AUXILIARY RAWNET: COMPLEMENTING HANDCRAFTED FEATURES WITH RAW WAVEFORM USING A LIGHT-WEIGHT AUXILIARY MODEL

Zhongwei Teng**, Quchen Fu***, Jules White***, Maria E. Powell*†*, Douglas C. Schmidt* *

* Dept. of Computer Science, Vanderbilt University
† Dept. of Otolaryngology–Head and Neck Surgery, Vanderbilt University Medical Center

ABSTRACT

An emerging trend in audio processing is capturing low-level speech representations from raw waveforms. These representations have shown promising results on a variety of tasks, such as speech recognition and speech separation. Compared to handcrafted features, learning speech features via backpropagation provides the model greater flexibility to represent data for different tasks theoretically. However, results from empirical studies show that handcrafted features are more competitive than learned features in some tasks, such as voice spoof detection. Instead of evaluating handcrafted features and raw waveforms independently, this paper proposes an Auxiliary Rawnet model to complement handcrafted features with features learned from raw waveforms. A key benefit of our approach is that it can improve accuracy at a relatively low computational cost. The proposed Auxiliary Rawnet model is tested using the ASVspoof 2019 dataset and the results from this dataset indicate that a light-weight waveform encoder can boost the performance of handcrafted-features-based encoders in exchange for a small amount of additional computational work.

Index Terms— Raw waveform, handcrafted features, spoof detection

1. INTRODUCTION

Fixed, handcrafted audio features, such as Mel-filter banks, exhibit high performance in capturing strong audio features in aspects of both auditory and machine learning [1, 2]. However, handcrafted features are often designed based on specific tasks, such as speech recognition. Therefore using these features to solve problems that they were not designed for may be suboptimal.

For example, Mel-filter banks apply triangular filter banks on a Mel-scale to spectrograms calculated using short-term Fourier transform (STFT) to represent the non-linear perception of the human hearing. The Mel-scale is derived from a set of perception experiments on humans. As a result, Mel-filter banks are coarse-grained at high-frequencies since humans are less sensitive to high frequency sound. This loss of signal energy (information) in high frequencies may lead to poor performance on tasks that rely on information in these higher frequencies [2].

Extracting audio features with backpropagation provides an alternative way to represent raw waveforms by using deep neural networks to learn task-specific features. Task-specific features can be learned for many problems, such as voice recognition[3, 4] or automatic speaker verification (ASV) Directly learning features from raw waveforms provides greater flexibility in handling unknown tasks, thereby overcoming some challenges of handcrafted features, which may lose signal energy needed by a specific task.

Previous research [1] indicates that representations learned from waveforms still have limitations on signal energy loss compared to the original raw signals they were learned from. On certain tasks, such as Voice Spoof Detection, models based on handcrafted data still show much better performance than models based on waveforms [5, 6]. Instead of relying on raw waveforms independently, therefore, a potential solution is to take advantage of both handcrafted *and* learned features.

For example, lost phase information in handcrafted features can be complemented by features learned from raw waveforms. There have been attempts to feed both handcrafted features and raw waveforms into networks for audio pattern recognition problems. Hoewver, little research has focused on the role of merging raw waveforms into arbitrary networks, as well as the trade-offs in model complexity of doing so.

This paper proposes the Auxiliary Rawnet (ARNet) architecture to combine learned features from raw waveforms with existing handcrafted features, by designing a lightweight auxiliary encoder. The proposed model was tested on the ASV Spoof 2019 dataset [7]. The model shows great promise in boosting the performance of single handcrafted-features-based networks that warrant further investigation on additional data sets and tasks.

This paper provides three contributions to research on complementing handcrafted features with raw waveforms using a light-weight auxiliary model. First, we elaborate on the problem of concatenating raw waveforms and handcrafted features in the speech field and propose an means to solve this problem efficiently. Second, we introduce the Auxiliary Rawnet architecture that attaches a light-weight auxiliary encoder to a model that relies on handcrafted features, thereby boosting model performance for the ASV spoof 2019 dataset that outperforms existing single systems. Third, we describe how our results show the potential of combining a light-weight waveform encoder with other encoders, providing an approach to balance the trade-off between performance and model complexity for models containing multiple encoders.

The remainder of this paper is organized as follows: Section2 discusses prior work in audio signal feature representation. Section3 explains the problem analyzed in this paper and describes the Auxiliary Rawnet structure. Section4 introduces the experimental dataset and tasks used in this paper. Section5 analyzes experimental results. Section6 presents concluding remarks and lessons learned.

2. RELATED WORK

Prior work has shown how the "front-end" of models, which extract features from raw data, can be improved by using deep neural networks [1, 8, 4, 2, 9] to directly learn features from raw signal data. Directly applying standard convolutional neural networks (CNNs) to process raw waveforms [10] has shown promising results in speech recognition, spoofing detection, and speech separation.

Convolutions on time-domain raw waveforms can be explained as finite impulse response filter banks [1]. Structured filters are applied to optimize standard CNNs based on digital signal processing theory, by initializing the first convolutional layer, which is believed to be the most important part, with known filter families [9, 11], so that a custom filter bank can be designed for a specific task.

Filter-based waveforms networks are emerging as excellent front-ends for many tasks [5, 2]. However, a theoretical analysis from Joakim et al. [1] shows that signal energy loss is still inevitable for features extracted from raw waveforms by a CNN. Their results show extracted features can carry up to 94.5% signal energy compared to the original waveforms. On the other hand, empirical research also indicates that handcrafted features are still competitive in specific questions, such as speech commands [2], voice spoof detection [7], and instrument classification [2].

Although there have been attempts to combine raw waveforms and handcrafted features in audio recognition [12], a general architecture for merging raw waveforms into networks that use handcrafted features, as well as the trade-offs in model complexity, has not been investigated thoroughly. This paper considers the use of waveforms as a supplement to handcrafted features and investigates their potential to boost performance with little additional computational cost.

3. THE AUXILIARY RAWNET ARCHITECTURE

3.1. Problem Formulation

Before introducing the Auxiliary RawNet (ARNet) architecture, we first formalize the problem it is intended to solve. Denote F_w as features of a raw waveform, and p as a problem to solve. We assume there is a constructive function f that can map $F_{p_{mag}}$, $F_{p_{phase}}$ and $S_{p_{noise}}$ into F_w, as described in Equation 1, where $F_{p_{mag}}$ is the ideal magnitude information needed to solve p, $F_{p_{phase}}$ is the ideal phase information needed to solve p, and $S_{p_{noise}}$ are signals with limited contribution to solving p (e.g., background noise).

$$F_w = f(F_{p_{mag}}, F_{p_{phase}}, S_{p_{noise}}) \tag{1}$$

Empirical studies [2] have shown the ability of handcrafted features to represent the strongest audio features for a variety of problems. Based on our assumption, the calculation of handcrafted features can be denoted as a mapping function g that can retrieve approximations of $F_{p_{mag}}$ or $F_{p_{phase}}$. For example, Mel-spectragrams can be described by the following equation:

$$F_{p_{mag}} \approx F_{mel} = g_{mel}(|STFT(F_w)|^2)) \tag{2}$$

When concatenating raw waveform data and handcrafted features to enhance model performance, our work is essentially to find a function, h, so that the total loss of $g(F_w)$ and $h(F_w)$ is smaller than a single $g(F_w)$. In other words, we want to find representations closer to the ideal solution $F_{p_{mag}} + F_{p_{phase}}$, as describe in Equation 3.

$$concat(g(F_w), h(F_w)) \approx F_{p_{mag}} + F_{p_{phase}} > g(F_w) \tag{3}$$

However, it is not clear how $g(F_w)$ interacts with $h(F_w)$. Inspired by observations from results regarding $g(F_w)$ and $h(F_w)$ on various tasks [2, 5], we make the following assumption about combining learned features and handcrafted features:

Assumption 1 (A1): *If a handcrafted feature, $g(F_w)$ shows strong results solving problem p, then there exists a $h(F_w)$ with size less than N in $concat(g(F_w), h(F_w))$ that will enhance overall performance. In other words, $h(F_w)$ can be an auxiliary component of $g(F_w)$ to improve performance with a bounded cost.*

3.2. The ARNet Structure

Based on the assumptions presented in section 3.1, we propose the ARNet architecturem which is shown in Figure 1. E_A, which processes the raw waveform, has a smaller bottleneck than E_M which processes handcrafted audio features, to make the raw waveforms play a supplementary role and bound the computational cost (e.g., bound N).

The Encoders. There are 3 encoders in the ARNet: the Main Encoder(E_M), Auxiliary Encoder(E_A), and Concatenate Encoder(E_C). E_M denotes the main encoder, whose inputs are the original handcrafted features that have shown good performance in solving the target problem. E_A is the encoder used to encode the raw waveforms in a light-weight way to compress F_w into F_a, where F_a are the features extracted by the auxiliary encoder. F_a and F_m (hand crafted features from the main encoder) are then concatenated in channels and further encoded by E_C.

Figure 1 shows the encoders used in our experiments on the ASVspoof 2019 dataset. We select the strided convolutional layer[4] as the first layer to directly process the raw waveforms. However, unlike previous raw waveforms networks, which include multiple CNN blocks with large kernels, the strided convolutional layer is only followed by three

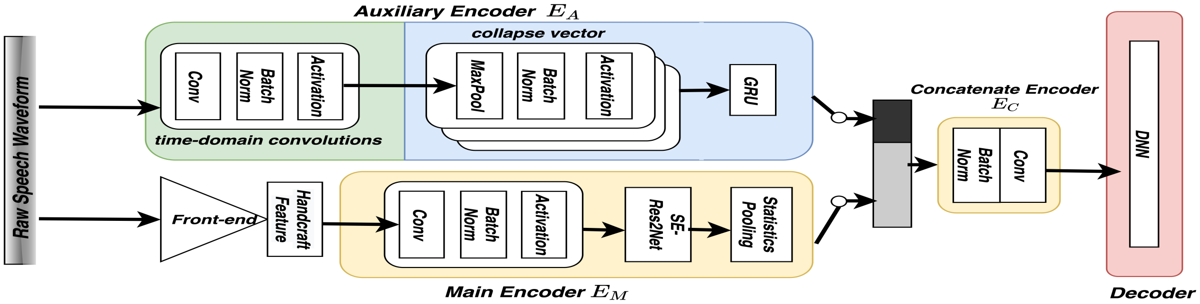

Fig. 1. The ARNet Architecture. E_A contains one strided CNN, 3 continuous max-pooling layers and a GRU. A TDNN-based model is illustrated here as an example of the E_M.

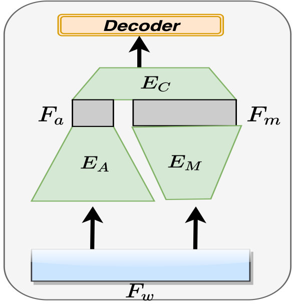

Fig. 2. Overview of the ARawNet. The model consists of a Main Encoder(E_M), Auxiliary Encoder(E_A), and Concatenate Encoder(E_C). E_A has a smaller bottleneck than E_M.

continuous pooling blocks to collapse vectors and remove any frame variance without further convolution. A GRU is used to encode frame-level features into utterance-level embeddings by keeping output vectors from the last time step.

The main encoder keeps layers before the statistical pooling layer, which will output utterance-level embeddings. Based on our assumption 1, we chose a narrow bottleneck for E_A. The dimension of the utterance-level embedding from E_A is designed to be smaller than the output dimension from E_M. Ultimately, E_C only contains a single Conv1d to encode concatenated results from E_A and E_M. The full architecture and model hyper-parameters are explained in Table 1.

Encoders	Blocks
Auxiliary Encoder	Conv(3,3,128) BN&LeakyReLu MaxPooling BN&LeakyReLu GRU(512)
Concatenate Encoder	BN Conv1D(1,1,256)

Table 1. The architecture Encoders.

The Decoder. In our problem, the decoder is a linear classifier layer that decodes embeddings from E_C to target classification.

3.3. Why do light-weight encoded raw waveforms augment handcrafted features?

Compared to the current filter-based architectures discussed in Section 2, we chose the strided convolutional receptive field, which is a standard CNN, as the first layer to process the raw waveforms. This layer consists of a set of time-domain convolutions, where all parameters(CNN kernel), are learned from the data. Calculation of the first CNN layer can be described as the following Equation [9], where x[n] is raw waveforms, h[n] is the filter and y[n] is filtered output:

$$y[n] = x[n] * h[n] = \sum_{0}^{L-1} x[l] \cdot h[n-l] \qquad (4)$$

As discussed in Section 3.1, concatenating $g(F_w)$ and $h(F_w)$ requires each encoder to have different attention to features in the raw waveforms so that they can complement each other. The standard convolutional layer with small kernels gives the E_A the least information about the signal processing mechanisms in $g(F_w)$, and thus potentially grants it the most flexibility to extract features, which do not overlap with $g(F_w)$.

In contrast to previous waveform-based networks [4, 5], the CNN blocks used in between the strided convolution layer and the GRU are completely removed, and only three continuous max-pooling layers with batch normalization are kept to collapse frame-level features step-by-step.

The first convolutional layer is considered the most critical part in processing raw waveforms. In deep networks it is also the most vulnerable to problems, such as vanishing gradients, without initializing filters [9]. However, based on our assumption 1, only significant frame-level features must be kept, indicating networks without deep CNN blocks can be used for E_A. Max pooling layers are used to collapse vectors and find significant pattern information that can be visualized after three pooling layers, as shown in Figure 3.

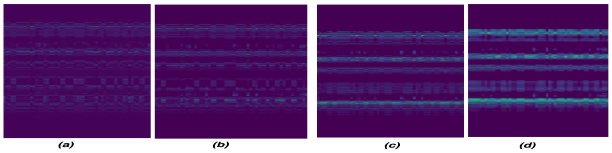

Fig. 3. Outputs visualization of the strided convolution layer and pooling layers. Outputs after 3 pooling layers(d) shows signification pattern information.

We test our assumption 1 based on the Theorem [13] from speech conversion problems, that if information bottlenecks between different encoders are precisely set, the model will decompose and produce disentangled representations of input speech signals. In our model, this Theorem can be described by the following equation:

$$E_M(F_w) = g(F_w), E_A(F_w) = h(F_w) \tag{5}$$

Thus, a narrow bottleneck is designed for E_A, which means the dimension of utterance-level embeddings dim_{E_A} is much smaller than dim_{E_M}.

4. EXPERIMENTAL SETUP

4.1. Experimental Dataset

The ASVspoof 2019 logical access (LA) dataset was developed to improve research on the growing threat of voice spoofing attacks on automated speech verification systems [7]. This dataset contains human-recorded audios and spoof audios generated from 19 sources (A01 - A19), including speech synthesis, voice conversion, and hybrid algorithms. We chose the ASVspoof 2019 LA dataset to validate the performance of our proposed model since:

- The performance of handcrafted features is limited by the difference in spoofing sources between the training and evaluation data.
- Current results on the ASVspoof 2019 challenge [7, 5] indicate that correct handcrafted features still provide the most competitive results from a single model compared raw waveforms approaches.
- Although pooling results of 19 spoof attacks is not satisfying, the waveforms-based network outperforms on the infamous A17 attacks [5].

4.2. Evaluation Metrics

Two metrics are used to evaluate the ASVspoof 2019 LA dataset including *min t-DCF* as the primary metric and $EqualErrorRate(EER)$ as a secondary metric, as described in [7]. The Tandem Detection Cost Function (t-DCF) [14] extends the conventional Detection Cost Function (DCF) in voice verification systems for spoofing attacks. The t-DCF measures the overall effect of CM systems combined with existing ASV systems. EER indicates the threshold of a CM system where the false positive and false negative rates are equal each to other.

4.3. Baseline Setup

Our experiments include one handcrafted feature-based system and one raw waveforms-based system respectively:

Res2net Architecture. The Res2net architecture [6] is the state-of-the-art single system in the ASVspoof 2019 challenge, which tested the performance of three handcrafted features: log power magnitude spectrogram (Spec), linear frequency cepstral coefficients (LFCC), and constant-Q transform (CQT).

RawNet2. The RawNet2 [5] is the first anti-spoofing model, which only relies on the raw waveforms as input. It shows good performance on the A17 attack.

5. RESULTS AND ANALYSIS

Table 2 shows the experimental results of the ARNet on the ASVSpoof 2019 dataset. These results demonstrate the effectiveness of adding a light-weight auxiliary encoder to the main encoder. Two handcrafted features, Mel-spectrogram and CQT [15], as well as two state-of-the-art models in the speaker verification problem (XVector [16, 17] and ECAPA-TDNN [18, 17]) were selected as main encoders in the ARNet architecture. Without modifying the hyper-parameters in the main encoder, we added the auxiliary encoder, as described in Table 1, in the network to evaluate our assumption. Overall, by introducing the auxiliary encoder, both EER and $min-tDCF$ are reduced by $\tilde{5}0\%$ in all combinations of front-end and main encoders. Specifically, CQT/ECAPA-TDNN with auxiliar encoder achieved the best performance on EER of 1.11% and $min-tDCF$ of 0.0364.

	Front-end	Main Encoder	E_A	EER	min-tDCF
[6]	Spec	Res2Net[6]	-	8.783	0.2237
	LFCC		-	2.869	0.0786
	CQT		-	2.502	0.0743
[5]	Raw waveforms	Rawnet2[5]	-	5.13	0.1175
Ours	Mel-Spectrogram	XVector	✓	**1.32**	0.03894
			-	2.39320	0.06875
Ours	Mel-Spectrogram	ECAPA-TDNN	✓	**1.39**	0.04316
			-	2.11	0.06425
Ours	CQT	XVector	✓	**1.74**	0.05194
			-	3.39875	0.09510
Ours	CQT	ECAPA-TDNN	✓	**1.11**	0.03645
		.	-	1.72667	0.05077

Table 2. Results on the ASVspoof 2019 dataset

Table 3 compares the number of trainable parameters and model complexity, multiply-and-accumulates (MACs) in our experiments. Compared to encoding handcrafted features (Res2Net), directly encoding raw waveforms (Rawnet2) increases model size and complexity by 2400% and 600%. In contrast, our auxiliary waveforms encoder only takes up 1.15M trainable parameters, which is a 19% increase in ECAPA-TDNN and the model complexity increases from 2.36 GMac to 3.19 GMac, i.e., the performance of our model increases by 28.2% with increments of 35.1% MACs.

Main Encoder	E_A	Parameters	MACs
Rawnet2	-	25.43 M	7.61 GMac
Res2Net	-	0.92 M	1.11 GMac
XVector	✓	5.81 M	2.71 GMac
XVector	-	4.66M	1.88 GMac
ECAPA-TDNN	✓	7.18 M	3.19 GMac
ECAPA-TDNN	-	6.03M	2.36 GMac

Table 3. Comparison of model complexity (MACs)

6. CONCLUDING REMARKS

This paper explored the problem of combining learned features and handcrafted featured in the audio field. We also described the ARNet architecture, which combines hand-crafted features and raw waveforms to complement each other without sacrificing model complexity. We tested two hand-crafted features (Mel-spectrogram and CQT) and two state-of-the-art models (XVector and ECAPA-TDNN) as the main encoder with our Auxiliary Encoder. Results from our experiments showed that raw waveforms have a general complementing ability to handcrafted features in the ASVspoof 2019 dataset. The code described here is available as open-source from `github.com/magnumresearchgroup/AuxiliaryRawNet`.

7. REFERENCES

[1] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[2] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[3] Jee-Weon Jung, Hee-Soo Heo, IL-Ho Yang, Hye-Jin Shim, and Ha-Jin Yu, "Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification," *extraction*, vol. 8, no. 12, pp. 23–24, 2018.

[4] Jee-weon Jung, Hee-Soo Heo, Ju-ho Kim, Hye-jin Shim, and Ha-Jin Yu, "Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," *arXiv preprint arXiv:1904.08104*, 2019.

[5] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[6] Xu Li, Na Li, Chao Weng, Xunying Liu, Dan Su, Dong Yu, and Helen Meng, "Replay and synthetic speech detection with res2net architecture," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6354–6358.

[7] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[8] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[9] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[10] Dimitri Palaz, Mathew Magimai Doss, and Ronan Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4295–4299.

[11] Paul-Gauthier Noé, Titouan Parcollet, and Mohamed Morchid, "Cgcnn: Complex gabor convolutional neural network on raw speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7724–7728.

[12] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[13] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[14] Tomi Kinnunen, Kong Aik Lee, Héctor Delgado, Nicholas Evans, Massimiliano Todisco, Md Sahidullah, Junichi Yamagishi, and Douglas A Reynolds, "t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," *arXiv preprint arXiv:1804.09618*, 2018.

[15] Christian Schörkhuber and Anssi Klapuri, "Constant-q transform toolbox for music processing," in *7th sound and music computing conference, Barcelona, Spain*, 2010, pp. 3–64.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[17] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[18] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.