Use of Paraconsistent Feature Engineering to support the Long Term Feature choice for Speaker Verification

Alex M. G. de Almeida, Claudineia H. Recco, Rodrigo C. Guido
Departament of Computer Science
São Paulo State University
Rua Cristóvão Colombo 2265, Jd Nazareth
15054-000, São José do Rio Preto - SP, Brazil
alex.marino@fatecourinhos.edu.br, chrecco@gmail.com, guido@ieee.org

Abstract

The state-of-art models for speech synthesis and voice conversion can generate synthetic speech perceptually indistinguishable from human speech, and speaker verification is crucial to prevent breaches. The building feature that best distinguishes genuine speech between spoof attacks is an open research subject. We used the baseline ASVSpoof2017, Transfer Learning (TL) set, and Symlet and Daubechies Discrete Wavelet Packet Transform (DWPT) for this investigation. To qualitatively assess the features, we used Paraconsistent Feature Engineering (PFE). Our experiments pointed out that for the use of more robust classifiers, the best choice would be the AlexNet method, while in terms of classification regarding the Equal Error Rate metric, the best suggestion would be Daubechies filter support 21. Finally, our findings indicate that Symlet filter support 17 as the most promising feature, which is evidence that PFE is a useful tool and contributes to feature selection.

Introduction

Among biometric strategies, such as fingerprint and face recognition, Automatic Speaker Verification (ASV) systems are essential in the real world. They are particularly intended to determine whether a voice recording belongs to a pre-registered speaker or not. Just as any other biometric system, ASV algorithms are subject to different types of attacks. Due to the rapid development in speech technology, voice conversion (Toda, Saruwatari, and Shikano 2001), and speech synthesis (De Leon et al. 2012) techniques make it possible to generate synthetic speech that is good enough to deceive an ASV system. Most efforts have been focused on finding new features. In the challenges organized about spoofing detection, a relevant number of papers focuses on building more appropriate feature representations such as phase spectrum (Wang et al. 2015), DWPT (Daqrouq et al. 2012), TL (Aravind et al. 2020) and constant Q cepstral coefficients (CQCC) (Todisco, Delgado, and Evans 2017). Even though ASV systems perform well among known attacks, they fail among unknown attacks. Thus, as shown in paper (Todisco, Delgado, and Evans 2017), we believe that the design of countermeasures should start with a search for a good set of discriminative features rather than complex classifiers.

Copyright © 2021by the authors. All rights reserved.

This work aims at investigating two sets of features, i.e., those coming from TL and DWPT, comparing with CQCC as a baseline, which is the one that best discriminates genuine and spoof speakers based on Equal Error Rates (EER). For this purpose, experiments were carried out based on the feature extraction from the entire segment of the input speech signals subsequently submitted to a Gaussian Mixture Model (GMM) classifier, according to the baseline experiments of ASVSpoof2017 (Kinnunen et al. 2017). In parallel, we perform Paraconsistent Feature Engineering (PFE) to all features to identify feature quality (Guido 2018). Specifically, signal energy-related methods A_1, A_2 and A_3, as defined in (Guido 2016), are experimented here under PFE application. Our key research questions are: (i) among the energy-related methods used, is it possible to determine the most appropriate?; (ii) is it possible, over PFE, to indicate the best set of features?; (iii) what set of experiments is more appropriate to discriminate genuine from spoofed speech?; (iv) is it possible to choose the promising potential feature considering the best PFE criteria and EER measure concomitantly?

The Set of Features

In this work, we use baseline CQCC, obtained from the Constant Q Transform of the signal under analysis followed by a uniform resampling and a Discrete Cosine Transformation (DCT), and two feature extraction methods, i.e., TL and DWPT.

TL is usually expressed through the use of pre-trained models. The computational cost of training deep models is computationally high, so it is common to reuse models from published literature, and a low-cost feature extractor (Tan et al. 2018). Our approach employs TL of a set of fast adaptation methods to the Mel-spectrograms extracted from inputs signal. Table 1 presents the output layers and the respective TL algorithm employed.

AlexNet	GoogleNet	ResNet18	ResNet50	ResNet101
fc7	loss3	fc1000	fc1000	fc1000

Table 1: TL algorithms output layer

DWPT has been used in recent works as a feature ex-

tractor in different research domains (Wang, Gan, and others 2018), it performs the recursive decomposition of the speech signal obtained by using a recursive binary tree. Given a signal, a pair of low-pass and high-pass filters were used to produce two sub-signals, i.e., trend and fluctuation, containing the original signal's relative energy features. Zero-padding was adopted to increase the signal length whenever required to allow for a mid-level decomposition tree.

Experimental Design

This section presents the proposed approach, divided into the following stages: Dataset, Evaluation Metrics, and Experimental Setup.

Dataset

All experiments performed in this work were conducted on ASVSpoof 2017 Dataset, which is focused on replay attack detection, for which details can be found in paper (Kinnunen et al. 2017). The Dataset was partitioned into three subsets: training, development, and evaluation. Each speech file in training and development was labeled as genuine or spoof, as shown in Table 2.

Subset	Genuine	Spoof	Total
Training	1508	1508	3016
Development	760	950	1710
Evaluation	1298	12922	14220

Table 2: Details about number of labeled samples on ASVspoof2017.

Evaluation Metrics

The primary metric is the *Equal Error Rate* (EER). Let $P_{ta}(\theta)$ and P_{miss} be the false alarm and miss rates at threshold 0 defined as:

$$P_{ta}(\theta) = \frac{\#\{\text{replay trials with score} > 0\}}{\#\{\text{total replay}\}}$$

and

$$P_{miss}(\theta) = \frac{\#\{\text{non-replay trials with score} \leq 0\}}{\#\{\text{total non-replay}\}},$$

so that $P_{ta}(\theta)$ and $P_{miss}(\theta)$ are monotonically decreasing and increasing functions of θ, respectively. The EER corresponds to to the θ for which two detection error rates are equal.

The second measure is the *Root Mean Square Error* (RMSE) which can be used interpreted as a reliability measure, is defined as:

$$RMSE(f) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2} \quad .$$

RMSE serves as an excellent general-purpose performance measure suitable for evaluating probabilistic classifiers (Japkowicz and Shah 2011).

Experimental Setup

The proposed approach contains three main steps, as shown in Figure 1: feature extraction, modeling, and feature evaluation, and PFE. For FE, we performed our experiments considering the baseline CQCC, TL comprising AlexNet, GoogleNet, ResNet18, ResNet50, and ResNet108 and, in addition, DWPT comprising Daubechies and Symmlet wavelet families with filter support sizes varying from 2 to 45 and from 2 to 35, respectively. In this step, three sets of feature vectors are produced and used to train the GMM models.

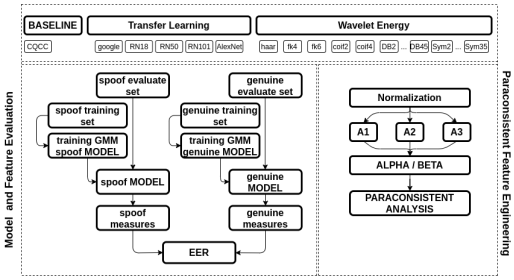

Figure 1: The proposed approach

The classification step uses the feature vectors as input to train each GMM model. For both the spoof and genuine subsets, the parameters considered are shown in Table 3. The

	genuine	spoof
mixture	512	512
iterations	10	10

Table 3: GMM setup

PFE step initially normalizes the input feature vectors obtained using the methods A1, A2, A3 (Guido 2016). Then, it performs intraclass and interclass verification and subsequently computes α, which expresses a level of faith, and β expresses a level of discredit. Where $0 \leq \alpha, \beta \leq 1$. Independence indicates that α and β are not complementary. Finally, it finds the point (G_1, G_2) and then computes Euclidean distance from the paraconsistent reticulated optimal point $(1, 0)$, which represents utmost faith and minimum discredit (Guido 2018).

Experimental Results

Our discussions start at the first question: *"(i) among the normalization methods used, is it possible to determine the most appropriate?"*. Observing Figure 2, it is possible to note that the set of experiments normalized with method A_3 presents median high $\beta = 0.83763$ criterion and low $\alpha = 0.33691$ criterion, implying a quiet faith concerning intraclass analysis and high discredit concerning interclass analysis. Proceeding, method A_1 presents a median high $\alpha = 0.87615$ criterion with a low $\beta = 0.23448$ criterion, whereas A_2 shows a median high $\alpha = 0.98330$ criterion with a low $\beta = 0.02938$ criterion. In contrast to A_3, the

features normalized with A_1 and A_2 better match intraclass similarity and interclass dissimilarities criteria. Considering A_1 and A_2, the latter is better than the former, given that its median is located in the A_2 outlier area. In this way, it is discernible that A_2 produces the best result concerning PFE. To answer the second question, i.e., *"(ii) is it possible over*

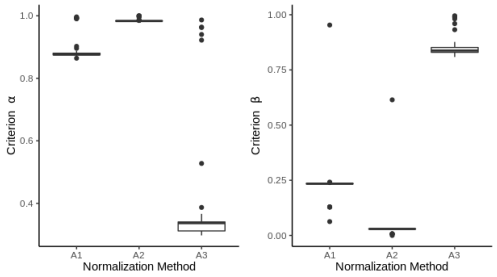

Figure 2: Boxplot with similarity (α criterion) and dissimilarity (β criterion) for the three energy-related methods used.

PFE indicate the best set of features?", and assuming the prevalence of A_2, it is necessary to observe Figure 3, noting that the set of TL features produces $0.99915 \leq \alpha \leq 1.00000$ and $0 \leq \beta \leq 0.00698$. In comparison, the set of Symmlet- and Daubechies-related features produces $0.98502 \leq \alpha \leq 0.98235$ and $0.029989 \leq \beta \leq 0.029390$, whereas the baseline CQCC feature produces $\alpha = 0.99513$ and $\beta = 0.61440$. This implies a high intraclass level of faith concerning the TL set and more significant than baseline CQCC, which is greater than that obtained with the set of DWPT features. Regarding the interclass dissimilarity, it is possible to note that baseline CQCC does not outperform the DWPT set, which shows a worse result in contrast with the TL set. Hence, we conclude that TL and DWPT set to produce better features for classification than the baseline CQCC. Proceed-

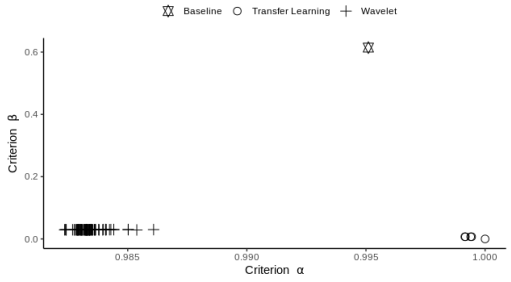

Figure 3: Feature set plot considering α and β criteria.

ing, question *"(iii) what set of experiments is more appropriate to discriminate genuine and spoof speakers?"*, essentially involves a classification result. When looking at Figure 4, particularly at the line graph, we observe that the best performance was obtained with the DWPT Daubechies Family with filter support 21, for which $EER = 11.99$. Still analyzing the line graph, the best result for the Symmlet family was obtained with a filter producing $EER = 12.91$. When observing the barplot of Figure 4, we also perceive that the experiments carried out with the TL set obtained a better result of $EER = 24.1$ for the ResNet101. Notably, the difference between the worst result obtained from the union of the DWPT Symmlet and Daubechies families of $EER = 17.97$ and the best TL result is almost twice the standard deviation of all experiments. Thus, joining the fact that the RMSE of the set of experiments with the DWPT is slightly higher than the third part of the RMSE of the TL experiments, as in Table 4, it is possible to suggest the prevalence of the set of results obtained with the DWPT set over the TL set.

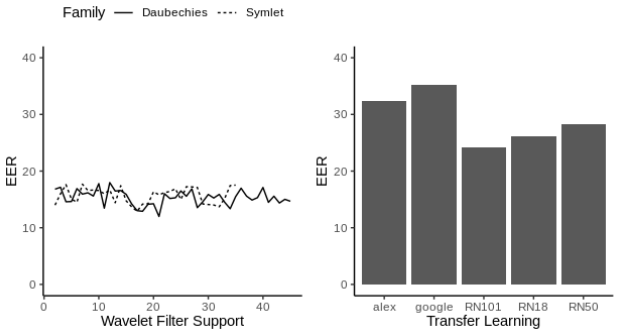

Figure 4: EER - wavelet families x TL

	TL set	DWPT set	All
MIN	24.15	11.99	5.36
MAX	35.22	17.96	35.22
AVERAGE	29.26	15.46	16.26
RMSE	4.05	1.42	4.03
STDDEV	4.52	1.36	3.64

Table 4: Global stats grouped by set of experiments.

Finally, to respond to the question: *"(iv) is it possible to choose the better feature regarding the joining of best PFE criteria and EER measure?"*, we observe that while the answer to the second research question suggests that the TL set produces a greater degree of faith and a lower degree of discredit, for the third research question we know that the set of DWPT features is more appropriate, creating a paradox. To solve it, we produced a comparative ranking for each experiment based on its classification in terms of feature quality. This comparative scale has a value of 1 associated with the best performance due to EER and PFE: the higher the ordinal value, the worse its relative performance. The ranking produces a scale that reveals no distance between the experiments. We used the average between the performances of the EER and PFE criteria that produced a final ranking. When looking at Table , we notice that the feature produced by Alexnet appears at the point of maximum and minimum discredit, i.e., $D(1, 0) = 0$. For the next four experiments, ResNet101, 50, and 18 present a $D(1, 0)$ slightly lower than AlexNet, reaffirming the advantage of the set of TL features over the DWPT features and baseline.

Feature	EER	Rank	D10	Feature
db21	11.99	1	0.000000	AlexNet
sym17	12.91	2	0.009913	ResNet101
db18	12.91	3	0.009926	ResNet50
db17	13.02	4	0.009954	ResNet18
db34	13.36	5	0.009958	googlenet
db11	13.46	6	0.047136	db43
db28	13.55	7	0.047142	db39
sym16	13.70	8	0.047319	db40
sym32	13.71	9	0.047366	db41
sym2	14.00	10	0.047383	db37

Table 5: Top 10 - EER and D10 Ranking

Contrary to this, and focusing again on Table , we note that among the TOP 10, there are only experiments carried out with features extracted with DWPT set, thus confirming its prevalence concerning the baseline and TL set. Differently, when looking at Table , we still notice that in the TOP 10, there are only experiments carried out with the DWPT set. However, the fact that there is no TL experiment is highlighted. Highlighting that the experiment carried out with DWPT Symmlet with filter support 17 produces the best combination in join EER classification and feature quality.

	Ranking		
Feature	EER	D10	General
sym17	2	11	6.5
sym2	10	13	11.5
db43	18	6	12
db34	5	20	12.5
db41	20	9	14.5
sym19	16	24	20
sym13	19	23	21
sym29	14	28	21
db39	36	7	21.5
db45	26	17	21.5

Table 6: Top 10 Global Ranking

Conclusion and Future Work

As an initial study, this article explored the application of PFE in conjunction with an experimental assessment of features extracted by using TL and Daubechies and Symmlet DWPT families with a wide range of filter support. We also evaluated the influence of signal energy-related methods on the result of PFE over the quality of features. Our experiments pointed out that A_2 method prevails over the others. They also suggested that, in terms of feature quality, the TL AlexNet method stood out with maximum α criterion and minimum β criterion, implying in full faith and minimum discredit, being better than the baseline. Contrary to this, it was outside the TOP 10 for EER. Such notes strongly suggest a high potential for discrimination, which would imply that if this work's future objective were directed towards exploring more robust classifiers, it would be the best choice.

Regarding the features extracted based on DWPT, Daubechies with support of Filter 21 and Symlet with support filter 17 was highlighted as presenting the best performances in terms of classification accuracy and EER. However, they are below the baseline. When evaluating feature quality and classification together, the results highlighted Symmlet with filter support 17 and, in the sequence, Daubechies with filter support 43. As future work, we intend to investigate the set of features that better discriminate genuine from spoofed speech. In this sense, we will abdicate the application of robust classifiers, so that the best set of features appears in exploring the wavelets bases. Hence, the suggestion of medium filter support close to 17 is a good indication for developing a new set of wavelet filters.

References

Aravind, P.; Nechiyil, U.; Paramparambath, N.; et al. 2020. Audio spoofing verification using deep convolutional neural networks by transfer learning. *arXiv preprint arXiv:2008.03464*.

Daqrouq, K.; Al-Qawasmi, A.; Daoud, O.; and Al-Sawalmeh, W. 2012. Self-organizing map weights and wavelet packet entropy for speaker verification. *International Journal of Circuits, Systems and Signal Processing* 6(1):12–20.

De Leon, P. L.; Pucher, M.; Yamagishi, J.; Hernaez, I.; and Saratxaga, I. 2012. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing* 20(8):2280–2290.

Guido, R. C. 2016. A tutorial on signal energy and its applications. *Neurocomputing* 179:264–282.

Guido, R. C. 2018. Paraconsistent feature engineering [lecture notes]. *IEEE Signal Processing Magazine* 36(1):154–158.

Japkowicz, N., and Shah, M. 2011. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press.

Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; and Lee, K. A. 2017. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, 270–279. Springer.

Toda, T.; Saruwatari, H.; and Shikano, K. 2001. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of straight spectrum. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 2, 841–844. IEEE.

Todisco, M.; Delgado, H.; and Evans, N. 2017. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language* 45:516–535.

Wang, L.; Yoshida, Y.; Kawakami, Y.; and Nakagawa, S. 2015. Relative phase information for detecting human speech and spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Wang, C.; Gan, M.; et al. 2018. Fault feature extraction of rolling element bearings based on wavelet packet transform and sparse representation theory. *Journal of Intelligent Manufacturing* 29(4):937–951.