# A Lightweight End-to-End Multi-task Learning System for Vietnamese Speaker Verification

*Anonymous submission to INTERSPEECH 2023*

## Abstract

Automatic speaker verification (ASV) in low-capacity devices utilized for industrial Internet of Things (IoT) applications is faced with two major challenges: lack of annotated training data and model complexity. To address these challenges, this paper introduces the first Vietnamese audio dataset for training a multi-task learning method named Vi-LMM that jointly per forms command detection, fake voice recognition, and speaker verification tasks. To optimize Vi-LMM for low-capacity de vices, we further employ knowledge distillation to reduce the number of parameters by 3.5 times. An empirical experiment is conducted to evaluate the effectiveness of the proposed method and the results show that Vi-LMM outperforms strong single task models in terms of both reducing the number of learnable parameters and achieving higher $F_1$ scores while maintaining comparable error rates.

**Index Terms**: speaker verification, multi-task learning, Viet namese

## 1. Introduction

Automatic speaker verification (ASV) is one of the most im portant fields of speech processing and is a fundamental com ponent of speech-based security systems [1, 2]. However, ASV for low-capacity devices, especially for Vietnamese, often faces two major challenges: the lack of dataset and model complexity.

ASVspoof 2019 [3] and ASVspoof 2021 [4] are two widely used datasets in English for anti-spoofing in automatic speaker verification, while FMFCC-A [5] is the largest publicly avail able dataset for synthetic speech detection in Mandarin. Vox Celeb [6] is an audio dataset consisting of more than 100,000 utterances from 1,251 celebrities, which is suitable for speaker verification tasks. However, for Vietnamese speaker verifica tion tasks, only two relatively small public datasets [7, 8] exist, but they are not publicly available to the research community. The VLSP2021 Challenge[1]provides a Vietnamese Speaker Ver ification dataset, but it is also not publicly available. This lack of high-quality annotated datasets in Vietnamese poses a sig nificant challenge for studies on ASV in the language. In this study, we aim to address this issue by developing a meticulously designed Vietnamese dataset suitable for novel AI model-based tasks that are currently receiving widespread attention from the research community worldwide.

Recent deep learning models for ASV have a significant number of parameters and require substantial computational re sources to perform accurately [9, 10, 11, 12, 13, 14]. Deep neural networks have been widely used for ASV, either as stan dalone models [15, 16, 8, 17, 18] or as feature extractors for other classifiers [19]. Moreover, previous studies on ASV tend to explore large pretrained speech presentation models [20, 21, 22]. Though these models achieve outstanding per formances, they have a massive size and long inference time. Thus, using such models in low-capacity IoT devices is chal lenging. Furthermore, current ASV systems focus on specific tasks, while real-world applications require the ability to han dle multiple tasks simultaneously. Using single model for each task significantly increases the operations that the hardware sys tem needs to perform, which thereby leads to certain latency and degrades the user experience. As a result, developing com pact and fast multi-task learning models that can be embedded in low-capacity devices is crucial for ASV applications. De spite previous efforts, no dataset or prior models address multi task learning for Vietnamese speaker verification. Additionally, most existing deep learning models are either exclusive to a sin gle task or have an excessive number of parameters, while our goal is to build a comprehensive and lightweight system suit able for low-capacity devices. In this study, we aim to address these challenges by developing multi-task learning models that are both compact and fast for Vietnamese speaker verification.

For all the aforementioned motivations, we push this re search field forward by introducing a new dataset and a lightweight model. Our dataset includes 6480 audio and text label pairs from 162 individuals and 65 types of AI synthetic voice for three key ASV tasks. Our Vi-LMM model is a lightweight multitasking model that incorporates an attention layer to integrate information between tasks. We have reduced the number of parameters of Vi-LMM by 3.5 times using recent advances in knowledge distillation. Our contributions to this field are summarized as follows:

- We introduce the first public Vietnamese dataset as training data of three sub-tasks of ASV, namely command detection, fake voice recognition, and speaker verification;
- We propose two lightweight models, termed Vi-LMM and Vi-LMM-S, for joint learning of the three tasks;
- Experimental results on our dataset show that (i) while re quiring a significantly smaller number of parameters, our proposed models exhibit comparable performance to other strong single baselines [23, 14, 8, 18, 12, 17] and (ii) our joint learning method improves the overall performance of the model on the three sub-tasks.

We publicly release our dataset and model implementations for research or educational purpose.[2]

[1]https:/vlsp.org.vn/vlsp2021/eval/vsv

We hope that our dataset and model can serve as a starting point for future Vietnamese speech processing research and applications.

[2]Our dataset and model will be released upon acceptance

| Type | Example |
|---|---|
| A | **Mở camera lên.**<br>**Turn on the camera.** |
| B | Anh ấy đến rồi, **mở camera lên.**<br>He has already come, **turn on the camera.** |
| C | Đến cảnh của cô ấy rồi, chuẩn bị đạo cụ cho tôi.<br>It's her scene, prepare the clothes for me. |
| D | **Camera** của điện thoại này tệ quá.<br>The **camera** of this phone is so bad. |

Table 1: *Examples of four speech categories in Vietnamese and their English translated version*

## 2. Our dataset

### 2.1. Multi-tasking dataset

Our goal is to develop a comprehensive dataset that can be used to train a multi-task learning model capable of performing three tasks: command detection, fake voice recognition, and speaker verification. The model should be able to differentiate between authentic user speech and different types of distractors, includ ing synthetic AI speeches, non-command speeches with simi lar patterns, and noises from other speakers. To achieve this, we select two widely-used commands in IoT applications, "Mở camera lên" (Turn on the camera) and "Đóng cửa lại" (Close the door), and define four categories of speech: **A**) exact com mand, **B**) conversational speech containing one command, **C**) speech with no command, and **D**) speech with similar words to the command but should not be identified as a correct one. Examples for each category are presented in the Table 1. The annotated dataset is divided into training, validation, and test sets in a 5/2/3 ratio, ensuring that the distribution of order utter ance types and gender is well-balanced across all subsets. The dataset statistics are displayed in Table 2.

### 2.2. Construction process

**Guideline construction.** We initially create a data collection protocol and record a small sample of audio recordings accord ing to it. After that, the team reviews the process and identifies any issues, leading to the development of the final data collec tion guideline that is used for the remaining dataset.

**Data collection.** In this stage, 170 Vietnamese participants be tween 18 and 25 years of age are recorded and labeled. Next, the subjects choose one out of two commands and prepare 20 tran script sentences, as described in Section 2.1. They prepare five transcripts for each of the four categories. The transcripts are re viewed and audited by the engineers before the subject records the 20 corresponding audios. The subjects are instructed to make sure that both the transcripts and the audios sound nat ural, convey meaningful sentences, and are common in real-life scenarios.

**AI synthetic speech generation.** After the data collection phase, we employ HiFiGAN [24] to perform an automatic speech generation task.The process finally obtains 3240 syn thetic data samples from 65 different AI voices.

**Revision.** We conduct a manual quality check of each audio and its label file to ensure consistency and remove samples that did not meet the criteria. For the verification task, we label each performer and their corresponding audio accordingly. For command detection, audios belonging to groups A and B are la beled as True, while others are labeled as False. Additionally, all speeches generated by HiFiGAN are labeled as AI synthesized

| Statistic | Total | A | B | C | D |
|---|---|---|---|---|---|
| | 810 | 810 | 810 | 810 | 810 |
| | 162 | 162 | 162 | 162 | 162 |
| | 1.85 | 0.75 | 1.88 | 1.54 | 0.96 |
| | | | | 6<br>5 | |
| | 3.43 | 2.38 | 4.43 | 4.00 | 2.79 |
| | 7 | 3 | 7 | 6 | 3 |
| | 37 | 3 | 37 | 29 | 23 |
| | 5.13 | 3.91 | | 11.64<br>11.29<br>6.02 | |
| | 810 | 810 | 810 | 810 | 810 |
| | 1382 | | 1382 | | 955245 |

| | | | |
|---|---|---|---|
| | | | 72<br>28<br>96<br>4 |

# audios 3240 # subjects 162 Minimum length (s) 0.75 Maximum length (s) 10.36 Average length (s) 3.38
Minimum length (t) 3 Maximum length (t) 37 Average length (t) 8.22 # AI synthetic audios 3240 Total duration (s)

Table 2: *Statistics of our Vietnamese dataset. (s) represents for "seconds" and (t) represents for "tokens".*

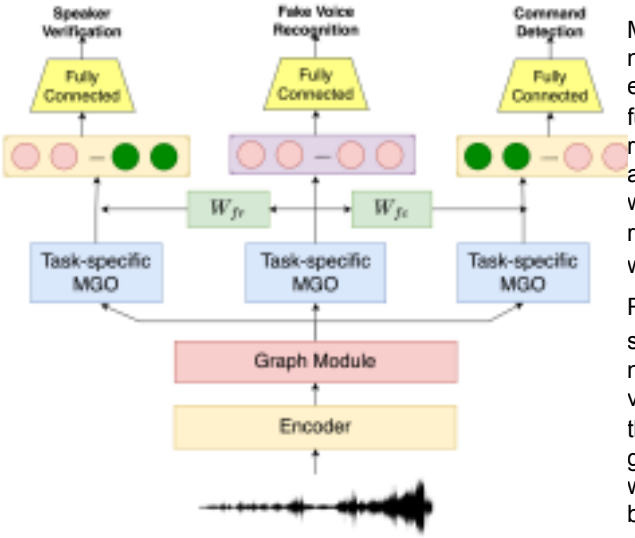


Figure 1: *The schematic overview of our Vi-LMM method*

speech for the fake voice recognition task. The final Vietnamese dataset contains 6480 audios from 162 subjects and 65 different AI voices.

### 2.3. Discussion

We generate an artificial dataset using a limited set of two spe cific commands and scripted speech, which may not fully rep resent the diversity and complexity of commands found in real world scenarios. It is essential to acknowledge that there are no publicly available Vietnamese speech datasets for the three tasks of command detection, fake voice recognition, and speaker ver ification. In our study, we aimed to simulate real-world speech and plan to compare the accuracy of our scripted speech dataset with genuine real-world speech in future work.

## 3. The proposed method

In this section, we present our proposed end-to-end method, named Vi-LMM, that simultaneously solves the three aforemen tioned sub-tasks. Figure 1 illustrates the architecture of Vi LMM, which consists of five main components including an au dio representation module, a graph module, three task-specific Max Graph Operation branches, a cross-task attention layer, and a decoding layer.

### 3.1. Audio Representation

To operate directly upon raw waveform inputs, we utilize a neural network-based encoder to take raw waveforms as in puts and then directly extracts the high-level representation $F \in R^{C\times S\times T}$, where $C$ is the number of channels, $S$ is the number of spectral bins, and $T$ is the temporal sequence length. For each channel in F, we extract the spectral and the temporal feature vectors by getting the maximum values along with the temporal and the spectral axis, respectively.

### 3.2. Graph Module

Motivated by recent studies that utilized graph neural networks to achieve state-of-the-art (SOTA) performance even with com pact models [25, 26, 27], we construct two fully-connected bidi rectional graphs $G_s$ and $G_t$ representing for the connections among spectral features and connections among temporal fea tures, respectively, where the features are outputs of the audio representation step. Since $G_s$ and $G_t$ are fully-connected, we can represent both graphs as matrices, i.e., $G_* \in R^{N_*\times D_*}$, where the subscript * represents $s$ and $t$ for spectral and temporal net works, and $N_*$ and $D_*$ are the number of nodes and the dimen sionality of each node vector in the corresponding graph, re spectively. We note that $N_s = N_t = C$, $D_s = T$, and $D_t = S$. Next, we adopt a graph attention layer [26] to assign a learnable attention weight to each edge for representing the relationship between two nodes.

After calculating the attention weights between nodes, a simple yet effective attentive graph pooling layer is applied to the output of each graph attention layer to reduce the number of nodes in both $G_s$ and $G_t$, which would reduce the compu tational complexity and improve discrimination between nodes. Finally, the model employs a graph combination to obtain a het erogeneous attention network $G_{st} \in R^{N_{st}\times D^{st}}$, where $N_{st} = N_s + N_t$ that contains both spectral and temporal information by combining two separated graphs $G_s$ and $G_t$.

### 3.3. Task-specific Max Graph Operation (MGO)

We adopt three MGO branches, originally introduced in [12], for multiple speech processing tasks. Each MGO consists of two modules, namely HS-GAL and feature synthesis. Unlike [12], each of the task-specific MGO branches in our study has two sequential HS-GAL modules.

Taking the heterogeneous graph $G_{st}$ obtained from the pre vious step, the HS-GAL module assigns additional attention weights to the edges connected between spectral and temporal feature nodes to learn high-level feature vectors that combine both spectral and

temporal information for classification. We denote the feature vectors as $h_f$, $h_c$, and $h_r \in R^{3D_{st}}$, which are the representation of fake voice recognition, command de tection, and speaker verification tasks, respectively.

### 3.4. Cross-task Attention

Based on the intuition that a representation of AI synthesis voice cannot be accepted as either a genuine speaker or a correct command, we design an additional attention layer to explic itly feed the information from the fake voice detection to the two remaining tasks, including command detection and speaker verification. Specifically, the cross-task attention layer inputs $\{h_f, h_c, h_r\}$ and produces specific cross-task attention weights that illustrate the influence of one task on another. Formally, to incorporate the information from fake voice detection to speaker verification, the layer first creates a cross information

concentrated vector $x_{fr} \in R^{3D_{st}}$ by multiplying a weight ma trix $W_{fr}$ with the feature vector $h_f$. Next, we compute the at tention weight between the two tasks using an attention weight $\lambda_{fr}$ and concatenates the result vector to the original $h_r$ vector as follows:

$$h'_r = (\lambda_{fr} x_{fr}) \;//\; h_r.$$

The layer produces the $h'_c$ to integrate useful information from fake voice recognition to the command detection task in a sim ilar manner. The vectors $h'_r$, $h'_c$, and $\mathbf{h}_f$ are then passed into a fully-connected layer (FC) where the output of each FC is the predicted label for each task.

### 3.5. Joint Learning

All attention weights and learnable matrices in our joint learn ing model are trained via a joint loss function $L$, which is the weighted sum of three single-task losses as follows:

$$L = \alpha L_C + \beta L_F + (1 - \alpha - \beta) L_R, \quad (1)$$

where $L_C$, $L_F$, and $L_R$ are cross-entropy losses computed based on labels from command detection, fake voice recogni tion, and speaker verification, respectively. The loss coeffi cients $\alpha$ and $\beta$ are fined tune during training to figure out the optimal loss function.

### 3.6. Vi-LMM Variant

Taking the advantage of knowledge distillation [28], we aim to further reduce the size of Vi-LMM whilst retaining the model's performance. To this end, our Vi-LMM acts as the teacher model and the encoder of the teacher with a significantly more lightweight encoding layer to construct a student model, termed Vi-LMM-S. Next, we transfer the teacher's knowledge to the student model using knowledge distillation techniques [28]. The training objective loss for Vi-LMM-S is the weighted sum of the student loss $L_{ST}$ and the distillation loss $L_{DI}$ as follows:

$$L_{LW} = L_{ST} + L_{DI},$$

where $L_{ST}$ can be established similarly as (1) and $L_{DI}$ are the weighted sum of cross-entropy losses computed based on soft labels from teacher model and soft predictions from student models.

## 4. Experiment

We conduct experiments on our dataset to study a quantitative comparison between Vi-LMM, Vi-LMM-S, and recent strong methods in terms of performance, model's size, and inference time.

### 4.1. Competitive Schemes

We compare our proposed models to five strong baselines across various domains, including:

- **Rawnet2** [23] & **Rawnet3** [14]: end-to-end DNNs classifier for raw waveform speaker recognition.
- **GFCC-ResNet101** [8]: a recent deep model designed for Vietnamese speaker authentication problem.
- **FastAudio** [18]: an end-to-end framework for audio classifi cation problem.
- **AASIST** [12]: current state-of-the-art model on the ASVspoof 2019 LA dataset.
- **AutoSpeech** [17]: a derived CNN architectures for the task of speaker verification on the VoxCeleb1 dataset.

| | **ers Inference** | | **c. Fake Voice Ver.** | | | | **Avg-E A** |
|---|---|---|---|---|---|---|---|
| | | $F_1$ | **EE R** | $F_1$ | **EE R** | $F_1$ | |
| 40.14M | 135ms | 90.72 | 15. 27 | 79.26 | 19. 83 | 73.1 8 | 17.55 |
| 52.38M | 223ms | 91.82 | 4.5 7 | 90.51 | | 32 8 | 9.19 |
| 128.4M | 630ms | **95.81** | 8.3 6 | 87.31 | 15. 32 | 78.5 4 | 11.84 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 40.2M | 150ms | 91.32 | 5.26 | 89.45 | | 03<br>2 | 9.65 |
| 41.4M | 174ms | 92.19 | **4.06** | **90.72** | **13.65** | 79.12 | **8.86** |
| 54M | 267ms | 93.27 | 7.92 | 87.65 | | 764 | 11.84 |
| 14M | 64ms | 93.58 | 4.58 | 90.45 | 13.87 | **79.43** | 9.22 |
| **4M** | **46ms** | 91.82 | 5.86 | 88.97 | 16.52 | 77.63 | 11.19 |
| 14M | 60ms | 93.21 | 4.87 | 89.95 | 14.03 | 78.96 | 9.45 |

**Model**-$F_1$

**Rawnet2** 81.05 **Rawnet3** 87.14 **GFCC-ResNet101** 87.22 **FastAudio** 85.56 **AASIST** 87.34 **AutoSpeech** 86.39 **Vi-LMM 87.82 Vi-LMM-S** 86.14 **Vi-LMM-C** 87.37

Table 3: *Results on the test set. "Command Dec.", "Fake Voice Rec.", "Speaker Ver." denote command detection, fake voice recog nition, and speaker verification, respectively. "Avg-EER" and "Avg-$F_1$" denote Average $F_1$ and Average EER, respectively. Here, Vi-LMM-S is the compact variant of Vi-LMM and Vi-LMM-C is Vi-LMM* without *the cross-task attention layer.*

*Note that*, our approach does not include large pre-trained mod els as encoders e.g. wav2vec2.0 [29], HuBERT [30]. There fore, models that utilize pre-trained models are incomparable to our system. Besides, we also perform an ablation study by re moving the cross-task attention layer to create a model termed Vi-LMM-S i.e. the feature vectors are fed directly into the clas sifiers after passing through the task-specific MGOs.

### 4.2. Experimental Settings

For Vi-LMM, we use a Rawnet2-based encoder [31] to extract the high-level audio representations from raw waveform inputs. For the Vi-LMM-S, we replace the Rawnet2-based encoder with MobileNetV2 [32], a widely used network in applications for low-resource devices such as [33, 34], to construct the student model.

To optimize our model's hyper-parameters, we performed a grid search on the validation set with the Adam optimizer. The results showed that a learning rate and weight decay of $10^{-5}$ were best, along with $\alpha$ and $\beta$ values of 0.3 and 0.4, respec tively. Our model was trained for 100 epochs with these hyper parameters.

For evaluation metrics, we adopt the standard $F_1$-score for all three tasks and the equal error rate (EER) for two security related tasks, including fake voice recognition and speaker ver ification. All our reported results are the average output over five experiments with different random seeds.

### 4.3. Main Results

Table 3 reports the performances of the chosen baselines and our system. It is worth noting that each baseline is trained specifically for each task. Thus, in order to make a fair compar ison, the number of parameters of each single model presented in Table 3 is tripled compared that of the original study.

In general, our findings indicate that both Vi-LMM and Vi-LMM-S demonstrate competitive performance compared to other strong baselines, while enjoying a significantly lower time and space complexity. Notably, Vi-LMM outperforms all other methods with the highest Average-$F_1$-score of 87.82%. In terms of Average-EER, Vi-LMM is the third-best performer follow ing AASIT and Rawnet3. It is noteworthy that Vi-LMM only requires 14 million parameters, whereas Rawnet2, which is the second-smallest method, requires 40.14 million parameters.

Our system performs comparably well to other models in terms of individual task performance. For command detec tion, Vi-LMM achieves an $F_1$-score of 93.58%, close to that

of the highest-performing model GFCC-ResNet101, which has approximately nine times more parameters. For fake voice recognition, Vi-LMM's performance is comparable to that of AASIST, the highest-performing model, in terms of EER and $F_1$-score, despite having significantly fewer parameters. For speaker verification, AASIST has the best EER, but Vi-LMM achieves the highest $F_1$-score of 79.43%, showing the effective ness of information feeding from the fake voice detection task.

To reflect the speed advantage of Vi-LMM, we also report the inference time for each model. It should be noted that other models require three runs to obtain outputs for a single data sample, whereas our model only requires one. Our results in dicate that Vi-LMM has the fastest inference time, taking only 64ms, while GFCC-ResNet101 and AASIST take 630ms and 174ms, respectively.

**Variants of Vi-LMM.** Results from Table 3 indicate that Vi-LMM-S performs comparably to Vi-LMM but with signifi cantly fewer parameters, making it suitable for low-capacity de vices. Conversely, removing the cross-task attention layer (Vi LMM-C) results in reduced performance across all three tasks. Notably, voice command detection sees a 0.3% reduction, while the remaining two tasks experience a 0.5% reduction in $F_1$-score. These findings suggest that the cross-task attention layer plays a vital role in multi-task learning.

## 5. Conclusions

In this study, we introduced the initial public dataset for Viet namese speaker verification, which comprises three sub-tasks: command detection, fake voice recognition, and speaker ver ification. In addition, we proposed two simple yet effective models, Vi-LMM and Vi-LMM-S, for jointly learning the three tasks. Particularly, Vi-LMM extends AASIST by integrating three task-specific MGO branches and a cross-task attention layer, while Vi-LMM-S employs knowledge distillation tech niques and has only 4 million parameters. The experimental evaluation shows that both models surpass most of the strong methods in terms of Average-$F_1$ while using significantly fewer parameters. Furthermore, we verified that joint learning of the three sub-tasks via a cross-task attention layer is beneficial to enhance the performance of all the tasks. We hope that our dataset and model can serve as a starting point for future Viet namese speech processing research and applications.

## 6. References


[1] Z. Saquib, N. Salam, R. P. Nair, N. Pandey, and A. Joshi, "A survey on automatic speaker recognition systems," *Communications in Computer and Information Science*, pp. 134–145, 2010.

[2] D. A. van Leeuwen, "Speaker verification systems and security considerations," in *Proceedings of 8th European Conference on Speech Communication and Technology (Eurospeech)*, 2003, pp. 1661–1664.

[3] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. W. D. Evans, T. H. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (INTER SPEECH)*, 2019, pp. 1008–1012.

[4] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado, "ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *Proceedings of Edition of ASVspoof*, 2021, pp. 47–54.

[5] Z. Zhang, Y. Gu, X. Yi, and X. Zhao, "Fmfcc-a: A challenging mandarin dataset for synthetic speech detection," *arXiv preprint arXiv:2110.09441*, 2021.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large scale speaker identification dataset," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2616–2620.

[7] T. P. Van, N. T. N. Quang, and T. M. Thanh, "Deep Learning Approach for Singer Voice Classification of Vietnamese Popular Music," in *Proceedings of the International Symposium on Infor mation and Communication Technology*, 2019, pp. 255–260.

[8] S. T. Nguyen, V. D. Lai, Q. Dam-Ba, A. Nguyen-Xuan, and C. Pham, "Vietnamese speaker authentication using deep mod els," in *Proceedings of the International Symposium on Informa tion and Communication Technology*, 2018, pp. 177–184.

[9] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, 2021.

[10] P. Aravind, U. Nechiyil, N. Paramparambath *et al.*, "Audio spoof ing verification using deep convolutional neural networks by transfer learning," *arXiv preprint arXiv:2008.03464*, 2020.

[11] H. Tak, Jung, Jee-Weon, J. Patino, M. R. Kamble, M. Todisco, and N. W. D. Evans, "End-to-end spectro-temporal graph atten tion networks for speaker verification anti-spoofing and speech deepfake detection," *Proceedings of 2021 Edition of the Auto matic Speaker Verification and Spoofing Countermeasures Chal lenge*, 2021.

[12] J. Jung, H.-S. Heo, H. Tak, H. Shim, J. S. Chung, B. Lee, H. Yu, and N. Evans, "Aasist: Audio anti-spoofing using inte grated spectro-temporal graph attention networks," *arXiv preprint arXiv:2110.01200*, 2021.

[13] X. Wang and J. Yamagishi, "A Comparative Study on Recent Neu ral Spoofing Countermeasures for Synthetic Speech Detection," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2021, pp. 4259–4263.

[14] J.-w. Jung, Y. J. Kim, H.-S. Heo, B.-J. Lee, Y. Kwon, and J. S. Chung, "Pushing the limits of raw waveform speaker recognition," *Proceedings of the 23rd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022.

[15] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spec tra information for spoofing attack detection," *IEEE/ACM Trans actions on Audio, Speech, and Language Processing*, pp. 2373– 2384, 2019.

[16] Z. Ge, A. N. Iyer, S. Cheluvaraja, R. Sundaram, and A. Ganap athiraju, "Neural network based speaker

classification and verifi cation systems with enhanced features," in *Proceedings of 2017 Intelligent Systems Conference (IntelliSys)*, 2017, pp. 1089–1094.

[17] S. Ding, T. Chen, X. Gong, W. Zha, and Z. Wang, "Autospeech: Neural architecture search for speaker recognition," in *Proceed ings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 916– 920.

[18] Q. Fu, Z. Teng, J. White, M. Powell, and D. C. Schmidt, "Fas taudio: A learnable audio front-end for spoof speech detection," *arXiv preprint arXiv:2109.02774*, 2021.

[19] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "Resnet and model fu sion for automatic spoofing detection." in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 102–106.

[20] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, "Large-scale self-supervised speech representation learning for automatic speaker verification," in *Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[21] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation," 2022.

[22] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *CoRR*, 2021.

[23] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *Proceed ings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.

[24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.

[25] H. Tak, J.-w. Jung, J. Patino, M. Todisco, and N. Evans, "Graph attention networks for anti-spoofing," *arXiv preprint arXiv:2104.03654*, 2021.

[26] P. Velickovi ˇ c, G. Cucurull, A. Casanova, A. Romero, P. Liò, and ´ Y. Bengio, "Graph Attention Networks," in *Proceedings of 6th International Conference on Learning Representations*, 2018.

[27] J.-W. Jung, H.-S. Heo, H.-J. Yu, and J. S. Chung, "Graph attention networks for speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Pro cessing (ICASSP)*, 2021, pp. 6149–6153.

[28] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[29] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representa tions," *CoRR*, 2020.

[30] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, 2021.

[31] H. Tak, J. weon Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detec tion," in *Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021.

[32] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Pro ceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[33] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. He manth, "Ssdmnv2: A real time dnn-based face mask detection system using single shot multibox detector and mobilenetv2," *Sus tainable cities and society*, 2021.

[34] M. Sukhavasi and S. Adapa, "Music theme recognition using cnn and self-attention," *arXiv preprint arXiv:1911.07041*, 2019.