

Multi-Microphone Speech Emotion Recognition Using the Hierarchical Token-Semantic Audio Transformer

Ohad Cohen

Faculty of Engineering
Bar-Ilan University
Ramat-Gan, Israel
ohad.cohen@biu.ac.il
0009-0002-4707-881X

Gershon Hazan

Faculty of Engineering
Bar-Ilan University
Ramat-Gan, Israel
gershon.hazan@biu.ac.il

Sharon Gannot

Faculty of Engineering
Bar-Ilan University
Ramat-Gan, Israel
sharon.gannot@biu.ac.il
0000-0002-2885-170X

Abstract—The performance of most emotion recognition systems degrades in real-life situations (“in the wild” scenarios) where the audio is contaminated by reverberation. Our study explores new methods to alleviate the performance degradation of Speech Emotion Recognition (SER) algorithms and develop a more robust system for adverse conditions. We propose processing multi-microphone signals to address these challenges and improve emotion classification accuracy. We adopt a state-of-the-art transformer model, the Hierarchical Token-semantic Audio Transformer (HTS-AT), to handle multi-channel audio inputs. We evaluate two strategies: averaging mel-spectrograms across channels and summing patch-embedded representations. Our multi-microphone model achieves superior performance compared to single-channel baselines when tested on real-world reverberant environments.

Index Terms—speech emotion recognition, human-robot interaction

I. INTRODUCTION

SER is widely studied in the literature. Most of the reported studies deal with clean speech data and do not consider additive noise and reverberant environments typical to real-life applications. Only a few studies address the influence of reverberation and noise on SER. In these studies, the reverberant data is artificially generated by convolving the clean utterances with Room Impulse Responses (RIRs), either simulated [1], [2] or recorded in a real environment with various reverberation levels [3], [4]. Noise may also be added to the reverberant signals.

A survey of Human-Robot Interaction (HRI) in a noisy environment can be found in [5], [6]. However, this survey does not address the SER task. Previous works have shown the significant challenges of detecting emotional speech in large and reverberant rooms. Reverberation can influence the speech signal and negatively affect the predicted results [7]. The literature on using multiple microphones for SER is very scarce. In [8], the robustness of Automatic Speech Recognition (ASR) systems to emotional speech in noisy conditions is addressed. A Binaural Emotional Speech Recognition (BESR) system is proposed, enabling the simultaneous acquisition of the speaker’s emotional state and transcribing the uttered speech signal. Devices equipped with multi-microphones are widely available nowadays. Adding the spatial information may improve the performance of audio processing tasks, including SER. However, learning-based algorithms are challenged by differences between training and test conditions, specifically a change in the microphone array constellations, e.g., when the number of microphones in train and test conditions is different.

In early works in the SER domain, various architectures such as Convolutional Neural Networks (CNN), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Long Short-term Memory (LSTM) were employed. Subsequently, combinations of CNN and RNN layers emerged, showcasing enhanced performance

as compared with traditional classification methodologies [9], [10], [11]. Additionally, the effectiveness of configurations utilizing blocks comprised of CNN, LSTM and Bidirectional Long Short-term Memory (BiLSTM), as elucidated in [12], [13], was proven effective. Most of the works mainly focus on unimodal learning of emotions, either text, speech, or video [14], [15], [16].

The Transformer architecture, initially formulated for Natural Language Processing (NLP) [17], [18], has found applications in the audio processing domain, including tasks such as speech separation [19] and audio classification [20]. The superior performance of the Audio Spectrogram Transformer (AST) model [20], an adaptation of the Vision Transformer (ViT) model [21], was demonstrated compared to CNN-based models. It is imperative to highlight that the AST model exclusively handles single-microphone data, whereas, in real-world scenarios, multiple microphones may often be available. A drawback of Transformer models is their reliance on large training data for convergence.

In the current work, we extend the HTS-AT [22] architecture to accommodate multi-channel inputs, thus enhancing robustness against reverberations. Since only a limited amount of data is available for our task, we resort to fine-tuning already-trained models. Moreover, our scheme can be fine-tuned with a certain number of microphones and tested with another microphone constellation, including microphone positions and the number of microphones. We evaluated the SER performance using three datasets, Ryerson audio-visual database of emotional speech and song (RAVDESS) [23], Interactive Emotional Dyadic Motion Capture (IEMOCAP) [24], and Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [25]. We used real-life RIRs from the Acoustic Characterisation of Environments (ACE) Challenge [26] to add reverberation to the speech utterances.

II. PROBLEM FORMULATION

Let $x(t)$ denote the anechoic signal in the discrete-time domain. An M -microphone array captures this signal after propagating in the acoustic enclosure. The received microphone signals are then given by $y_i(t) = \{x * h_i\}(t)$, $i = 1, 2, \dots, M$, where $h_i(t)$ is the RIR from the source position to the position of microphone i . This work aims to classify the emotion given the observations $y_i(t)$; $i = 1, \dots, M$.

III. PROPOSED MODEL

The proposed SER model is based on the Swin-Transformer [27]¹, a state-of-the-art hierarchical ViT [21] architecture, used for a variety of computer vision tasks, that utilizes shifted windows to capture long-range image dependencies. We adopted and modified a Swin Transformer architecture, namely the Hierarchical Token-semantic

¹github.com/microsoft/Swin-Transformer

Audio Transformer (HTS-AT) [22], which aims to improve the performance and scalability of audio tasks, such as the AudioSet dataset [28]. The HTS-AT model is designed to achieve the best performance by reducing the number of parameters, requiring fewer GPU resources, and less training time than the AST architecture [22]. In the current work, we propose modifying the HTS-AT architecture to suit the audio processing requirements better. The main change is to adapt the model to the multi-channel audio processing task, expanding its usefulness beyond the original single-channel design. Two alternative multi-channel pre-processing strategies are examined. The first strategy applies a summation of patch tokens derived from mel-spectrograms, and the second strategy applies averaging of mel-spectrograms. Our proposed approach first fuses the multi-channel information into a single stream. Then, the regular HTS-AT architecture is applied.

A. Preprocessing and Input Features

We assume that the sampling rate of the audio signals is 16 kHz. Each microphone signal is first analyzed by a short-time Fourier transform (STFT), with a window size of 1024 samples and a hop size of 160 samples. The log-absolute value of the STFT bins are then aggregated to construct the mel-spectrograms.

B. Architecture

We will now elaborate on the HTS-AT architecture. The standard transformer architecture requires extensive computational resources due to the unmodified input token sequence length across all layers. This includes maintaining a large global self-attention matrix and calculating outputs and gradients at each step. The HTS-AT architecture is introduced to address these challenges. The HTS-AT is designed for supporting multiple audio tasks, e.g., classification, sound event detection, and source localization. It introduces two critical architectural optimizations: a hierarchical structure and a windowed attention mechanism. The input audio mel-spectrogram is split into localized patch tokens using a convolutional Patch-Embed CNN layer of kernel dimensions $P \times P$, in which the patches are ordered by time segment and frequency bin. Then, tokens propagate through a series of Swin Transformer encoder groups. At the terminus of each group, a Patch Merging layer reshapes the token sequence into its original 2D mel-spectrogram. This layer merges neighboring patches and then embeds them back into a latent space of reduced length. Consequently, the memory requirements decrease exponentially with depth. Within each Swin Transformer block, window attention is restricted to non-overlapping $W \times W$ squares, partitioning the token sequence. Calculating self-attention within each window subset substantially reduces computational complexity relative to full global attention while capturing localized relationships. As patch size increases downstream, windows encapsulate larger temporal and frequency contexts. Finally, the HTS-AT incorporates a token-semantic CNN layer after the last transformer block. This layer refines the output by grouping tokens, thus capturing information about their time frames and frequency bins. Consequently, this enhances classification by exploiting token relationships.

C. Multi-Channel Methods

We propose two multi-channel approaches that enable the HTS-AT model to handle single- and multi-channel inputs through fine-tuning. Both methods maintain HTS-AT's core architecture to allow flexibility in microphone numbers during fine-tuning and evaluation.

Patch-Embed Summation: Inspired by the study in [29], we embraced a Patch-Embed scheme where each channel undergoes a

shared embedding layer followed by summation. This maintains flexibility in the microphone numbers used during fine-tuning and testing without altering the core HTS-AT architecture. On the left side of Fig. 1, each of the M channels is analyzed to generate mel-spectrograms, which are then concatenated along the channel depth. SpecAugment [30] is applied collectively across all M mel-spectrograms. Subsequently, each mel-spectrogram is reshaped to the dimensions of a 256×256 image and passes through the shared Patch-Embed layer. Next, a summation operation is performed across all M -encoded channels, making this input suitable for the pre-trained HTS-AT. The summation consolidates inter-channel information, enabling more robust representations suitable for reverberant conditions. The summed embeddings are then fed to the HTS-AT feedforward block for the actual emotion classification.

Average Mel-Spectrograms: On the left side of Fig. 2, the mel-spectrogram of each channel is calculated, followed by an averaging stage that produces a single-channel representation, regardless of the original number of channels. Since reverberation affects each channel differently, the averaged spectrogram will tend to be less reverberant. The averaged mel-spectrogram is then augmented and structured into an image. Subsequently, the image is fed to the Patch-Embed encoder layer, generating a feature tensor with dimensions of 4096×96 . Similarly to the first architecture, the HTS-AT feedforward block implements the emotion classification.

IV. EXPERIMENTAL STUDY

In this section, we describe the experimental setup and present a comparative study of the proposed scheme and a baseline method.

A. Datasets

Our study comprised three speech emotion recognition datasets, namely RAVDESS[23], IEMOCAP[24] and CREMA-D[25] datasets.

The **RAVDESS** dataset comprises 24 actors, evenly distributed between male and female speakers, each uttering 60 English sentences. Hence, there are 1440 utterances expressing eight different emotions: ‘sad’, ‘happy’, ‘angry’, ‘calm’, ‘fearful’, ‘surprised’, ‘neutral’, and ‘disgust’. All utterances are transcribed in advance. Consequently, emotions are more artificially expressed than in spontaneous conversation. In this dataset, we decided to merge the emotions ‘neutral’ and ‘calm’ as representations of ‘neutral’. Therefore, the output of our model is on seven classes instead of eight. A significant drawback of the dataset is its small number of utterances.

The **IEMOCAP** dataset comprises approximately 12 hours of speech expressing four emotions: ‘happy’, ‘sad’, ‘angry’ and ‘neutral’. It consists of conversations between two people that are either improvised or played according to a pre-determined transcript chosen to evoke different emotions.

The **CREMA-D** is a dataset of 7442 original clips from 91 actors comprising 48 male and 43 female actors. Speech utterances were selected from a set of 12 sentences. The sentences were presented using one of six different emotions ‘anger’, ‘disgust’, ‘fear’, ‘happy’, ‘neutral’ and ‘sad’.

For the multi-channel experiments, we first split the original three datasets as follows: 80% as a training set, 10% as a validation set, and 10% as a test set. We note that the same actor may appear in several splits but not the same utterance. Since multi-microphone datasets for SER are unavailable, we generated our own dataset. We used the ‘gpuRIR’ Python package² to simulate reverberant multi-channel microphone signals (setting the number of microphones to

²github.com/DavidDiazGuerra/gpuRIR

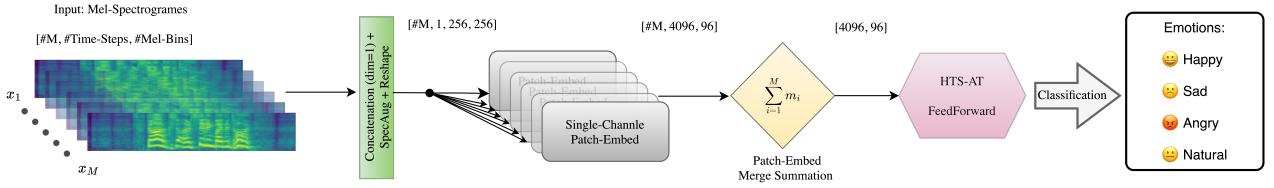


Fig. 1: Scheme of Patch-Embed Summation.

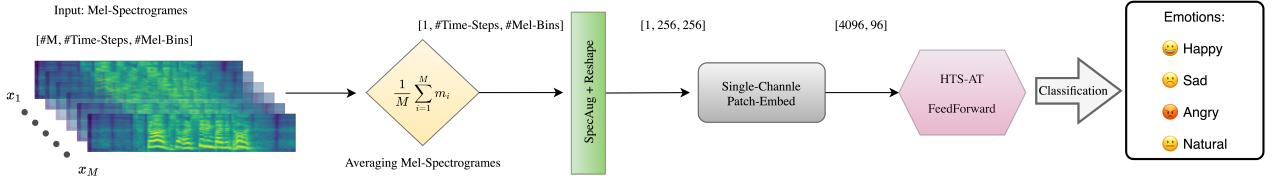


Fig. 2: Scheme of Average Mel-Spectrograms.

TABLE I: The results for the non-reverberant single-microphone case. The fine-tuned HTS-AT model is compared with [12]. We present the weighted average accuracy (in percent) of 20 models that achieved the highest accuracy on the validation set and tested on the test set.

Models	RAVDESS	IEMOCAP	CREMA-D
Fine-tuned HTS-AT	90%	70.93%	75.86%
BiLSTM + Attention[12]	80%	66%	-

$M = 4$) with reverberation time in the range of $T_{60} = 200 - 800$ ms with different room sizes and randomized microphone locations. The room dimensions were uniformly distributed between 3 and 8 meters, with an aspect ratio of 1 to 1.6. The room's height was fixed at 2.9 meters. The sound source was placed at a constant height of 1.75 meters, while the microphones were positioned at a fixed height of 1.6 meters. The x and y coordinates were randomly determined within the room for both the sound source and microphones while keeping them at least 0.5 meters away from the room walls. By doing so, we were also able to enlarge the datasets. For RAVDESS, 6863 reverberant speech samples were generated for training, and 852 samples were used for validation. We follow the same procedure for IEMOCAP with 7356 train samples and 2107 validation samples and for CREMA-D with 5945 train samples and 1487 validation samples. Our objective encompasses evaluating our model in real-world reverberant environments. To achieve this, we conducted tests utilizing the ACE RIR database [26], which comprises seven distinct rooms characterized by varying dimensions and exhibiting diverse ranges of T_{60} . We used the subset captured by a mobile phone equipped with three microphones, which are frequently used for speech interaction, making them a practical choice for real-world SER applications. The model was fine-tuned with the synthesized RIRs and evaluated with the test sets of the various datasets convolved with the ACE RIRs.

B. Algorithm Setup

We fine-tuned our model starting from the original HTS-AT model that was pre-trained utilizing the AudioSet dataset, which comprises over two million audio samples. Each sample is 10 seconds long and is categorized into 527 distinct sound event categories. This pre-training provides a strong foundation for audio feature extraction, which we then adapt to our specific SER. In order to adapt the pre-

trained AduioSet model to RAVDESS, IEMOCAP, and CREMA-D, we adjusted the number of output classes to 7, 4, and 6, respectively. In the three datasets, pre-processing and warm-up strategy were carried out as demonstrated in [22] by providing the HTS-AT with 64 mel-bins to compute the STFT and mel-spectrograms features with 160 hop size and 1024 window size. We modified the original AdamW optimizer of HTS-AT to the traditional Adam optimizer with a learning rate of $1e^{-3}$, and batch size of 128. We set the number of epochs to 150 for all datasets. We used cross-entropy loss as the metric. We also follow the same original HTS-AT hyperparameter settings. The patch dimensions are set to 4×4 , the patch window length is 256 frames, and the attention window size is 8×8 . The architectural configuration comprises four network groups, each comprising 2, 2, 6, and 2 Swin-Transformer blocks, respectively. The initial patch embed is linearly projected to a dimension of $D = 96$, and correspondingly, after each transformer group, the dimension increases exponentially to $8D = 768$, aligning seamlessly with the principles of AST. In part of our experiments, we reduced the depth of each network group by half, thereby reducing the number of trainable parameters by half to prevent overfitting. In addition, we added an early stopping strategy with a patience of 50 for RAVDESS and 25 for IEMOCAP and CREMA-D. The overall number of parameters for the fine-tuned HTS-AT was 15.7M for the RAVDESS and 28.6M for IEMOCAP and CREMA-D. Since only fine-tuning is required, we could train the models in less than two hours on an A6000 RTX GPU.

C. Results

We begin our experimental study with the single-microphone case. We compare the fine-tuned HTS-AT model with another model developed in our group that utilizes BiLSTM combined with an Attention mechanism [12], which has achieved very good results on the RAVDESS and IEMOCAP datasets. We applied both schemes to a non-reverberant single-microphone signal. The results are presented in Table I.

Furthermore, to complete establishing the baseline using the RAVDESS and IEMOCAP datasets, we fine-tuned another single-channel HTS-AT model by fine-tuning on the artificially reverberated datasets with a uniformly distributed $T_{60} = 200 - 800$ ms. Then, we compared the three models by computing the Accuracy with

TABLE II: Results on the test sets of the RAVDESS, IEMOCAP, and CREMA-D datasets convolved with RIRs of three microphones from the ACE database. The ‘HTS-AT’ columns are fine-tuned on reverberant single-channel audio. The ‘Avg mel’ columns depict results where mel-spectrograms were averaged across four channels during fine-tuning and tested on three channels. The ‘Sum PE’ columns are the Patch-Embed Summation approach fine-tuned and tested on three channels.

Room (T_{60} [ms])	RAVDESS			IEMOCAP			CREMA-D		
	HTS-AT	Avg mel	Sum PE	HTS-AT	Avg mel	Sum PE	HTS-AT	Avg mel	Sum PE
Lecture 1 (638)	77.3 (70.6-84.0)	80.6 (74.6-86.6)	81.3 (74.6-87.3)	61.3 (57.1-65.2)	67.0 (63.0-71.0)	67.4 (63.6-71.2)	63.2 (60.1-66.6)	66.4 (63.2-70.1)	67.4 (64.2-71.0)
Lecture 2 (1220)	77.3 (70.6-83.3)	78.6 (71.3-84.6)	82.0 (76.0-88.0)	63.4 (59.5-67.4)	66.1 (62.4-70.0)	68.7 (65.0-72.4)	65.4 (62.0-68.9)	67.3 (64.0-70.8)	66.0 (62.5-69.3)
Lobby (646)	78.6 (72.0-84.6)	82.0 (76.0-88.0)	84.0 (78.0-89.3)	61.8 (58.0-65.8)	64.0 (60.0-67.8)	65.1 (61.3-68.8)	64.2 (60.9-67.8)	66.2 (62.8-69.8)	66.8 (63.3-70.2)
Meeting 1 (437)	73.3 (66.6-80.0)	80.6 (74.0-86.6)	82.6 (76.6-88.0)	62.9 (58.9-66.9)	66.7 (62.9-70.6)	68.7 (65.1-72.4)	64.4 (61.0-67.8)	65.7 (62.5-69.3)	65.7 (62.2-69.0)
Meeting 2 (371)	82.0 (76.0-87.3)	83.3 (76.6-89.3)	85.3 (80.0-90.6)	59.5 (55.7-63.6)	66.3 (62.4-70.3)	64.2 (60.2-68.1)	65.6 (62.4-69.0)	66.6 (63.4-69.8)	66.8 (63.6-70.2)
Office 1 (332)	76.0 (69.3-82.6)	83.3 (77.3-88.6)	80.6 (74.0-86.6)	63.6 (59.7-67.6)	68.3 (64.7-72.3)	67.2 (63.6-71.0)	64.1 (60.1-67.6)	68.8 (65.4-72.4)	67.3 (64.0-71.0)
Office 2 (390)	78.6 (72.6-84.6)	81.3 (75.3-87.3)	82.6 (76.0-88.0)	59.5 (55.5-63.8)	64.7 (61.0-68.8)	65.4 (61.5-69.6)	62.6 (59.4-66.1)	64.0 (60.6-67.6)	64.1 (60.6-67.7)

statistical significance measure.³ The evaluation was performed on the reverberant test set using the ACE database only applied to one of the microphones. Examining Fig. 3 for both the IEMOCAP

channel scheme, 2) ‘Avg mel’ - mel-spectrograms averaged across four channels, and 3) ‘Sum PE’ - Patch-Embed Summation across three channels. In the table, we report the accuracy and the associated confidence interval for all seven rooms and three databases. Examining these results indicates the consistent, albeit modest, improvements achieved by the ‘Sum PE’ approach. Embedding each channel separately before summing, the model captures subtle inter-channel differences that might be lost in averaging. This allows the model to emphasize the most informative aspects of each channel for emotion recognition, leading to improved performance. Importantly, these benefits come without a considerable increase in computational complexity compared to single-channel models.

V. CONCLUSIONS

In this paper, we presented a multi-microphone transformer-based model for SER in reverberant environments. Based on the HTS-AT architecture, the model employs two strategies for handling multi-channel audio inputs: averaging mel-spectrograms across channels and summing patch-embedded representations. When tested on real-world reverberant environments, the results show improved SER accuracy compared to single-channel schemes. By leveraging spatial information from multiple microphones, our model was able to exhibit a more robust behavior of the SER in challenging acoustic conditions. The consistent, but not large, improvements of the proposed multi-microphone schemes show promise for developing SER systems that can perform reliably in noisy and reverberant scenarios.

REFERENCES

- [1] T. Guo, S. Li, M. Unoki, and S. Okada, “Investigation of noise-reverberation-robustness of modulation spectral features for speech-emotion recognition,” in *Asia-Pacific Signal and Information Processing (APSIPA)*, 2022, pp. 39–46.
- [2] L. N. S. Wijayasingha and J. A. Stankovic, “Robustness to noise for speech emotion classification using CNNs and attention mechanisms,” *Smart Health*, 2020.
- [3] F. Eyben, F. Weninger, and B. Schuller, “Affect recognition in real-life acoustic conditions: a new perspective on feature selection,” in *INTERSPEECH*, Lyon, France, 2013.
- [4] M. Y. Ahmed, Z. Chen, E. Fass, and J. Stankovic, “Real time distant speech emotion recognition in indoor environments,” in *Int. Conf. on Mobile and Ubiquitous Systems: Computing, Networking and Services*, 2017, pp. 215–224.
- [5] S. Zhao, Y. Yang, and J. Chen, “Effect of reverberation in speech-based emotion recognition,” in *IEEE Int. Conf. on the Science of Electrical Engineering in Israel (ICSEE)*, 2018.
- [6] H. W. Löllmann, H. Barfuss, A. Deleforge, S. Meier, and W. Kellermann, “Challenges in acoustic signal enhancement for human-robot communication,” in *Speech Communication; ITG Symposium*, 2014.
- [7] P. Heracleous, K. Yasuda, F. Sugaya, A. Yoneyama, and M. Hashimoto, “Speech emotion recognition in noisy and reverberant environments,” in *Int. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 262–266.

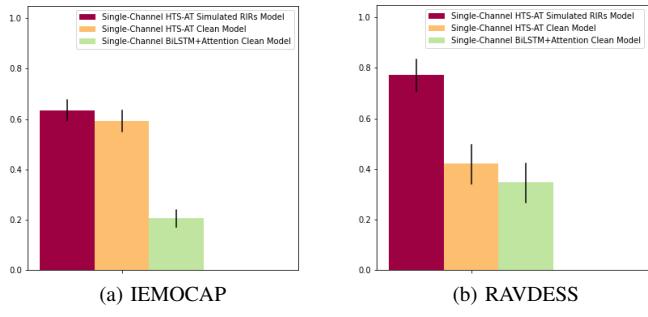


Fig. 3: Accuracy and Confidence Interval on test sets convolved with ACE RIR Lecture Room 2 ($T_{60} = 1220$ ms). The results of two HTS-AT fine-tuned on either clean or simulated RIRs datasets compared with [12] trained on the clean datasets.

and RAVDESS datasets, it is evident that the two HTS-AT variants outperform the approach in [12]. It is also clear that for RAVDESS, the HTS-AT fine-tuned with reverberant speech significantly outperforms the HTS-AT fine-tuned using clean data. While this is also true for IEMOCAP, the differences are less significant. In addition, the significant gap observed for the RAVDESS dataset compared to IEMOCAP is due to differences in the nature of the datasets. RAVDESS contains more controlled, acted emotions, while IEMOCAP is more naturalistic.

We now turn to the evaluation of the multi-channel schemes. In all experiments, we tested all three datasets with the utterances convolved with the RIRs of the ACE database. We only report the results for the remote source case to emphasize the reverberation effects. The average mel-spectrogram scheme was fine-tuned using four microphones. The Patch-Embed Summation scheme was fine-tuned using three microphones. While we used different microphone counts ($M = 3$ and $M = 4$) for fine-tuning, this was done to demonstrate the flexibility of our approach in handling varying numbers of input channels. However, we acknowledge that using the same number of microphones for all models would provide a more direct comparison. In both cases, the RIRs were generated using the simulator, as explained above. All experimental procedures followed identical training-validation splits, maintaining consistent model configurations, sizes, and hyperparameters for each dataset.

Table II provides a comparative assessment of three model variants, all fine-tuned with reverberant speech: 1) ‘HTS-AT’ - a single-

³github.com/luferrer/ConfidenceIntervals

- [8] M. Bashirpour and M. Geravanchizadeh, "Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments," *EURASIP J. on Audio, Speech, and Music Processing*, vol. 2018, no. 1, pp. 1–13, 2018.
- [9] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Asia-Pacific Signal and Information Processing (APSIPA)*, 2016.
- [10] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Interspeech*, 2018, pp. 3683–3687.
- [11] C.-W. Huang, S. Narayanan *et al.*, "Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition," *arXiv preprint arXiv:1706.02901*, 2017.
- [12] D. Sherman, G. Hazan, and S. Gannot, "Study of speech emotion recognition using BLSTM with attention," in *European Signal Processing Conf. (EUSIPCO)*, Helsinki, Finland, Sep. 2023.
- [13] S. Kwon *et al.*, "CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network," *Mathematics*, vol. 8, no. 12, p. 2133, 2020.
- [14] S. K. Bharti, S. Varadhanapathy, R. K. Gupta, P. K. Shukla, M. Bouye, S. K. Hingaa, and A. Mahmoud, "Text-based emotion recognition using deep learning approach," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 2645381, 2022.
- [15] A. Bhavan, P. Chauhan, R. R. Shah *et al.*, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.
- [16] M. Abdullah, M. Ahmad, and D. Han, "Facial expression recognition in videos: An cnn-lstm based model for video classification," in *Int. Conf. on Electronics, Information, and Communication (ICEIC)*, 2020.
- [17] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for NLP tasks," in *Conf. on Computer Science and Information Systems (FedCSIS)*, 2020, pp. 179–183.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [19] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 21–25.
- [20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Interspeech*, 2021, pp. 571–575.
- [21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. on Learning Representations (ICLR)*, 2021.
- [22] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 646–650.
- [23] S. R. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
- [24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [25] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenкова, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [26] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1681–1693, 2016.
- [27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF Int. Conf. on computer vision*, 2021, pp. 10 012–10 022.
- [28] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE Int. Conf. on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 776–780.
- [29] A. Eliav and S. Gannot, "Concurrent speaker detection: A multi-microphone transformer-based approach," in *European Signal Processing Conf. (EUSIPCO)*, France, Aug. 2024.
- [30] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, 2019, pp. 2613–2617.