ON THE IMPORTANCE OF DIFFERENT FREQUENCY BINS FOR SPEAKER VERIFICATION

Aiwen Deng[1], *Shuai Wang*[2*], *Wenxiong Kang*[1], *Feiqi Deng*[1]

[1]South China University of Technology, Guangzhou, China
[2] Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

The majority of modern speaker verification systems take spectral analysis-based features as input, which contains multiple frequency bins. Naturally, there would be a question of whether all different frequency bins contribute equally to the speaker verification system performance? In this paper, we propose the frequency reweighting layer (FRL) to automatically learn and balance the importance of different frequency bins. This new layer can be freely inserted into the original speaker embedding learner once or multiple times at different layers, with an ignorable number of new parameters. Based on the proposed novel architecture, a set of experiments are designed and carried out on the VoxCeleb1 dataset, which not only achieves superior performance but also exhibits an interesting weight distribution – the lower frequencies matter more.

Index Terms— frequency bin, attention, speaker verification

1. INTRODUCTION

Speaker verification aims to recognize one's identity via his or her speech. The key to performing this task is to accurately model the speaker's identity, which is closely related to the feature design and model architecture.

From the early Gaussian Mixture Model (GMM) era [1, 2, 3], to the current cutting-edge deep embedding learning methods [4, 5, 6], the mainstream features used are based on spectral analysis. However, spectral features such as MFCC and the log mel filter-bank features (Fbank) were initially designed for the speech recognition task, which means they are not necessarily the most suitable for the speaker recognition task. To this end, different features were proposed for the speaker verification task. For instance, raw-wave-based input was investigated in [7, 8], showing promising results on the VoxCeleb dataset. In this framework, the authors count on the carefully designed network architecture to automatically learn the useful information from the raw wave signals. However, new feature design poses a high expertise requirement for researchers. Ideal features for speaker recognition should be speaker-discriminative, robust to variation caused by phonetic information, channel effects, .etc. Consequently, although different features have been proposed, from the most recent speaker verification challenges such as SRE [9, 10] and VoxCeleb [11, 12], the spectral analysis based feature such as Fbank is still the most popular and competitive one, while different feature enhancement methods were proposed (Will be reviewed in Section. 2).

It's well-known the Fbank feature is computed through the Short-Term Fourier Analysis (STFT), and each dimension corresponds to the information encoded in different frequency ranges. In the speaker embedding learning process, all dimensions are usually treated as a whole input to the model without considering the importance of different frequency bins. However, we believe not all frequency ranges contribute to the system performance equally, it might be beneficial if we treat different frequency dimensions differently and explicitly guide the network to consider this difference by model design. Thus, we propose the Frequency Reweighting Layer (FRL), which learns the importance of different frequency dimensions from the training set automatically and reweights them accordingly.

Our contribution in this paper can be listed as follows,

- We proposed the frequency reweighting layer (FRL), which automatically learns the importance of different frequency bins of features with ignorable computation cost in the inference process.
- We proposed the multi-layer FRL integration architecture, which greatly boosts the system performance.
- We illustrated the importance distribution of different frequency bins via the weight visualization and mask probing experiment.
- We demonstrated that the new proposed architecture has superior robustness.

The rest of this paper is organized as follows, Section. 2 briefly reviews the related work which considered the impact of different frequencies, Section. 3 introduces our proposed frequency reweighting network. We then describe our experimental setups and results, including detailed analysis on the

* Shuai Wang is the corresponding author

importance of different frequency bins in Section 4 and 5. Finally, Section. 6 concludes the paper.

2. RELATED WORK

The human voice can be mapped into multi-dimensional spectral features via signal processing algorithms. Different bins usually correspond to the information encoded in different frequency bins (low to high). As mentioned in the introduction, such features are not optimal for the speaker recognition task. Thus many researchers have investigated different feature enhancement methods. Such methods are usually done by corresponding network architecture design. [13] used Context Aggregation Network (CAN) to generate Time-Frequency (TF) bins attention mask to denoise the input spectral features. [14] utilized multiple Conv-MS blocks to generate a ratio mask matrix to weight the input spectrogram by multiplying it by the corresponding frequency bins and time frames. In addition, [15] proposed fine-grained early frequency attention, enabling the network to attend to information items frequency bins without a significant increase in the number of parameters and complexity. Our proposed frequency reweighting network is similar, which can also be regarded as a feature enhancement method. However, it should be noticed that our method mainly differs from the above-mentioned attention-based methods considering the following aspects,

- The above-related works require query-key operations given the input features, but our proposed method is input-independent.
- The extra parameters introduced by our method are tuned on the whole dataset, only considering different feature dimensions.

The above properties of our proposed method give two main advantages, 1) There would be nearly no additional computation cost compared to the original network architecture. 2) Our method learns a more general pattern of the importance of different feature dimensions, which eases further analysis on the importance of different feature dimensions.

3. FREQUENCY REWEIGHTING NETWORK FOR SPEAKER VERIFICATION

In this section, we will first describe the proposed Frequency Reweight Layer (FRL) module and then introduce the overall speaker embedding learning architecture where the FRLs are integrated.

3.1. Frequency Reweight Layer

The proposed FRL is designed to focus on each frequency bin individually by generating a scale parameter for each frequency bin. The entry of each frequency dimension is multiplied by the corresponding scale parameter to obtain the intermediate output. To reflect the general impact of different frequency bins of input spectral features on SV tasks, we design the generated scale to be independent of the input features. From this perspective, the FRL is different from the attention-based reweight methods[15] which requires input features as keys to calculate attention weights. This design implies that the automatically learned reweighting parameters would be independent of individual samples and learns the feature importance distribution from the whole dataset.

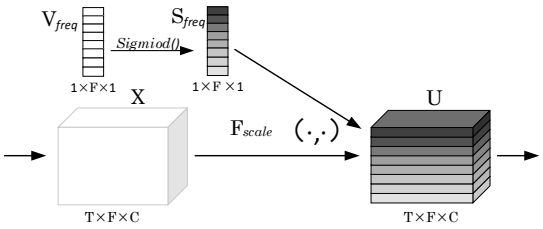

Fig. 1. The frequency reweighting layer architecture

The FRL structure is illustrated in Figure 1. It is a reweighting operation where the input feature $\mathbf{X} \in R^{C\times F\times T}$ is multiplied by the scale weight vector $\mathbf{S}_{\text{freq}} = [s_1, s_2, \ldots, s_F]$ to obtain the output $\mathbf{U} \in R^{C\times F\times T}$, and C, F, T represents the number of input channels, feature frequency dimension and sequence length, respectively. The calculation formula is as follows.

$$\mathbf{U} = F_{scale}(\mathbf{S}_{\text{freq}}, \mathbf{X}) \tag{1}$$

Where F_{scale} denotes the corresponding scale factor multiplied by each frequency dimension features of the input features $\mathbf{X}$. The scale vector $\mathbf{S}_{\text{freq}} = [s_1, s_2, \ldots, s_F]$ is obtained from the learnable network parameter vector $\mathbf{V}_{\text{freq}} = [v_1, v_2, \ldots, v_F]$ via the Sigmoid function.

$$s_i = Sigmoid(v_i) \tag{2}$$

where $i \in [1, F]$. In addition, if we integrate multiple FRL modules at different layers, residual connections are added to avoid information loss, leading to a final output $\mathbf{Y}$ formulated as:

$$\mathbf{Y} = \mathbf{X} + \mathbf{U} \tag{3}$$

3.2. FRL integration for speaker embedding learning

ResNet-based architecture has been widely used as a speaker embedding extractor in SV tasks [16, 17, 18]. In this work, it's also used as our speaker embedding learner. The original design of FRL is to reweight the input spectral features. However, in practice, the proposed FRL can be inserted after the original input spectrogram or any layer of the network to reweight the feature maps' frequency dimension. So we proposed the multi-layer FRL integration architecture, as shown

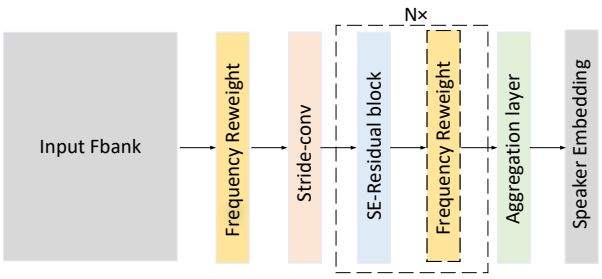

Fig. 2. Overall system structure, the newly proposed frequency reweight layer can be freely inserted into original architecture

in Figure 2. The FRL in the proposed framework is mainly placed after the original input spectral features and each SE-Residual block. In fact, it is unnecessary to add the FRL after each SE-Residual block, because the resolution of the frequency dimension of the feature map is lower for the later blocks due to the network downsampling. In this case,the feature frequency dimensions are highly correlated, so it is difficult to analyze them at a fine-grained level. Therefore, the proposed structure only considers combining the FRL on the first two blocks, and the specific effect can be seen in the experimental section.

4. EXPERIMENTAL SETUP

4.1. Dataset

All the following experiments are carried out on the VoxCeleb1 dataset. VoxCeleb is a large scale text-independent speaker recognition dataset comprised of two releases, VoxCeleb1 [11] and VoxCeleb2 [12]. Note that we only use VoxCeleb1 in this paper. VoxCeleb1 contains over 150000 utterances from 1251 different celebrities. For the speaker verification task, part 1 was split into the training part and the evaluation part. The training part contains 148642 utterances from 1211 celebrities, while the evaluation set contains about 4874 utterances from the rest 40 celebrities. The standard trial list for the verification contains 37720 pairs.

4.2. Implementation Details

For the feature preparation, 80-dimensional Fbank features with a frame length of 25 ms and a 10ms frameshift were used as input features. A Fbank with a fixed length of 200-frames is randomly selected from the input utterances as input during training. Neither voice activity detection (VAD) nor data augmentation is used for training, and data preparation is done with the Kaldi toolkit.

For the network architecture, we adopt a Fast-ResNet34 network[16] with Squeeze-and Excitation (SE) block[19] as our speaker embedding learning back-bone. The Fast-ResNet34 is the same as the original ResNet with 34 layers, except that it uses only one-quarter of the channels in each residual block and the strides are earlier in order to reduce computational cost. Attentive-statistics pooling (ASP) [20] is used as an aggregation layer and is passed through the fully connected layer to obtain 512-dimensional embeddings. The loss function is a combination of AM-softmax[21] and Angular Prototypical loss[16].

All of the models are trained using the Adam optimizer, and the weight decay is set to $5e$-5. The learning rate is initially 0.005 and is decayed to half of its until convergence with patience of 5. Each mini-batch contains 128 speakers and 256 utterances for Angular Prototypical loss which the size of both the support set and the query set are set to 1. The margin and scaling factor of AM-Softmax loss are set to 0.1 and 30.

For the evaluation, standard metrics like Equal Error Rate (EER) and Minimum Decision Cost Function (minDCF) at $P_{\text{target}} = 0.01$ and $C_{\text{FA}} = C_{\text{Miss}} = 1$ are used. Cosine similarity is used as the distance metric.

5. RESULTS

5.1. Effect of the Frequency Reweighting Layer

Using the model described in section. 4 as the baseline model, we first verify the effect of inserting the proposed FRL at different layers. We only verify that the FRL is inserted after the input layer and the first two blocks because the frequency resolution of the feature map gets smaller as the layer deepens. The frequency features are decomposed into channel information.

System	EER	minDCF	New Params
Baseline	3.35	0.383	0
+ inp	3.23	**0.367**	80
+ lay1	3.24	0.382	40
+ lay2	3.27	0.370	20
+ inp + lay1 + lay2	**2.99**	0.372	140

Table 1. EER and MinDCF performance of the systems on the standard VoxCeleb1-test. The inp, lay1, lay2 denote the FRL insertion after the input Fbank features, the first SE-Residual block, and the second layer SE-Residual block, respectively.

A performance overview of the baseline systems and the FRL is given in Table 1. In this experiment, we try to insert the FRL on the baseline model after the input Fbank features, the first and the second SE-Residual block, respectively, to verify the effect of FRL. It can be seen that FRL works for all the cases considering different insertion positions. The numbers of new parameters added to the model are only 80, 40, 20,

respectively. Further, the multi-layer architecture combining these three layers of FRL can achieve 10.75% EER relatively reduction, with only 140 extra parameters are needed. This demonstrates that the effectiveness of the proposed FRL.

5.2. Visualization of learned weight

As described in Section 3, the weight of the FRL is input independent, so it is able to reflect the importance of each frequency bin of the input Fbank features for the SV task. The visualization of the FRL weights distribution of the input layer after the sigmoid function is shown in Figure 3 (a). It can be seen that for the SV task, the low-frequency part of the input Fbank features obtains significantly higher weights compared with the high-frequency part. To verify the generality of this weight distribution curve, we retrain the model by swapping the first 40 dimensions of the input Fbank features with the last 40 dimensions. The obtained FRL weight distribution is shown in Figure 3 (b), it shows that the original low-frequency parts of Fbank will be still given a higher weight. This demonstrates that the FRL can learn the speaker task-related frequency weight distribution from the training data.

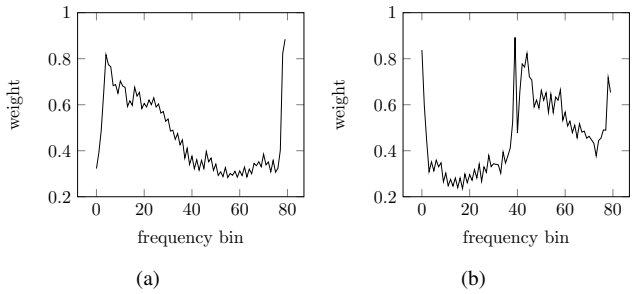

Fig. 3. Visualization of the learned weight. (a) the FRL weight distribution for normal input Fbank features; (b) the FRL weight distribution of the input spectrum by swapping the first 40 dimensions of the input Fbank features with the last 40 dimensions.

5.3. Effect of masking different frequency bins

In order to verify whether the weight distribution can match the importance distribution of different frequency bins of the input Fbank features, we mask a single frequency bin of the input Fbank features for all test samples, and the results of the EER reflect the importance of that frequency for the SV task. The curve of mask frequency bin vs. EER results for the baseline model and the proposed Frequency Reweight Network(FRN) model is shown in Figure 4 (a) and (b). The horizontal axis is the frequency bin of the input Fbank features being masked. The mask result curves of both the baseline model and the model with the FRL reflect that the impact of low-frequency features on the model performance is more significant than that of high-frequency parts. This is generally consistent with the FRL weight distribution.

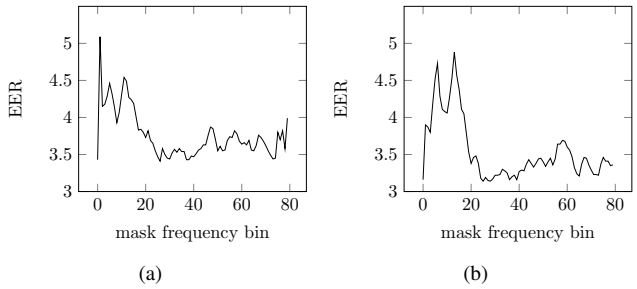

Fig. 4. (a)the baseline model frequency bin masking results; (b)the FRN model frequency bin masking results.

Moreover, we further test the robustness of the FRL against frequency perturbations by performing random frequency masking on the test samples. Using the frequency mask module of SpecAugment to set different maximum mask lengths, the results obtained are shown in Table 2. As the results show, the model's performance with FRL receives less impact as the maximum mask length grows.

Max mask length	0	5	10	15	20
Baseline	3.35	5.03	8.25	12.40	16.62
Baseline+inp	3.23	4.73	7.00	10.45	15.06
ΔEER	0.12	0.3	1.25	1.95	1.56

Table 2. Frequency mask robustness test results. The values in the table indicate EER. The inp denote the FRL insertion after the input Fbank features.

6. CONCLUSION

As the most popular feature used for the speaker verification task, there is not much effort towards analyzing the importance of the information encoded by different frequency bins. To this end, we proposed a novel frequency reweighting network architecture to automatically learn the importance of different frequency dimensions. Furthermore, we apply this architecture to different layers of the speaker embedding learner, along with residual connections, leading to impressive performance improvement. Via the weight visualization and the mask probing experiment, we show that the system performance attributes more to the lower frequencies.

7. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant 61573151 and Grant 61976095, in part by the Science and Technology Planning Project of Guangdong Province under Grant 2018B030323026.

8. REFERENCES

[1] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[3] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[4] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[5] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in *Sixteenth annual conference of the international speech communication association*, 2015.

[6] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification.," in *Interspeech*, 2017, pp. 999–1003.

[7] Jee-Weon Jung, Hee-Soo Heo, Il-Ho Yang, Hye-Jin Shim, and Ha-Jin Yu, "A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5349–5353.

[8] 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.

[9] Alvin F Martin and Craig S Greenberg, "The nist 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, Jaime Hernandez-Cordero, et al., "The 2016 nist speaker recognition evaluation.," in *Interspeech*, 2017, pp. 1353–1357.

[11] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[12] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[13] Saurabh Kataria, Phani Sankar Nidadavolu, Jesús Villalba, Nanxin Chen, Paola Garcia-Perera, and Najim Dehak, "Feature enhancement with deep feature losses for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7584–7588.

[14] Yanpei Shi, Qiang Huang, and Thomas Hain, "Robust speaker recognition using speech enhancement and attention model," *arXiv preprint arXiv:2001.05031*, 2020.

[15] Amirhossein Hajavi and Ali Etemad, "Knowing what to listen to: Early attention for deep speech representation learning," *arXiv preprint arXiv:2009.01822*, 2020.

[16] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," *arXiv e-prints*, pp. arXiv–2003, 2020.

[17] Chengfang Luo, Xin Guo, Aiwen Deng, Wei Xu, Junhong Zhao, and Wenxiong Kang, "Learning discriminative speaker embedding by improving aggregation strategy and loss function for speaker verification," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[18] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[19] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[20] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive statistics pooling for deep speaker embedding," *arXiv preprint arXiv:1803.10963*, 2018.

[21] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.