Multiroom Speech Emotion Recognition

Erez Shalev Israel Cohen
Andrew and Erna Viterbi Faculty of Electrical & Computer Engineering
Technion–Israel Institute of Technology, Haifa 3200003, Israel
{erezsh@ef, icohen@ee}.technion.ac.il

***Abstract*—Automated audio systems, such as speech emotion recognition, can benefit from the ability to work from another room. No research has yet been conducted on the effectiveness of such systems when the sound source originates in a different room than the target system, and the sound has to travel between the rooms through the wall. New advancements in room-impulse-response generators enable a large-scale simulation of audio sources from adjacent rooms and integration into a training dataset. Such a capability improves the performance of data-driven methods such as deep learning. This paper presents the first evaluation of multiroom speech emotion recognition systems. The isolating policies due to COVID-19 presented many cases of isolated individuals suffering emotional difficulties, where such capabilities would be very beneficial. We perform training, with and without an audio simulation generator, and compare the results of three different models on real data recorded in a real multiroom audio scene. We show that models trained without the new generator achieve poor results when presented with multiroom data. We proceed to show that augmentation using the new generator improves the performances for all three models. Our results demonstrate the advantage of using such a generator. Furthermore, testing with two different deep learning architectures shows that the generator improves the results independently of the given architecture.**

***Index Terms*—Emotion recognition, acoustics, room impulse response, multiroom, augmentation.**

I. Introduction

Given an audio segment of speech, the task of detecting the emotional state of the speaker is called speech emotion recognition (SER). Such technology can help with customer support review and analysis, human-machine interaction, mental health monitoring, etc. Mental health monitoring is even more critical with COVID-19 influences [1], and additional options, such as monitoring isolated personals' distress. The vast research explores many aspects of SER. For example, [2] explored the input features for SER classifiers. Several supporting modalities were examined, such as visual-cues [3], bio-signals [4], and textual information [5]. Many classification models were tested, such as, Hidden Markov Models (HMM) [6], Gaussian Mixture Model (GMM) [7], and support vector machines (SVM) [8].

Recently, deep learning methods have shown promising SER results. Several deep neural networks (DNN) architectures were offered, amongst which are convolutional neural networks (CNN), recurrent neural networks (RNN) using long short term memory (LSTM) or gated recurrent units (GRU), time-delay neural networks (TDNN), residual networks (ResNet), dilated residual networks (DRN), to name a few.Knowledge transfer between models was also studied for different data and different domains [9]. More architectures are constantly tested, such as graph convolution networks (GCN)-based architecture and attention rectified linear units GRU (AR-GRU) [10], [11]. Many more aspects of SER are being explored, e.g., reduction of the computational complexity [12].

Despite extensive research in the SER field, none of the existing techniques considers the scenario where the speech sounds source is located in a different room from the SER system, and the audio travels through a joint wall. This issue is becoming vital with the increasing distress situations introduced by COVID-19's social isolation policy. For example, a rise in domestic violence has been documented [13], a higher suicide rate [14], and other emotional responses with children, the elderly community, people coping with mental conditions, and the generally lonely personals. In such cases, multiroom emotion recognition can be very beneficial. Unfortunately, such audio scenes are not included in existing SER datasets. The recordings are either made in a clean environment or have single room reverberation characteristics inherent to the data and cannot be controlled by the researcher.

In this paper, we study the problem of SER given a source and a receiver located in different coupled rooms. We present three models for SER and evaluate their performances on real data recorded from another room. We train the models using our augmentation method presented in [15] and assess the influence of this augmentation on SER performance. Our results show the benefit of using the multiroom impulse response generator [15]. We show that the performances without a generator's augmentation are not better than a guess for this use-case. We demonstrate this on two different architectures.

The rest of this paper is organized as follows: Section II elaborates on the signal representation and evaluation metrics. Section III details the datasets used and the pre-processing procedure. The architectures for the classification models are described in Section IV. Section V presents the experiments and results. Finally, we conclude with a discussion in Section VI.

II. Problem Formulation

We consider an audio signal $x(t)$, which is a recording of emotional speech in a clean recording environment with minimum reverberation. For the acoustic environment simulation, we produced a room impulse response (RIR), $h(t)$. In this case, the input to our model is given by

$$y(t) = h(t) * x(t) + w(t), \tag{1}$$

where $*$ is the convolution operation and $w(t)$ is white Gaussian noise (WGN). When training without the environmental influences, the input signal to the system is simply

$$y(t) = x(t). \tag{2}$$

We consider a labeled dataset of N speech samples $x_i(t)$, with the respective label $c_i \in \mathcal{C}$, where $\mathcal{C}$ is the set of all possible labels, and $0 < i \leq N$. Given a classification model M, the predicted label $\hat{c}_i$, is given by the activation of the model on the input sample $y_i(t)$,

$$\hat{c}_i = M\{y_i(t)\}. \tag{3}$$

For the evaluation of the models, we use three metrics. We follow measures from previous work, such as [9]–[12]. Given the true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) predictions, the balanced accuracy (BA) is defined as

$$P_{BA} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{4}$$

the unbalanced accuracy (UA) is defined as

$$P_{UA} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right), \tag{5}$$

and the F1 measure is defined as

$$P_{F1} = \frac{2TP}{2TP + FP + FN}. \tag{6}$$

The P_{BA} measure is a simple success rate, counting the percentage of correct classification. The P_{UA}, is an accuracy measure meant for a similar assessment over an unbalanced dataset, where different classes contain a different number of samples. For a balanced dataset, the P_{BA} and P_{UA} will converge to the same score. The P_{F1} score is the harmonic mean between precision and recall. The maximal score for all of these measures is 100%.

III. Datasets

We use a combination of three datasets for the training and evaluation of the SER models: Berlin Emotional Dataset (EmoDB) [16], Toronto Emotional Speech Database (TESS) [17], and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [18]. For simplicity, we limit the datasets to the specific emotion labels ['angry', 'sad', 'neutral', 'ps', 'happy'] (where 'ps' is pleasant surprise). Each audio file was loaded with a sample rate of 48 kHz, and each sample is zero-padded up to 309500 samples (approximately 6.448 seconds). The zero paddings are used to fix the input length and the number of time bins T. We proceed to extract both Mel-frequency cepstral coefficients (MFCC) and Mel-spectrogram coefficients, which are extracted from a short-time Fourier transform (STFT). The STFT was calculated with 2048 frequency bins, using a Hanning window of 2048 samples with 512 samples hop length. This results in a $T = 605$ time bins. A total of 45 and 128 Mel-frequency cepstral and Mel-spectrogram coefficients were calculated, respectively. Both MFCC and Mel-spectrogram features were concatenated to create a $(T \times F) = (605 \times 173)$ feature map. We have extended the third dimension with a size 1 to represent a single channel, where the model requires a channel dimension.

In addition to the three datasets, we have recorded sounds in a real multiroom audio scene. This real-test set comprises 75 samples, 15 for each of the five labels. In this real-test, the receiver room was of size $(H \times D \times W) = (2.75 \times 1.5 \times 2.5)$ m, and the source room was of size $(H \times D \times W) = (2.75 \times 3.1 \times 2.5)$ m. Both rooms are real adjacent rooms containing various furniture and other absorbing and reflecting objects. Between recordings, we altered the locations of the source and receiver within their respective rooms. We used two standard cellular phone devices for our recordings, which should best represent the data quality that such a system is most likely to encounter in real life. We kept the sampling rate and used mono-recordings in MP4 file format.

IV. Models

We used three DNN architectures, namely a CNN AlexNet model [19], and two custom RNN models. All three were trained with and without augmentation, and were evaluated using the metrics defined in Section II.

A. AlexNet

The original AlexNet is designed for a square, 227×227 image as an input. To fit our feature map into the network, we have altered the first layer to receive a single channel, $(T \times F)$, image. We kept the original $(11, 11)$ filter size but altered the stride to $(11, 3)$ so that it fits our $(T \times F)$ size. As a result, the rest of the model remains unchanged. When generating the dataset as described in Section III, we added the channel dimension. The AlexNet was trained using 150 epochs, using an Adam optimizer, with early stopping. A patience of ten epoches was used as a condition for the early stopping.

B. RNN

We created 2 RNN customized networks. All the parameters were empirically chosen with respect to the datasets and task use-case. The first two layers are RNN layers in both networks with 128 bi-directional units. The difference between the RNNs is the type of RNN units, namely GRU or LSTM. Each RNN layer is followed by a dropout layer with the value 0.3 and 0.2, for GRU and LSTM, respectively. The subsequent two layers are dense, fully connected layers, with rectified linear units (ReLU) activation, followed by the same dropout layer as the respective RNN unit. Lastly, we added a final dense layer with soft-max activation and a one-hot labeling output. The RNN models were trained for 70 epochs, using Adam optimizer, with early stopping.

C. Architecture discussion

AlexNet is a standard classifier and was also used in [15] for audio classification. A CNN architecture considers the input as an image, where one axis represents the time bins, and

TABLE I: Performance on the combined datasets with respect to a synthetic evaluation test-set. The models were trained without any augmentation.

Architecture	UA	BA	F1
AlexNet	84.18%	84.44%	84.45%
LSTM-RNN	83.74%	83.04%	83.54%
GRU-RNN	**85.94**%	**85.61**%	**85.86**%

TABLE II: Performance on the real data. The models were trained without any augmentation.

Architecture	UA	BA	F1
AlexNet	20%	20%	6.66%
LSTM-RNN	21.33%	21.33%	9.22%
GRU-RNN	**26.66**%	**26.66**%	**18.09**%

the other represents the frequency bins. Since the convolution is two-dimensional, the architecture can extract both inter-frequency and sequential data and the connections between them. However, the filter size is fixed, limiting the time length of the sequential feature extraction.

Conversely to CNNs, RNNs are specifically tailored for the extraction of sequential data. The advantage of RNN is in learning the length of contextual sequence relevant to the task at hand. Considering our task, we note that emotion in speech typically has a sequential factor and dependency between consequent words and tonality. Even the tempo of the sentence can contain valuable classification information. Therefore RNN architectures are also prime candidates for this task, due to their relative benefits in classifying sequential data. To evaluate adjacent room SER, we designed two custom bi-directional RNN models. We have to keep in mind that the RNN units cannot extract inter-frequency information, and the architecture depends on the following dense units for that purpose.

Current literature is indecisive regarding the optimal architecture for audio tasks, RNN or CNN [20]. The literature is also unclear on which unit is the best RNN unit [21]. The GRU has less control over the amount of memory exposed to other units in the network. Also, both use a different method for controlling the amount of memory from previous steps, exposed for the calculation of the current activation. While both CNN and RNN architectures have pros and cons, and both GRU and LSTM units achieve comparable performances, we have to test all three of our models to find the optimal method for SER in general, and for SER from an adjacent room in particular.

D. Augmentation method

It is possible to train a simple model for an SER task using the existing datasets, and then evaluate its performance when the audio arrives from another room. However, it is highly beneficial to integrate a simulation of the audio environment into the training phase [15]. A new RIR generation method, namely the structure image method (StIM) [15], enables to simulate the transition of sound through a wall between two adjacent rooms.

StIM starts by iteratively imaging the source on the source-room walls, similar to the image method [22]. However, due to the source and receiver residing in different rooms, simply imaging the source creates artifact reflections that do not exist in reality. To eliminate these artifacts, StIM tests whether the line between the imaged source and the original receiver goes through the source room. Sources that do not satisfy this criterion are artifacts, thus being eliminated. For non-artifact sources, StIM proceeds and creates images of the receiver on the receiver-room walls. Artifact receiver images are also created and need to be eliminated. The criterion for receiver elimination is whether the line between the current-iteration source (imaged or original) and the current iteration's imaged receiver goes through the original receiver room. The attenuation and delay of the signal are calculated as a function of the reflection order and traveling distance. The resulting RIR is a superposition of the delayed and attenuated reflections.

Building on this generator, we wish to create an SER training dataset that follows a multiroom use case. For this purpose, we are using a set of 1000 generated multiroom RIRs. Each RIR represents an audio scene, where the source and receiver reside in two adjacent rooms. The dimensions of the coupled rooms are randomised, where the height, width, and depth of each room (H, W, D) are constrained (in meters) by $2 \leq H \leq 6$, $1.5 \leq D \leq 4$, $1.5 \leq W \leq 4$. For simplicity, we assume an alignment of the ceiling and floor of both rooms (meaning they are of the same height). We also align a single wall between both rooms. These alignment assumptions are a common and reasonable case in indoor environments. The locations of the receiver and source inside their respective rooms are also randomized. All RIRs are of length 4096 samples. For each sample in the dataset, we randomize k RIRs. The sample is then convolved with each of these k RIRs to represent how it reacts with respect to k acoustical scenarios of traveling between rooms. Thus, we create k new samples for each original sample. We consider this a k fold augmentation of the original data.

V. Experimental Results

We start by training all three models without augmentation. We used 20% of the clean data as an evaluation test-set. The performances of all three models are presented in Table I. The GRU-RNN model seems to perform slightly better than other architectures with compatible results. Such minor differences may be caused by neglectable factors such as weight initialization or similar training parameters. Thus, all three models can be considered to perform equally.

We proceed with evaluating these models using our real data. Note that since the real data is balanced, the scores converge for balanced and unbalanced accuracy. The results, presented in Table II, show significant performance degradation for all three models. As evident by the confusion matrices presented in Figure 1, all three models predict mostly the same

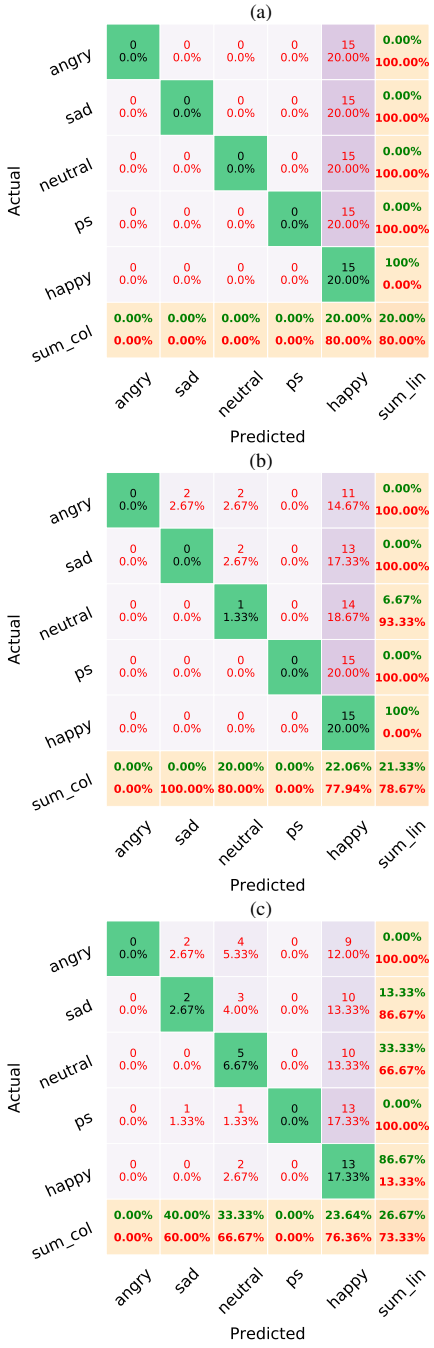

Fig. 1: Confusion matrices of classifiers trained without augmentation on real data. (a) AlexNet, (b) RNN with LSTM, and (c) RNN with GRU. All models mostly guess a single label.

single label for all the test samples, reducing the performances to a level of guesses and not classifications. Some of the acoustic features like pitch can be altered by the transition through a wall while other features are invariant such as speaking rate. These features could influence SER performances [23]. The GRU-RNN model seems slightly better in this task, achieving

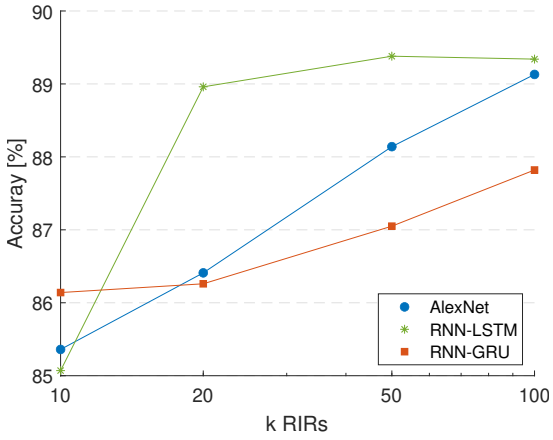

Fig. 2: Performance on real data with respect to k.

a small number of correct labels for more than one class, but still failing in general. A possible explanation is that the GRU exposes all of its memory to other units in the network. This memory may contain the required information for classification, which is late due to the delays. The LSTM, which controls the exposure of memory may be preventing this information from flowing to other units. Training LSTM with a compatible simulation will force the unit to learn an exposure amount of memory that will be adequate for this specific use-case.

As evident by the low performance of the multiroom SER task, an augmentation method is needed. Therefore we proceed to train all three models using $k = [10, 20, 50, 100]$ augmentation folds. All models are trained using the three datasets. We also use a test set from these datasets for evaluation purposes. However, we focus our discussion on the performances on the real-recorded evaluation set. The results on the test set, with respect to k, are presented in Figure 2.

The performances of each of the models trained with augmentation are given in Table III. The results show that all three architectures greatly benefit from integrating the augmentation into the dataset. It seems that the best results are achieved using $k = 50$ folds of augmentation, and are saturated for higher values of k. We note that in [15], the value of k was chosen when the evaluation test-set reached performance saturation. Given our real-recorded samples, we find the value of k with respect to saturation on the real-recorded set. However, such a method is not always possible and depends on the availability of enough real-recorded samples.

As evident from Table III, even augmentation by a factor of $k = 10$ greatly improves the performance on the real-recorded set. The sequential models show better results over the CNN for all k values, while the best performing model is the LSTM-RNN. This could imply that the sequential characteristics of the data in the SER task from an adjacent room have more impact on the performance than the inter-frequency features. This is reasonable, as reverberations are a product of delaying and summing the original signal. Such

TABLE III: Performances on the real data. The models were trained with augmentation.

Architecture	k	UA	BA	F1
AlexNet	10	33.33%	33.33%	23.45%
AlexNet	20	38.66%	38.66%	35.79%
AlexNet	50	50.66%	50.66%	43%
AlexNet	100	49.33%	49.33%	47.73%
LSTM-RNN	10	42.66%	42.66%	37.68%
LSTM-RNN	20	57.33%	57.33%	54.45%
LSTM-RNN	50	**61.33**%	**61.33**%	58.23%
LSTM-RNN	100	60%	60%	**58.8**%
GRU-RNN	10	44.44%	44.44%	38.84%
GRU-RNN	20	46.66%	46.66%	40.01%
GRU-RNN	50	52%	52%	42.74%
GRU-RNN	100	53.33%	53.33%	50.12%

a process has a sequential nature, with a varying window of time delay, depending on the room shape: large rooms will have long delays, while in small rooms, the receiver will experience incidents by more frequent reflections. The RNN architecture is capable of learning the optimal sequence length to be observed. In the case of SER in a multiroom audio scene, the sequence length significantly affects the overall performance.

VI. Conclusions

We have presented the first evaluation of an SER task between coupled rooms. The results of our study show that current models cannot identify the emotional state of a speaker from another room without augmentation during the training phase. In light of the findings, it is crucial to use generators capable of simulating situations as close as possible to the target situation. For the case of audio transition between rooms, StIM provides a suitable simulation method. The performance improvement is not limited to any specific architecture, and the results show that the StIM improves both CNN and RNN performances. With StIM, existing datasets can be used without recording a new dataset specifically for the multiroom case. Moreover, the low computational complexity of the system allows parameters to be controlled quickly and easily, enabling different experimental scenarios without the necessity of actual recordings. LSTM-RNN achieved the best performance for SER from another room when trained using StIM augmentation with $k = 50$ rooms. Many other architectures can be evaluated, such as TDNN, ResNets, and GCN. Some of these may perform better in specific cases, such as polyphonic audio. It may be useful to pursue future work in multiroom SER using methods other than deep learning, such as HMMs and SVMs, and evaluate the performance improvement provided by StIM augmentation.

References

[1] M. Czerwinski, J. Hernandez, and D. McDuff, "Building an AI that feels: AI systems with emotional intelligence could learn faster and be more helpful," *IEEE Spectrum*, vol. 58, no. 5, pp. 32–38, 2021.

[2] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[3] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *Proc. 18th Internat. Conf. on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 1136–1139.

[4] J. Kim, "Bimodal emotion recognition using speech and physiological changes," *Robust speech recognition and understanding*, vol. 265, p. 280, 2007.

[5] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP).*, vol. 2. IEEE, 2003, pp. II–1.

[7] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[8] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proc. Internat. Conf. on Electronic & Mechanical Engineering and Information Technology*, vol. 2. IEEE, 2011, pp. 621–625.

[9] Y. Gao, J. Liu, L. Wang, and J. Dang, "Domain-adversarial autoencoder with attention based feature level fusion for speech emotion recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 6314–6318.

[10] A. Shirian and T. Guha, "Compact graph architecture for speech emotion recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 6284–6288.

[11] S. T. Rajamani, K. T. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2021, pp. 6294–6298.

[12] S. Kwon *et al.*, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, 2020.

[13] B. Boserup, M. McKenney, and A. Elkbuli, "Alarming trends in US domestic violence during the COVID-19 pandemic," *The American Journal of Emergency Medicine*, vol. 38, no. 12, pp. 2753–2755, 2020.

[14] M. D. Griffiths and M. A. Mamun, "COVID-19 suicidal behavior among couples and suicide pacts: Case study evidence from press reports," *Psychiatry research*, vol. 289, no. 113105, 2020.

[15] E. Shalev, I. Cohen, and D. Lvov, "Indoors audio classification with structure image method for simulating multi-room acoustics," *The Journal of the Acoustical Society of America*, vol. 150, no. 4, pp. 3059–3073, 2021.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th European Conference on Speech Communication and Technology*, 2005.

[17] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[18] 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.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[20] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[23] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, 2013.