17	A18	A19	P1	P2
2011-6403	**0.00**	0.00	0.00	0.22	0.25	0.79	0.25	0.11	0.37	0.03	0.00	0.55	0.03	0.18	0.31	0.17	0.15	0.25	0.41	0.60	0.93	0.34	13.28
2410-5604	0.00	**0.00**	0.00	0.31	0.42	0.91	0.33	0.13	0.35	0.06	0.00	0.54	0.08	0.23	0.31	0.16	0.16	0.33	0.58	0.87	0.99	0.39	15.50
2011-5604	0.00	0.00	**0.00**	0.27	0.31	0.82	0.29	0.12	0.37	0.06	0.00	0.56	0.08	0.22	0.28	0.18	0.17	0.30	0.48	0.81	0.99	0.38	14.75
3209-8000	0.00	0.00	0.00	**0.00**	0.15	0.00	0.03	0.01	0.00	0.00	0.00	0.54	0.00	0.41	0.71	0.10	0.23	0.00	0.47	0.29	0.00	0.26	10.59
15.62-4806	0.00	0.00	0.00	0.16	**0.00**	0.51	0.18	0.08	0.45	0.01	0.00	0.54	0.07	0.09	0.09	0.12	0.13	0.16	0.33	0.60	0.85	0.31	12.31
3608-8000	0.00	0.00	0.00	0.00	0.12	**0.00**	0.04	0.01	0.00	0.00	0.00	0.55	0.00	0.40	0.79	0.09	0.31	0.00	0.55	0.24	0.00	0.27	11.55
full-band	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.11	0.07	0.06	0.06	0.00	0.35	0.07	0.00	**0.09**	**03.50**

Table 3: *Performance for the ASVspoof 2019 evaluation partition in terms of pooled min t-DCF and pooled EER for top-performing systems (T05, T45, T60 and T24), four different approaches to fusion (boldface) and baseline systems (B1, B2).*

System	min-tDCF	EER
T05	0.0069	0.22
T45 [30]	0.0510	1.86
GMM fusion	**0.0740**	2.92
SVM fusion (polynomial kernel)	**0.0748**	2.92
T60 [11]	0.0755	2.64
Optimised LFCC (full-band)	0.0904	3.50
Linear fusion	**0.0911**	3.38
T24	0.0953	3.45
Multinomial logistic regression fusion	**0.1182**	4.50
LFCC:B2 [13]	0.2116	8.09
CQCC:B1 [13]	0.2366	9.57

5.2. Performance comparison and discussion

Table 3 also shows results for the two ASVspoof 2019 baseline systems (B1 and B2, last two rows) and the top-performing four (out of 48 submissions) challenge results [13]. The latter are signified by their anonymous ASVspoof 2019 identifiers T05, T45, T60 and T24 [13]. Only T45 [30] and T60 [11] system details are in the public domain but from these and the description of the ASVspoof 2019 challenge in [13], it is known that all four of these competing systems are based upon an ensemble of comparatively complex neural network based architectures, as opposed to a simple GMM-based solution used in our work. Furthermore, they used a combination of *multiple, different* front-end parameterisations, unlike the use of the *single, same* base front-end used in our work. While we acknowledge that this comparison is between evaluation and post-evaluation results, both non-linear GMM-based and SVM approaches to fusion outperform all but two of the 48 competing systems. Even though the gap is not substantial, the two linear approaches to fusion are outperformed by the two non-linear approaches.

6. Conclusions

The work reported in this paper investigated whether spoofing attacks leave sub-band artefacts that require specific spoofing countermeasures for detection. In addition, it sought to determine whether non-linear fusion approaches offer better potential to combine the scores produced by ensemble of sub-band classifiers. Extending our prior work, we used a high-resolution base front-end that is adapted using a crude center of mass technique to identify 6 different, additional sub-band front-ends, all of which are used with a GMM-based back-end that is relearned for each feature set. Fusion was performed with a variety of different techniques, both linear and non-linear.

Excellent results obtained using a high-resolution, full-band classifier alone demonstrate the importance of the front-end. This finding could be beneficial to other anti-spoofing researchers that use neural networks with standard, low-resolution front-ends. A switch to high-resolution front-ends may improve performance; even advanced neural network solutions cannot recover information that is already lost, e.g. in spectro-temporal decomposition. Our results also show that sub-band classifiers can detect reliably all attacks in the development data upon which sub-bands classifiers were learned. Even though evaluation results are far less promising, fusion results still show that the use of sub-band classifiers helps to improve performance beyond what can be achieved with a full-band classifier alone and that non-linear fusion outperforms linear fusion.

Despite its simplicity, our approach outperforms all but two competing challenge systems. Noting that our approach is learned used only training data and not combined training and development data, as was permitted by ASVspoof 2019 rules, noting also that we did not optimise the approach used for sub-band selection nor tackle spectral overlap in any way, this is a particularly satisfactory result. Other experiments for which we do not have the space to report lend further support to our approach. Leave-one-out fusion experiments showed the consistent benefit of sub-band classifiers and non-linear fusion. Also, the use of linearly partitioned sub-bands in an otherwise identical setup gave worse performance and show the merit of attack-specific sub-bands, a finding supported by other authors in concurrent work [31], albeit for replay spoofing attacks.

Our future work will investigate non-linear probabilistic linear discriminant analysis back-end techniques; linear approaches which assume matching within-class co-variance proved unsuccessful since the within-class co-variance of bona fide and spoofed data is different. Score normalisation before fusion could also be explored as a means to improve performance using linear fusion. Another natural extension is to explore the use of high-resolution and sub-band front-ends with neural network based architectures. The goal of this work would be to see if localised spectral information is being used in the same way and hence to improve upon the interpretability and explainability of complex neural network techniques.

7. Acknowledgements

The work was partially supported by the Voice Personae and RESPECT projects, both funded by the French Agence Nationale de la Recherche (ANR).

8. References

[1] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. INTERSPEECH*, Dresden, Germany, 2015, pp. 2087–2091.

[2] K. Sriskandaraja, V. Sethu, P. N. Le, and E. Ambikairajah, "Investigation of sub-band discriminative information between spoofed and genuine speech," in *Proc. INTERSPEECH*, San Francisco, USA, 2016, pp. 1710–1714.

[3] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio replay attack detection using high-frequency features." in *Interspeech*, 2017, pp. 27–31.

[4] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay attack detection using DNN for channel discrimination." in *Interspeech*, 2017, pp. 97–101.

[5] L. Lin, R. Wang, and Y. Diqun, "A replay speech detection algorithm based on sub-band analysis," in *International Conference on Intelligent Information Processing (IIP)*, Nanning, China, 2018, pp. 337–345.

[6] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, 2019.

[7] S. Garg, S. Bhilare, and V. Kanhangad, "Subband analysis for performance improvement of replay attack detection in speaker verification systems," in *International Conference on Identity, Security, and Behavior Analysis (ISBA)*, 2019, pp. 1–7.

[8] J. Jung, H. Shim, H. Heo, and H. Yu, "Replay attack detection with complementary high-resolution information using end-to-end DNN for the ASVspoof 2019 challenge," in *Interspeech*, 2019, pp. 1083–1087.

[9] H. Tak, J. Patino, A. Nautsch, N. Evans, and M. Todisco, "An explainability study of the constant Q cepstral coefficient spoofing countermeasure for automatic speaker verification," *Speaker Odyssey Workshop*, 2020.

[10] Z. Ji, Z. Y. Li, P. Li, M. An, S. Gao, D. Wu, and F. Zhao, "Ensemble learning for countermeasure of audio replay spoofing attack in ASVspoof2017." in *INTERSPEECH*, 2017, pp. 87–91.

[11] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1118–1112.

[12] J. Monteiro, J. Alam, and T. H. Falk, "An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers," in *ICASSP*, Barcelona, May 4-8, 2020.

[13] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1008–1012.

[14] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. Lee *et al.*, "The ASVspoof 2019 database," *arXiv preprint arXiv:1911.01601*, 2019.

[15] ASVspoof 2019: the automatic speaker verification spoofing and countermeasures challenge evaluation plan. [Online]. Available: http://www.asvspoof.org/asvspoof2019/asvspoof2019_evaluation_plan.pdf

[16] T. Kinnunen, K. Lee, H. Delgado, N. Evans, M. Todisco, J. Sahidullah, M.and Yamagishi, and D. A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Speaker Odyssey Workshop*, Les Sables d'Olonne, France, 2018, pp. 312–319.

[17] T. Kinnunen, H. Delgado, N. Evans, K. A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, "Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals," *IEEE/ACM Transactions on Audio Speech and Language Processing (TASLP)*, 2020, submitted.

[18] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.

[19] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Speaker Odyssey Workshop*, Bilbao, Spain, 2016, pp. 249–252.

[20] ——, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language*, vol. 45, pp. 516–535, 2017.

[21] J. C. Brown, "Calculation of a constant Q spectral transform," *The Journal of the Acoustical Society of America (JASA)*, vol. 89, no. 1, pp. 425–434, 1991.

[22] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Proc. Audio Engineering Society International Conference on Semantic Audio*, London, UK, 2014.

[23] J. R. Deller Jr, J. G. Proakis, and J. H. Hansen, *Discrete time processing of speech signals*. Prentice Hall PTR, 1993.

[24] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[25] R. P. Feynman, R. B. Leighton, and M. Sands, *The Feynman lectures on physics; New millennium ed.* New York, NY: Basic Books, 2010, originally published 1963-1965. [Online]. Available: https://cds.cern.ch/record/1494701

[26] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[27] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[28] N. Brümmer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for binary classifier score processing," 2011.

[29] M. Todisco, H. Delgado, K. Lee, M. Sahidullah, N. Evans, T. Kinnunen, and J. Yamagishi, "Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion," in *Proc. INTERSPEECH*, Hyderabadd, India, 2018, pp. 77–81.

[30] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 1033–1037.

[31] B. Chettri, T. Kinnunen, and E. Benetos, "Subband modeling for spoofing detection in automatic speaker verification," *Speaker Odyssey Workshop*, 2020.