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 biannual 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: CM1 and CM2. Countermeasure CM1 is tuned to detect artefacts present within a lower sub-band, at 0-4 kHz for example. Countermeasure CM2 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 A1, characterised by artefacts predominantly at low frequencies (blue points, top-left); attack A2, characterised by artefacts at high frequencies (red points, bottom right); attack A3, 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 CM1 and CM2 and their respective thresholds 21 and 22 are tuned for the detection of spoofing attacks A1 and A2 respectively, spoofing attack A1 provokes mostly low scores for CM1 and mostly high scores for CM2, and vice versa for attack A2. Attack A3 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.
configuration includes 20ms frame-blocking with 10ms shift, a filterbank with 20 linearly-scaled filters and 20 static, velocity ($\Delta$) and acceleration ($\Delta^2$) 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.

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,
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.

<table>
<thead>
<tr>
<th>Filters ($N$)</th>
<th>min t-DCF</th>
<th>EER (%)</th>
<th>$D_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.2110</td>
<td>2.71</td>
<td>0.1338</td>
</tr>
<tr>
<td>30</td>
<td>0.0000</td>
<td>0.79</td>
<td>0.1706</td>
</tr>
<tr>
<td>40</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1770</td>
</tr>
<tr>
<td>50</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1785</td>
</tr>
<tr>
<td>60</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1793</td>
</tr>
<tr>
<td>70</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1826</td>
</tr>
<tr>
<td>80</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1875</td>
</tr>
<tr>
<td>90</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1823</td>
</tr>
<tr>
<td>100</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1830</td>
</tr>
<tr>
<td>120</td>
<td>0.0000</td>
<td>0.00</td>
<td>0.1820</td>
</tr>
</tbody>
</table>

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 $f_{\text{min}}$ (x-axis) and $f_{\text{max}}$ (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 $r_i = [f_{\text{min}}, f_{\text{max}}]$ and mass $m_i = (\text{min t-DCF}_i)^{-1}$. The coordinates $R = [f_{\text{CoM}}^m, f_{\text{CoM}}^m]$ of the CoM satisfy the condition $\sum_{i=1}^n m_i(r_i - R) = 0$. Solving for $R$ yields:

$$R = \frac{1}{M} \sum_{i=1}^n m_i r_i$$  (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 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 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; multinoval 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 CM scores. This is because spoofing artefacts that are localised in the spectrum may be detected only by sub-band CM scores. This is because spoofing artefacts that are localised in the spectrum may be detected only by sub-band CM scores. This is because spoofing artefacts that are localised in the spectrum may be detected only by sub-band CM scores.

\footnote{Linear and residual basis function kernels were also tested and yielded inferior results. These experiments are not reported here.}
of which are used with a GMM-based back-end that is rehearled 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).

---

**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.

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

**Table 3:** Performance of 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).

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

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 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...
8. References


