FAtNet: Cost-Effective Approach Towards Mitigating the Linguistic Bias in Speaker Verification Systems

Divya V Sharma
IIIT-Delhi
divyas@iiitd.ac.in

Arun Balaji Buduru
IIIT-Delhi
arunb@iiitd.ac.in

Abstract

Linguistic bias in Deep Neural Network (DNN) based Natural Language Processing (NLP) systems is a critical problem that needs attention. The problem further intensifies in the case of security systems, such as speaker verification, where fairness is essential. Speaker verification systems are intelligent systems that determine if two speech recordings belong to the same speaker. Such human-oriented security systems should be usable by diverse people speaking varied languages. Thus, a speaker verification system trained on speech in one language should generalize when tested for other languages. However, DNN-based models are often language-dependent. Previous works explore domain adaptation to fine-tune the pre-trained model for out-of-domain languages. Fine-tuning the model individually for each existing language is expensive. Hence, it limits the usability of the system. This paper proposes the cost-effective idea of integrating a lightweight embedding with existing speaker verification systems to mitigate linguistic bias without adaptation. This work is motivated by the theoretical hypothesis that attentive-frames could help generate language-agnostic embeddings. For scientific validation of this hypothesis, we propose two frame-attentive networks and investigate the effect of their integration with baselines for twelve languages. Empirical results suggest that frame-attentive embedding can cost-effectively reduce linguistic bias and enhance the usability of baselines.

1 Introduction

Mitigating the linguistic bias in Deep Neural Network (DNN) based models is one of the critical challenges in Natural Language Processing (NLP). The linguistic bias, specifically in the security systems, such as speaker verification models, is a far more critical problem requiring much research. Speaker verification systems are biometric authentication systems that use speech signals to authenticate a speaker. These systems use the fact that every speaker has unique traits in their voice (Hansen and Hasan, 2015). Such systems have real-world applications in e-commerce, forensics, law, business, and access control mechanisms (Hansen and Hasan, 2015). These systems can be text-dependent or text-independent (Hansen and Hasan, 2015). Text-independent speaker verification systems are more user-friendly than text-dependent systems. These systems authenticate a speaker without any constraint on the content of speech.

However, speaker verification models often tend to be language-dependent (Auckenthaler et al., 2001). The reason is that a robust speaker verification system would require memory to analyze the sequential speech data and capture relevant discriminatory information. Memory helps in remembering past information. Remembering the past and predicting the future can contribute to linguistic content in the embedding (Shain and Elsner, 2020). Therefore, the generated embedding for speaker verification may contain linguistic detail.

Language-dependent speaker verification models perform relatively well on test sets containing speech recordings in the same language as the training set. However, the performance of these systems degrades on test sets containing speech recordings in different languages. The majority of the publicly available speech datasets are in English. It is a tedious task to get labeled datasets for various low-resource languages. Most of the previous works use domain adaptation to improve the performance of speaker verification models only for a limited set of languages (Rohdin et al., 2019; Xia et al., 2019; Chen et al., 2020). It is also costly to fine-tune a pre-trained speaker verification model individually for each existing language. Further, studies show that the linguistic content in the embedding increase with the temporal scope of representations (Chrupała et al., 2020).

Our proposed work is based on the theoretical hypothesis that frame-level features contain less lin-

Findings of the Association for Computational Linguistics: NAACL 2022, pages 1247 - 1258
July 10-15, 2022 ©2022 Association for Computational Linguistics

guistic information due to the low-temporal scope of frames. Thus, frame-level features may help generate a language-agnostic embedding. Furthermore, an intelligent selection of frame-level features may help in enhancing the model's generalizability to out-of-domain testing. We aim to address the problem of language dependency in text-independent speaker verification systems cost-effectively, without the overhead of domain adaptation. We propose an idea that incorporates a lightweight embedding with existing speaker verification systems which may help in improving the generalizability of these systems to out-of-domain testing. To scientifically validate the theoretical hypothesis, we present and investigate two variants of frame-attentive networks: FAtNet-v1 and FAtNet-v2. Our proposed models accept two speech recordings as input and determine if they belong to the same speaker. The speakers in the trial pair may be unknown. We qualitatively compare the generalization ability of our proposed models with two strong baselines on four publicly available data sets. We perform quantitative experiments on 12 languages to assess the integration of our proposed FAtNet embeddings with the baselines on publicly available out-of-domain test sets without domain adaptation. We have released the code[1] to encourage more research on this problem.

We summarize our main contributions below:

1. Investigate cognitive ideas such as attention, residual connection for memory, and learning parameters to generate language-agnostic embeddings.
2. To validate the theoretical hypothesis scientifically, propose two novel frame-attentive networks: FAtNet-v1 and FAtNet-v2.
3. Perform qualitative and quantitative experiments for twelve languages using two strong baselines and four publicly available datasets.

2 Background and Motivation

Language dependency in speaker verification: The current state-of-the-art explores deep neural networks (DNN) to solve speaker recognition problems (Hansen and Hasan, 2015; Li et al., 2018; Jung et al., 2020, 2019; Nagrani et al., 2017; Snyder et al., 2018; Nagrani et al., 2020; Guzewich et al., 2018; Zhao et al., 2019; Gao et al., 2018). However, most DNN-based feature extractors are language-dependent (Oleg et al., 2016). Language dependency can make the system less usable as users may belong to different geographic locations and speak varied languages. It is challenging to get labeled datasets for various low-resource languages (Brignatz et al., 2021). Moreover, when tested on multilingual datasets and features, the models that show consistent behavior may be helpful in other applications (such as code-switching) through information sharing (Belinkov et al., 2019). We know that there are approximately 7,000 languages in the world (Huang et al., 2021). One of the critical challenges in Natural Language Processing (NLP) is to develop techniques to overcome this linguistic bias and enhance the usability of the model across the globe (Huang et al., 2021).

Recent works: Transfer learning is a solution to address the problem of domain mismatch. However, it is challenging to get labeled datasets for various low-resource languages (Brignatz et al., 2021). Recent works investigate adversarial domain adaptation techniques for solving cross-lingual speaker verification problems (Rohdin et al., 2019; Xia et al., 2019; Chen et al., 2020; Brignatz et al., 2021). However, most of these approaches can improve the performance of speaker verification models for a limited set of languages as these approaches require an additional overhead of domain adaptation.

To the best of our knowledge, (Chojnacka et al., 2021) is the closest work related to our problem statement where the authors attempt to reduce linguistic bias in speaker verification without domain adaptation. In (Chojnacka et al., 2021), the authors suggest that training a speaker verification model in multiple languages can increase its generalizability to out-of-domain languages. However, they trained the model on an extensive training set consisting of 1,96,000 speakers and 2,06,18,000 utterances. Training on such an extensive dataset requires significant computational requirements, often not feasible in a realistic scenario. In addition to that, their proposed work involves a combination of text-dependent and text-independent speaker verification systems. Our proposed method involves integrating a lightweight embedding with the existing text-independent speaker verification models to reduce linguistic bias in those systems.

Linguistic components in a frame: Speech signals are non-stationary, and hence they are divided into frames. A speech signal is assumed to be stationary within a frame (Malek, 2020). The tem-

[1] https://github.com/vdivyas/FAtNet.git

poral scope of a frame is usually a few milliseconds. Researchers have investigated the role of temporal scope in their study of phonology in neural models (Chrupała et al., 2020). Studies show that Representational Similarity Analysis (RSA) applied to local representations results in lower correlations between phonemes and neural activation patterns (Chrupała et al., 2020). Thus, considering the low temporal scope of frames, it is intuitive that they carry less linguistic information.

Theoretical Hypothesis: Our proposed work investigates the theoretical hypothesis that utterance-level embedding captures more linguistic information than frame-level embedding. Therefore, a frame-level embedding can be more language-robust than an utterance-level embedding. This abstract knowledge of frame-level features can allow the model to learn relevant discriminatory information from frames and ignore the linguistic information from speech. Researchers have stated that some frames tend to be more critical than others for the final-encoded representation of speech (Havard et al., 2019). Attention mechanisms are popular in state-of-the-art speaker verification models (Zhu et al., 2018; Okabe et al., 2018). We explore the effectiveness of attention for an intelligent selection of features within a frame.

3 Proposed Approach

To investigate the theoretical hypothesis, we propose the following two variants of Frame-Attentive Networks: FAtNet-v1 and FAtNet-v2.

As illustrated in Figure 1 and Figure 2, the time-delay neural network (TDNN) paths are similar in both the FAtNet versions and the details are as follows: The model accepts a pair of Mel-frequency cepstral coefficients[2] (MFCC) for speaker verification (Chen et al., 2020; Zhu et al., 2018; Khoury et al., 2014). MFCCs finetune the features to what human beings hear (Lyons). Let d be the dimension of input MFCCs and l_1 and l_2 be the number of frames in the given pair for speaker verification. The values of l_1 and l_2 may differ due to the duration variability issue. The models were trained on 3-second chunks of speech (Nagrani et al., 2020). Eighty-dimensional MFCCs of shape (94,80) generated using these audio clips for training are input to the model. As shown in Figure 1 and Figure 2, we pass each of the input MFCCs to the AdaptiveAvgPool2d layer to get features of shape (b, 94, 80), where b is the batch-size (Yu et al., 2019). It allows the model to accept variable duration speech recordings during test time without any special augmentation strategy. It facilitates easy integration of FAtNet embedding with other speaker verification models and enhances the usability of the models.

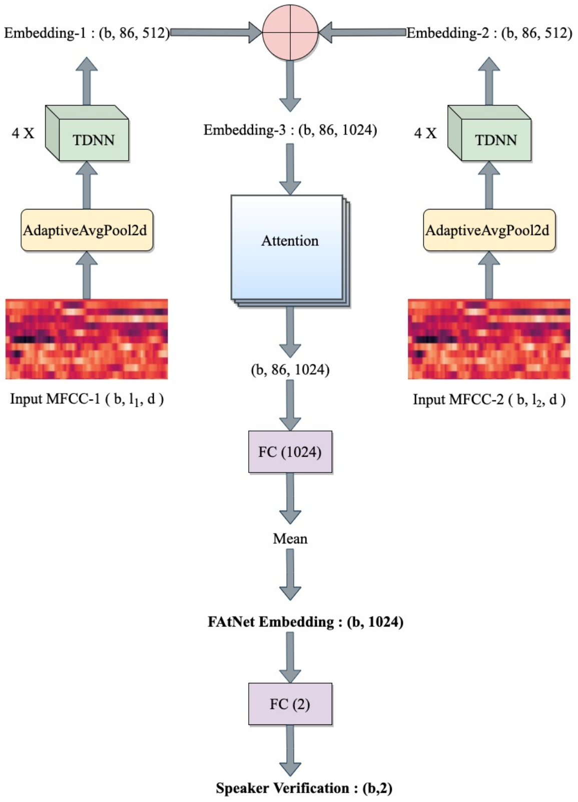
Embedding-1 : (b, 86, 512)
Embedding-2 : (b, 86, 512)
4 X
TDNN
Embedding-3 : (b, 86, 1024)
4 X
TDNN
AdaptiveAvgPool2d
Attention
AdaptiveAvgPool2d
Input MFCC-1 (b, l_1, d)
(b, 86, 1024)
Input MFCC-2 (b, l_2, d)
FC (1024)
Mean
FAtNet Embedding : (b, 1024)
FC (2)
Speaker Verification : (b,2)

Figure 1: Architecture diagram for FAtNet-v1.

The next step is to compute frame-level features for further analysis. The abstract knowledge of frame-level features can reduce the linguistic information in the final embedding. We use four stacked TDNN[3] layers to extract the frame-level features (Vijayaditya Peddinti, 2015). Given two speech recordings as input, the problem is to determine if they belong to the same speaker or not. Thus, we have two such TDNN paths for input audio clips.

FAtNet-v1: We concatenate the frame-level features obtained for both the input speech recordings as shown in Figure 1. We apply batch normalization. We further pass these concatenated features through an eight-head frame-level attention block. In each frame, attention gives more weight to relevant features.

[2] We compared the performance of 80-dimensional MFCCs with 300-dimensional spectrogram (Nagrani et al., 2017) as inputs to FAtNet-v1. Details are present in the Appendix section.

[3] Hyper-parameter detail for the stacked TDNN layers is present in the Appendix section.

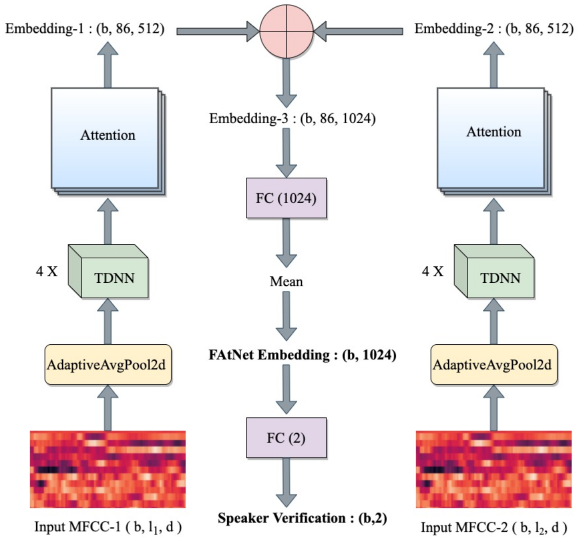
Embedding-1 : (b, 86, 512)
Embedding-2 : (b, 86, 512)
Embedding-3 : (b, 86, 1024)
Attention
Attention
FC (1024)
4 X
TDNN
Mean
4 X
TDNN
AdaptiveAvgPool2d
FAtNet Embedding : (b, 1024)
AdaptiveAvgPool2d
FC (2)
Input MFCC-1 (b, l1, d)
Speaker Verification : (b,2)
Input MFCC-2 (b, l2, d)

Figure 2: Architecture diagram for FAtNet-v2.

FAtNet-v2: We pass the frame-level features obtained for the input speech recordings through separate four-head attention blocks. Thus, we get embeddings 1 and 2 as shown in Figure 2. In each frame, the attention block gives more weight to relevant features. We then concatenate the outputs to get embedding 3.

Attention: Attention mechanisms are popular in state-of-the-art speaker verification models (Zhu et al., 2018; Wu et al., 2020). Our proposed FAtNet attention mechanism is inspired by (Vaswani et al., 2017; Moshnoi). The frame-level features are sent through a multi-head residual self-attention block, as shown in Figure 1 and Figure 2, respectively. The input to the attention block can be a tensor of shape (b, l, d) where b is the batch size, l is the number of frames, and d is the number of features or dimensions in each frame. Let dv be the dimension of linear space where the input needs to be projected, and nv is the number of heads in the multi-head self-attention block. We pass the same tensor to the attention block as the query tensor, key tensor, and, value tensor. The idea is to use the query tensor and key tensor to generate a weight tensor for the value tensor. The first step involves passing each of the three tensors: query, key, and value through separate fully-connected layers consisting of $dv * nv$ output units and applying the $ReLU$ activation function to get the modified query, key and value tensors, say Q, K, and V, respectively. After adequate reshaping, the dimensions of Q, K, and V should be (b, l, nv, dv).

For each example i, the following computation is performed within the attention block using Q, K, and V :

1. $K_{\text{i}}{}^{\text{permute}} := K_{\text{i}}.permute(0, 2, 1)$
2. $prob_{\text{i}} := Q_{\text{i}} * K_{\text{i}}{}^{\text{permute}}$
3. $prob_{\text{i}}{}^{\text{scaled}} := \frac{prod_{\text{i}}}{\sqrt{dv}}$
4. $weights_{\text{i}}{}^{\text{attn}}$:=Softmax($prod_{\text{i}}{}^{\text{scaled}}$, dim = -1)
5. $rprod_{\text{i}} := weights_{\text{i}}{}^{\text{attn}} * V_{\text{i}}$

We further include a residual connection that acts as a memory to combine the initial set of frame-level features with $rprod_{\text{i}}$. So, to compute the output of the residual attention block, we add $rprod_{\text{i}}$ to the original query tensor and pass it through a fully-connected layer consisting of d_{out} neurons and apply $ReLU$.

The remaining layers are similar in both FAtNet-v1 and FAtNet-v2. The details are as follows: The generated attentive-frames are batch normalized. Then these are passed through a fully connected layer for fine-grained analysis. We further apply a leaky-relu activation function with L_2-normalization. We aggregate these frame-level features by computing a mean and then pass them through a fully connected layer consisting of two units for speaker verification.

4 Experimental Setup

4.1 Datasets

Training datasets: We trained separate models of FAtNet-v1 on publicly available VoxCeleb-1[4] and VoxCeleb-2 dev sets[5] (Nagrani et al., 2017; Chung et al., 2018). FAtNet-v2 was trained on VoxCeleb-2 dev set (Zhao et al., 2019). VoxCeleb-1 speech corpus contains recordings from 1,251 speakers, out of which 799 and 215 speakers belong to the USA and UK, respectively, where English is a dominant language. It consists of utterances from 1,211 speakers in the dev set and 40 speakers in the test set. The VoxCeleb-2 dataset consists of 5,994 speakers in the dev set. We used the same dev-test split as given in (Nagrani et al., 2017; Chung et al., 2018). The VoxCeleb datasets contain mostly English speech recordings (Chen et al., 2021). Details about the training setup are present in the Appendix section.

Test datasets: Experiments were conducted using trial pairs from the following publicly available

[4] VoxCeleb-1: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox1.html

[5] VoxCeleb-2: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html

datasets: VoxCeleb-1[6] test set (mostly English) (Nagrani et al., 2017), LibriSpeech[7] test set (English) (Panayotov et al., 2015), Aishell-1 test set[8] (Non-English) (Bu et al., 2017), and Voxforge[9] test set (Non-English) (Voxforge.org). Aishell-1 is a Mandarin speech corpus. Voxforge test set contains speech recordings in 10 different languages, namely, Bulgarian, Dutch, French, German, Greek, Italian, Portuguese, Russian, Spanish, and Turkish. We randomly generated the trial pairs for LibriSpeech, Aishell-1, and Voxforge from these publicly available datasets. The VoxCeleb-1 test set, LibriSpeech test set, Aishell-1 test set, and the Voxforge test set contain 37720, 47402, 23800, and 51856 trial pairs, respectively. The majority of the publicly available speech datasets are in the English language. VoxCeleb (used to train the models) datasets contain speech recordings in mostly English. Therefore we primarily investigate the effectiveness of this work on Non-English test sets without domain adaptation.

4.2 Baselines

We performed experiments using two publicly available baselines: RawNet-2[10] (Jung et al., 2020) and VGG-M[11] (Nagrani et al., 2017).

RawNet-2: RawNet-2 is an improved version of RawNet (Jung et al., 2019). It takes raw waveform as input and extracts speaker embedding. The model is pre-trained on VoxCeleb-2 for the speaker identification task to obtain 1024-dimensional embedding (Jung et al., 2020). Speech recordings from the trial pair are fed to the model individually as inputs. Thus, we get two 1024-dimensional embeddings for each input audio in the trial pair. We compute a cosine-similarity score of these two embeddings for the speaker verification task.

VGG-M: The VGG-M model was trained on the entire VoxCeleb-1 dataset for speaker identification (Nagrani et al., 2017). It generates a 4096-dimensional discriminative embedding. We used this pre-trained model to construct a siamese network for speaker verification. We fine-tuned the siamese network on VoxCeleb-1 dev for speaker verification. Speech recordings from a trial pair are inputs to the VGG-M models (frozen weights). We concatenate the generated 4096-dimensional embeddings to get a single 8,192-dimensional embedding. After batch normalization, we pass this embedding through a fully connected layer (consisting of 512-units) and apply the ReLU activation function. Then, after performing L_2-normalization, this 512-dimensional embedding (VGG-embedding) is fed to another fully connected layer consisting of two units for speaker verification.

4.3 Input strategy

For simplicity, we feed the input features to the model without any test time augmentation. The adaptive average pooling layer of the FAtNet model handles the duration variability issue. FAtNet is not a siamese network, and the weights of both the TDNN paths are learned separately. We pass the features from each input audio clip in the trial pair through both the TDNN paths. We further compute a mean as shown below: Assuming $mfcc_1$ and $mfcc_2$ are the MFCC features obtained for clips in the trial pair.

FAtNet-v1:

1. prob_1 := model(mfcc_1, mfcc_2)
2. prob_2 := model(mfcc_2, mfcc_1)
3. $\text{prob}_{\text{final}}$:= mean(prob_1, prob_2)

FAtNet-v2:

1. $\text{emb}_{1\text{a}}$,$\text{emb}_{2\text{a}}$:= model(mfcc_1, mfcc_2)
2. $\text{emb}_{2\text{b}}$,$\text{emb}_{1\text{b}}$:= model(mfcc_2, mfcc_1)
3. emb_1 := mean($\text{emb}_{1\text{a}}$, $\text{emb}_{1\text{b}}$)
4. emb_2 := mean($\text{emb}_{2\text{a}}$, $\text{emb}_{2\text{b}}$)
5. $\text{prob}_{\text{final}}$:= CosineSimilarity(emb_1, emb_2)

4.4 Evaluation Metric

Equal Error Rate (EER) is a standard evaluation metric for biometric systems (Hansen and Hasan, 2015). Therefore, we investigate the effectiveness of this work in terms of EER. A lower EER score indicates a better performance.

5 Experiments and Results

5.1 Experimental validation of hypothesis

Our proposed approach aims to reduce the linguistic bias in the existing speaker verification systems by integrating a language-agnostic embedding. To

[6] VoxCeleb-1 : https://www.robots.ox.ac.uk/~vgg/data/voxceleb/meta/veri_test.txt

[7] LibriSpeech: https://www.openslr.org/12

[8] Aishell-1: https://www.openslr.org/33/

[9] Voxforge: http://www.voxforge.org/

[10] We used the pre-trained RawNet-2 model available in https://github.com/Jungjee/RawNet

[11] We used the pre-trained VGG-M model available in https://github.com/Derpimort/VGGVox-PyTorch

scientifically validate the theoretical hypothesis, we explore the effect of integrating FAtNet embedding with baselines for out-of-domain test sets.

Consider a trial pair $(clip1.wav, clip2.wav)$ having MFCCs, say, $(mfcc_1, mfcc_2)$ and spectrograms, say, $(spec_1, spec_2)$.

VGG-M $\oplus$ FAtNet-v1: In this integration, we directly pass MFCC's for the trial pair, say $(mfcc_1, mfcc_2)$, through the FAtNet-v1 model to get 1024-dimensional FAtNet embedding. We also pass spectrograms, say $(spec1, spec2)$, through the VGG-M siamese baseline to get the 512-dimensional VGG-embedding. After concatenating these embeddings, we pass it to a fully connected layer of 1024 neurons. Finally, after applying $ReLU$ and L_2-normalization, we pass it through another fully connected layer consisting of 2 units for speaker verification. We fine-tuned the last two fully connected layers on the VoxCeleb training set for speaker verification.

RawNet-2 $\oplus$ FAtNet-v2: For this integration, we obtained the 512-dimensional FAtNet embeddings for speech recordings in trial pair by feeding $(mfcc_1, mfcc_2)$ through the FAtNet-v2 model. We get these FAtNet embeddings from steps 3 and 4 of the input strategy (described in section 4.3) for FAtNet-v2. We obtained the 1024-dimensional RawNet embeddings for recordings by feeding them to the RawNet-2 baseline. We compute cosine similarity after concatenating audio1's FAtNet-v2 embedding with its RawNet-2 embedding and audio2's FAtNet-v2 embedding with its RawNet-2 embedding.

As illustrated in Table 1, we observe significant improvements in the performance of baselines on out-of-domain test sets after integration with FAtNet embeddings. This observation suggests that with a very little overhead, our proposed FAtNet embeddings may help improve the performance of these baselines on out-of-domain test sets without domain adaptation.

5.2 Language-specific analysis

For an extensive validation of the observations from the previous experiment, we created separate test sets for 11 different languages using the Voxforge dataset. The Bulgarian test set consists of 3,110 trial pairs. The other test sets contain 20,000 trial pairs each. Figure 3 and Figure 4 show that integrating the baselines with our proposed FAtNet embedding consistently reduced the equal error rate. This observation further verifies that the FAtNet embedding can help in reducing the language dependency of baselines and increase their generalizability on out-of-domain test sets. We observe an absolute improvement of 2.64% on the Dutch (Non-English) test set after integrating FAtNet-v1 with VGG-M. We observe an absolute improvement of 3.65% on the Portuguese (Non-English) set after integrating RawNet-2 with FAtNet-v2. Thus, the highest absolute improvements observed in the language-specific test sets were in Non-English test sets (Dutch and Portuguese).

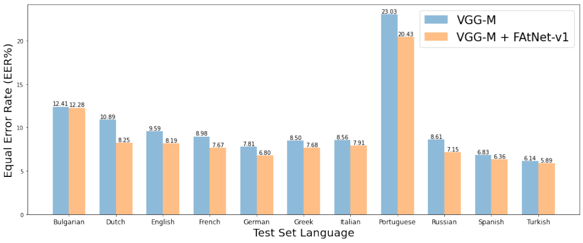

Figure 3: Figure showing that integrating VGG-M with FAtNet-v1 consistently reduced the EER on test sets generated for speech in different languages.

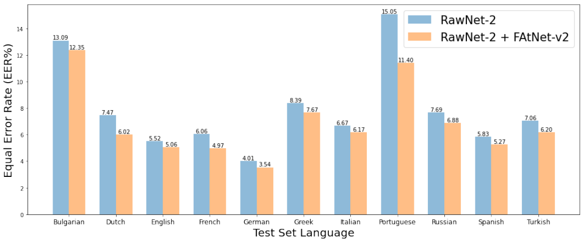

Figure 4: Figure showing that integrating RawNet-2 with FAtNet-v2 consistently reduced the EER on test sets generated for speech in different languages.

5.3 Linguistic study with augmentation

To get more linguistic insights, we study the performance of standalone FAtNet models using test-time augmentation (TTA) to feed input data to the models. We call the input strategy described in section 4.3 as S_0. In TTA Strategy, each input audio recording in the test set is either repeated several times or clipped to make its duration equal to 30-second (Nagrani et al., 2020). We further clip them into 3-second chunks. We create a batch of all possible pairs. So, we get a batch of 100 pairs. Finally, we feed the entire batch to the model (as shown for $\mathbf{S_0}$). For FAtNet-v1, we average out the probabilities in the end. For FAtNet-v2, we compute the average of 100 sets of obtained embedding$_1$'s to get the

Model	FAtNet Train set	Test Set	EER (%)	Rel. Imp. (%)
VGG-M	-	Voxforge	9.190	-
VGG-M $\oplus$ **FAtNet-v1**	VoxCeleb-1 dev	Voxforge	7.665	**+16.594%**
VGG-M $\oplus$ **FAtNet-v1**	VoxCeleb-2 dev	Voxforge	7.618	**+17.106%**
VGG-M	-	Aishell-1	9.999	-
VGG-M $\oplus$ **FAtNet-v1**	VoxCeleb-1 dev	Aishell-1	9.139	**+8.601%**
VGG-M $\oplus$ **FAtNet-v1**	VoxCeleb-2 dev	Aishell-1	6.866	**+31.333%**
RawNet-2	-	Voxforge	7.012	-
RawNet-2 $\oplus$ **FAtNet-v2**	VoxCeleb-2 dev	Voxforge	5.341	**+23.831%**
RawNet-2	-	Aishell-1	6.202	-
RawNet-2 $\oplus$ **FAtNet-v2**	VoxCeleb-2 dev	Aishell-1	3.832	**+38.213%**

Table 1: Table showing the relative improvements in the performance of VGG-M and RawNet-2 baselines after integration with FAtNet embeddings.

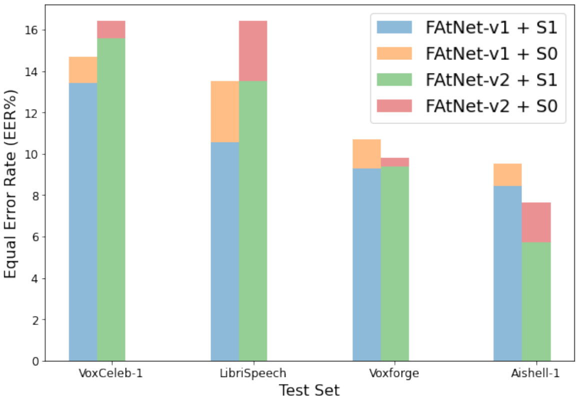

Figure 5: Figure showing the stability of the proposed FAtNet models on out-of-domain test sets. We achieved a better performance using the S_1 test-time augmentation strategy as compared to S_0.

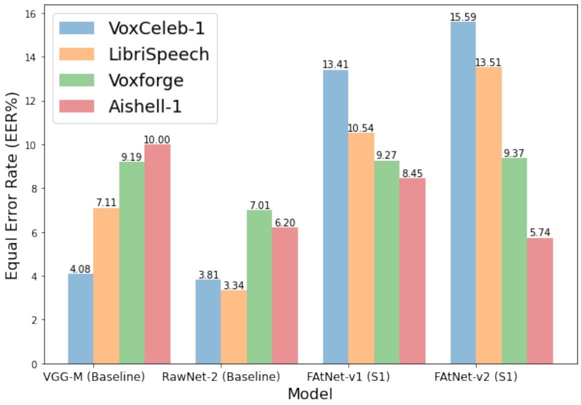

Figure 6: Figure showing that baselines' performance degraded on Non-English test sets without domain adaptation. On the contrary, the performance of FAtNet models improved on those sets without adaptation.

final tensor for embedding$_1$. We do the same for embedding$_2$. Then, we compute the cosine similarity between final embedding$_1$ and embedding$_2$. We call this input strategy S_1.

Observations: Figure 5 shows that we achieved a better performance using S_1 as compared to S_0. Interestingly, FAtNet-v2 achieved better performance on out-of-domain (Non-English) test sets than FAtNet-v1. It is reasonable as FAtNet-v2 contains two 4-head attention blocks, whereas FAtNet-v1 contains only a single 8-head attention block. It indicates that an intelligent selection of frame-level features individually from each audio clip enhances the language robustness of the model for out-of-domain sets.

5.4 Qualitative comparison with the baselines

In this experiment, we compare the generalization capabilities of our proposed networks with the baselines. We observe from Figure 6 that the baseline models performed relatively well on the VoxCeleb-1 test set (mostly English) and LibriSpeech (English) test set. However, the performance of baselines degraded on the other two out-of-domain multilingual test sets, namely, Aishell-1 (Mandarin) and Voxforge (Non-English).

On the contrary, we observe that the performance of FAtNet models improved for out-of-domain multilingual test sets. FAtNet models generalized well to out-of-domain test sets without domain adaptation. It illustrates the language dependency in VGG-M and RawNet-2 baselines. The relatively poor performance of FAtNet models on VoxCeleb-1 and LibriSpeech test sets may be due to other variability issues in these datasets. VoxCeleb-1 is a dataset collected in noisy unconstrained conditions. LibriSpeech corpus (derived from audiobooks) consists of high prosodic variations. It suggests that frame-attentive networks are generalizable but not robust to noise or prosodic variations.

5.5 Ablation Study

Intending to investigate which component of FAtNet models makes them language robust, we did an ablation study. FAtNet models contain two parts: the TDNN module and the attention block. We trained a simple TDNN-based model on the VoxCeleb-2 dev set. The architecture of this model remains the same as FAtNet models, except that this TDNN-based model does not contain the attention block. It is observed from Figure 7 and Figure 8 that our proposed FAtNet models outperformed the TDNN model on most of the test sets. The plots suggest that frame-level features make the model language robust. An intelligent selection of features using attention can help in enhancing performance. Thus, our proposed FAtNet models consisting of this TDNN model and attention blocks are language robust.

Interestingly, the TDNN model outperformed FAtNet-v1 and FAtNet-v2 on Aishell-1 and LibriSpeech. FAtNet-v1 outperformed TDNN on LibriSpeech, but FAtNet-v2 could not show that FAtNet-v1 can better handle prosodic variations than FAtNet-v2. It is reasonable as FAtNet-v1 contains a single attention block that selects intelligent features after concatenating the frame-level features of both the audio clips. Thus, the number of dimensions of each frame sent to the FAtNet-v1 attention block is higher than FAtNet-v2. However, FAtNet-v2 performed better on out-of-domain test sets than TDNN, whereas FAtNet-v1 performed poorly on the Aishell-1 test set compared to TDNN. The presence of two separate attention blocks (specific for each audio clip) in FAtNet-v2 makes it more language robust than FAtNet-v1.

6 Discussion

This work investigates the cost-effective idea of integrating the lightweight frame-attentive embedding with heavier and stronger baselines to mitigate the linguistic bias in such baselines without adaptation. After comprehensive experimentation on twelve languages and ablation studies, we observed that the proposed method showed significant and consistent improvements in reducing the linguistic bias in the baselines. Some final considerations:

Model complexity: Table 2 illustrates that the FAtNet models consist of fewer parameters as compared to the baselines. Thus, the proposed frame-attentive networks take less time to train. The VGG-M and the RawNet-2 occupy 71.7MB and 53.6MB of disk space. Our proposed FAtNet-v1 and FAtNet-v2 occupy 41MB and 32.6MB of space. Thus, FAtNet models are lighter than the baselines.

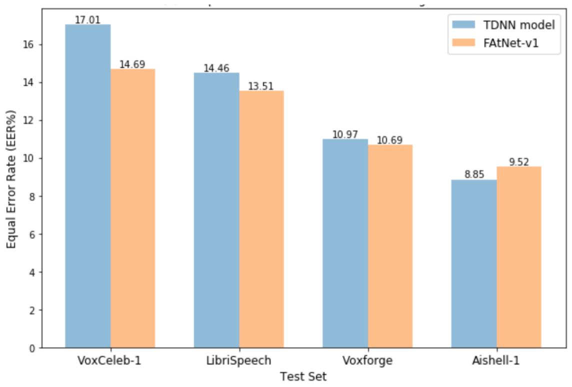

Figure 7: Comparing the performance of TDNN model with FAtNet-v1. We used the S0 strategy (FAtNet-v1 version) for both these models.

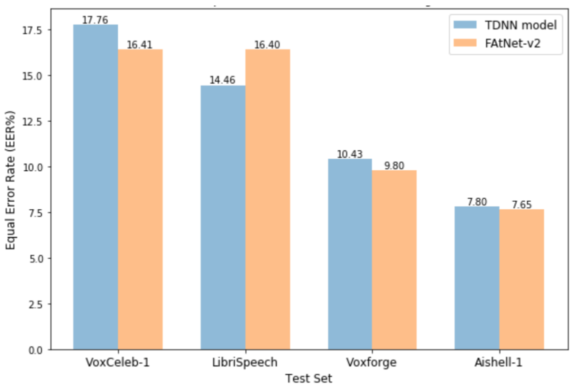

Figure 8: Comparing the performance of TDNN model with FAtNet-v2. We used the S0 strategy (FAtNet-v2 version) for both these models.

Model	#Parameters
VGG-M	17909219
RawNet-2	13379378
FAtNet-v1	10226690
FAtNet-v2	8127490

Table 2: Table showing details about the number of parameters in the proposed networks and the baselines.

Cost-effectiveness: Mitigating the linguistic bias without adaptation is crucial for enhancing the usability of the model across the globe. However, it is also an extremely challenging problem that requires complex decision-making. Consequently, it requires more parameters in the network and some additional overhead. Below we discuss the overhead of some of the popular methods to mitigate the problem of linguistic bias:

1. Fine-tuning the pre-trained model individually for each language could help mitigate this issue. However, considering that there are approximately 7,000 languages in the world (Huang et al., 2021), it is costly to adapt the models individually for each existing language.

2. Training heavy and highly complex models on extensive datasets could help enhance the model's generalizability to out-of-domain testing. However, this approach requires significant computational overhead and storage space.

FAtNet models have lower complexity than the baseline models. We purposefully integrated our proposed lightweight embeddings with heavier and stronger baselines. We observed significant improvements after integrating the lightweight FAtNet embeddings with the baselines. Hence, with a very little overhead, FAtNet embeddings may help enhance the generalizability of baselines. Therefore, as compared to the overhead of the above two approaches, our proposed approach cost-effectively enhances the usability of baseline models across the globe.

7 Conclusions and Future Work

In this paper, we introduced the cost-effective idea of utilizing a lightweight frame-level embedding for reducing linguistic bias in existing speaker verification systems without the overhead of domain adaptation. We also explored applying attention to individual frames to focus on relevant frame-level discriminative information. For an in-depth analysis of our proposed theoretical hypothesis, we proposed two variants of frame-attentive networks: FAtNet-v1 and FAtNet-v2. We investigated the effect of their integration with the baselines for twelve languages. Empirical results showed consistent improvements in the performance of baselines on out-of-domain test sets without domain adaptation after integration with the FAtNet embedding. Qualitative comparison with baselines suggested that the proposed models are comparatively more generalizable. Additionally, we did an ablation study of our proposed networks. It turns out that frame-level embedding captures less linguistic information from speech than utterance-level embedding. An intelligent selection of features from frames can further improve the performance of speaker verification models.

Our analysis points to vital problems for future work. For instance, it may be worthwhile to explore standalone domain-invariant architectures. This work focuses on the scenario where the input speech recordings within a trial pair are in the same language. However, it may be helpful to explore this idea further for bilingual speakers where the input speech recordings in the trial pair are in different languages.

References

R. Auckenthaler, M. J. Carey, and J. S. D. Mason. 2001. Language dependency in text-independent speaker verification. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 441–444 vol.1.

Yonatan Belinkov, Ahmed Ali, and James Glass. 2019. Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 81–85.

Vincent Brignatz, Jarod Duret, Driss Matrouf, and Mickael Rouvier. 2021. Language adaptation for speaker recognition systems using contrastive learning. In *Speech and Computer*, pages 91–99, Cham. Springer International Publishing.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline.

Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2020. Adversarial Domain Adaptation for Speaker Verification Using Partially Shared Network. In *Proc. Interspeech*, pages 3017–3021.

Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2021. Self-supervised learning based domain adaptation for robust speaker verification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5834–5838.

Roza Chojnacka, Jason Pelecanos, Quan Wang, and Ignacio Lopez-Moreno. 2021. Speakerstew: Scaling to many languages with a triaged multilingual text-dependent and text-independent speaker verification system. *CoRR*, abs/2104.02125.

Grzegorz Chrupała, Bertrand Higy, and Afra Alishahi. 2020. Analyzing analytical methods: The case of phonology in neural models of spoken language. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4146–4156, Online. Association for Computational Linguistics.

J. S. Chung, A. Nagrani, and A. Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.

Zhifu Gao, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai. 2018. An improved deep embedding learning method for short duration speaker verification. In *Proc. Interspeech 2018*, pages 3578–3582.

Peter Guzewich, Stephen Zahorian, Xiao Chen, and Hao Zhang. 2018. Cross-corpora convolutional deep neural network dereverberation preprocessing for speaker verification and speech enhancement. In *Proc. Interspeech 2018*, pages 1329–1333.

J. H. L. Hansen and T. Hasan. 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99.

William N. Havard, Jean-Pierre Chevrot, and Laurent Besacier. 2019. Word recognition, competition, and activation in a model of visually grounded speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 339–348, Hong Kong, China. Association for Computational Linguistics.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, and Alexander Hauptmann. 2021. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2443–2459, Online. Association for Computational Linguistics.

Jee-Weon Jung, Hee-Soo Heo, Ju-Ho Kim, Hye-Jin Shim, and Ha-Jin Yu. 2019. Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. pages 1268–1272.

Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu. 2020. Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. *Proc. Interspeech 2020*, pages 3583–3587.

Elie Khoury, Laurent Shafey, and Sébastien Marcel. 2014. Spear: An open source toolbox for speaker recognition based on bob. pages 1655–1659.

Na Li, Deyi Tuo, Dan Su, Zhifeng Li, and Dong Yu. 2018. Deep discriminative embeddings for duration robust speaker verification. In *Proc. Interspeech 2018*, pages 2262–2266.

J. Lyons. Mel frequency cepstral coefficient (mfcc) tutorial. http://practicalcryptography.com /miscellaneous/ machine-learning/guide-mel-frequency-cepstralcoefficients-mfccs/.

Ayoub Malek. 2020. Signal framing. https://superkogito.github.io/blog/ SignalFraming.html.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python.

Ion Moshnoi. All you need is attention — computer vision edition. https://medium.com/@moshnoi2000/all-you-need-is-attention-computer-vision-edition-dbe7538330a4. 04/19/2018.

A. Nagrani, J. S. Chung, and A. Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. 2018. Attentive statistics pooling for deep speaker embedding. In *Proc. Interspeech 2018*, pages 2252–2256.

Kudashev Oleg, Sergey Novoselov, Timur Pekhovsky, Konstantin Simonchik, and Galina Lavrentyeva. 2016. Usage of dnn in speaker recognition: Advantages and problems. In *Advances in Neural Networks – ISNN 2016*, volume 9719, pages 82–91.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Johan Rohdin, Themos Stafylakis, Anna Silnova, Hossein Zeinali, Lukas Burget, and Oldrich Plchot. 2019. Speaker verification using end-to-end adversarial language adaptation. In *ICASSP 2019*, pages 6006–6010.

Cory Shain and Micha Elsner. 2020. Acquiring language from speech by learning to remember and predict. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 195–214, Online. Association for Computational Linguistics.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. pages 5329–5333.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Sanjeev Khudanpur Vijayaditya Peddinti, Daniel Povey. 2015. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, pages 3214–3218.

Voxforge.org. Free and open source speech recognition engines (on linux, windows and mac). http://www.voxforge.org/. Accessed 06/25/2014.

Yanfeng Wu, Chenkai Guo, Hongcan Gao, Xiaolei Hou, and Jing Xu. 2020. Vector-Based Attentive Pooling for Text-Independent Speaker Verification. In *Proc. Interspeech 2020*, pages 936–940.

Wei Xia, Jing Huang, and John H. L. Hansen. 2019. Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 5816–5820. IEEE.

Ya-Qi Yu, Lei Fan, and Wu-Jun Li. 2019. Ensemble additive margin softmax for speaker verification. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6046–6050.

F. Zhao, H. Li, and X. Zhang. 2019. A robust text-independent speaker verification method based on speech separation and deep speaker. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6101–6105.

Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey. 2018. Self-attentive speaker embeddings for text-independent speaker verification. In *Proc. Interspeech 2018*, pages 3573–3577.

A Training Setup

We cropped the silent parts of speech recordings and clipped them into chunks of 3-seconds. We generated MFCCs of 80-dimensions as acoustic inputs to the model using the Librosa library (McFee et al., 2015). We randomly generated separate training sets consisting of 5,25,000 and 23,88,000 trial pairs for models trained on VoxCeleb-1 dev and VoxCeleb-2 dev sets, respectively. The training set consists of an equal number of positives and negatives. We shuffled the training examples before each epoch. The batch size was 128. We trained the proposed models under the joint supervision of softmax loss and center loss. Center loss helps in reducing intra-speaker variations, whereas the softmax loss helps in increasing the inter-speaker variations.(Li et al., 2018). We use Adam and RMSProp for FAtNet and center loss, respectively (Li et al., 2018; Paszke et al., 2017). For training on VoxCeleb-1 dev, the learning rate was 0.005 and 0.2 for Adam (for FAtNet) and RMSprop (for center loss), respectively. The learning rates were reduced after every ten epochs using $step\_lr$ scheduler with the gamma value of 0.5 and 0.3 for Adam and RMSprop, respectively. For training on VoxCeleb-2 dev, the learning rate was as low as 0.0005 for both Adam (for FAtNet) and RMSprop (for center loss), respectively. It is due to more steps being performed in one epoch while training on VoxCeleb-2 dev. We computed the total loss as follows: $total\_loss := softmax\_loss + 0.01 * center\_loss$ (Li et al., 2018). We used GeForce GTX 1080 GPU.

B Effect of training set size on model performance

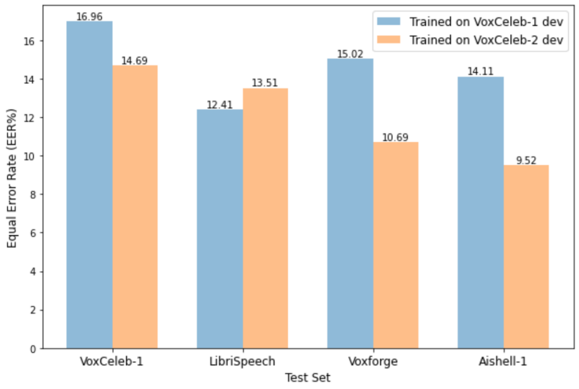

Figure 9: Figure showing the effect of training set size on the performance of FAtNet-v1. Observation: The FAtNet-v1 model trained on the VoxCeleb-2 dev set performs better than the model trained on the VoxCeleb-1 dev set for most of the test sets. It indicates that increasing the size of the training set can help in improving the performance of speaker verification models.

In this experiment, we study the effect of training set size on the model performance. It can help visualize the improvements in increasing the number of speakers in the training set. For this experiment, we trained separate models of FAtNet-v1 on VoxCeleb-1 dev and VoxCeleb-2 dev respectively. As shown in Figure 9, the model trained on the VoxCeleb-2 dev set outperformed that trained on the VoxCeleb-1 dev set for most of the test set. It is reasonable as the VoxCeleb-2 dev set is more diverse and multilingual than the VoxCeleb-1 dev set. For the LibriSpeech test set, the model trained on the VoxCeleb-1 dev set performed slightly better than that trained on the VoxCeleb-2 dev set. It is reasonable due to the high proportion of English speech recordings in VoxCeleb-1 and LibriSpeech being an English speech corpus. Thus, we used the FAtNet models trained on the VoxCeleb-2 dev set for experiments (unless explicitly specified otherwise).

C Choice of Input Features

We compared the performance of 80-dimensional MFCCs with 300-dimensional spectrogram (Nagrani et al., 2017) as inputs to FAtNet-v1. As illustrated in Table 3, we achieved better performance using the MFCC features.

Test Set	MFCC	Spectrogram
VoxCeleb-1	14.682	20.668
LibriSpeech	13.527	20.919
Voxforge	10.694	20.274
Aishell-1	9.521	21.143

Table 3: Table showing the EER(%) on training FAtNet-v1 using MFCC and Spectrogram features, respectively.

D Hyper-parameter details for the stacked TDNN

Table 4 illustrates the hyper-parameter detail for the four stacked TDNN layers in the proposed frame-attentive networks. We set the dropout and stride to 0.1 and 1, respectively, for all four layers.

Hyper-parameter	#1	#2	#3	#4
Input dimension	80	512	512	512
Output dimension	512	512	512	512
Context-size	3	5	3	1
Batch-norm	False	False	True	True

Table 4: Hyper-parameter detail for the stacked TDNN layers in FAtNet models.