A DATA-DRIVEN APPROACH FOR ACOUSTIC PARAMETER SIMILARITY ESTIMATION OF SPEECH RECORDING

Mattia Papa, Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Stefano Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

ABSTRACT

Speech audio acquisitions exhibit different quality and reverberation properties depending on the recording setup and environment. For this reason, it is expected that speech analysis systems that work correctly on certain audio recordings may fail on others acquired in different acoustic contexts. Therefore, to be able to tell whether a track under analysis shares the same acoustic characteristics of a reference one may be useful to understand if it can be successfully processed by a given speech analysis system. Alternatively, in a forensic scenario, an estimate of acoustic parameter similarity between two tracks can be used to verify whether the recordings have been likely acquired in the same environment or not. In this work, we propose two methods to estimate acoustic parameter similarity between a speech recording under analysis and a reference one. The first method relies on the estimation of channel-based acoustic indicators that are then compared to extract a similarity measure. The second method directly learns a parameter similarity measure through siamese neural networks.

Index Terms— Acoustic similarity, siamese neural networks, reverberation time, clarity index

1. INTRODUCTION

In recent years, due to the widespread diffusion of portable devices with audio recording capabilities, speech and audio analysis has received significant attention and has been used for multiple applications, like speech recognition in voice user interfaces. The input of these systems is often a speech signal acquired with a single microphone in environments with unpredictable acoustic characteristics [1]. Depending on the recording context, noise level and reverberation behaviour may change drastically, thus the ability to monitor the acoustic characteristics of the environment is crucial for the effectiveness of speech analysis systems [2, 3]. For this reason, the possibility to evaluate acoustic parameter similarity between a reference audio recording and a track under analysis is interesting in different contexts.

As an example, a method that estimates acoustic similarity can be embedded in data-driven robust speech recognition systems to improve their performances. To overcome the problem of mismatch between clean training data and noisy in-the-wild speech signals, data augmentation techniques are often used to increase the robustness of the systems [4, 5]. Assessing the similarity between real-world audio signal under analysis and a reference track from the training / evaluation set can help in the interpretation of potential errors and in re-defining the training data. Estimating acoustic parameter similarity can also help in a preliminary phase to select the most suitable speech analysis system or parameter tuning depending on the acoustic context [6].

Acoustic similarity estimation can help in facing challenges encountered in audio forensics as well. A forensic investigator often verifies not only the content and the speaker identity of a speech audio evidence, but also the environment in which the recordings has taken place [7]. The analysis of the similarity between the audio track under analysis and a reference one allows to verify the match between the claimed environment and the actual one in which the recording has been performed. Moreover, audio evidences can be maliciously manipulated applying splicing, i.e., concatenating multiple segments from different audio tracks [8]. If the acoustic recording conditions of the spliced segments are different, the analysis of acoustic parameter similarity can highlight inconsistencies and localize splicing points [9].

In this work we propose two data-driven methods to assess acoustic parameter similarity between two single-channel speech audio recordings.

The first method employs a Convolutional Neural Network (CNN) that maps a time-frequency representation of the two inputs to five different acoustic indicators: Signal-to-Noise Ratio (SNR), reverberation time (T_{60}), Direct-to-Reverberant Ratio (DRR), and two different clarity indices (C_{50} and C_{80}) [10]. Acoustic similarity is then defined as the euclidean distances between the parameters estimated from the reference signal and the signal under analysis.

The problem of non-intrusive acoustic parameter estimation from monaural speech signals has been explored in the audio analysis research community. In [11] sub-band decomposition is combined with statistical analysis to extract T_{60} and DRR directly from speech signal. In [12], a deep learning approach is used for reverberation time approximation. In [13] the authors estimate C_{50}, proved to be highly correlated to performances of phoneme recognition systems, using short-term features and decision tree learning. Another popular strategy is to jointly estimate several different parameters with a single estimator, rather than just one. In [6] a set of features based on Gabor filters is fed to a multi-layer perceptron network to estimate both T_{60} and DRR. In [14] a set of frame-based features are extracted and fed to a Recurrent Neural Network (RNN) to model the temporal correlation between features and outputs, i.e., DRR and T_{60}. Recently, the authors of [15] proposed to jointly estimate three different acoustic indicators (i.e., T_{60}, SNR and DRR) using a CNN trained on simulated reverberant speech samples. The authors assert that approximating multiple outputs helps data-driven approaches to be robust in noisy conditions. The results are interesting and inspired the multi-task learning methodology adopted in this work.

The second data-driven method we propose addresses directly the problem of acoustic parameter similarity estimation, without explicitly extracting acoustic parameters indicators. In this case a metric learning approach is tested adopting a siamese architecture.

The two methods are evaluated on a large dataset of simulated Room Impulse Responses (RIRs) convolved with speech signals cor-

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

rupted by noise. Results show that the second method outperforms the first one in estimating acoustic similarity between the two audio inputs. However, the first method provides more interpretable responses in terms of acoustic parameters.

2. BACKGROUND

In this section we introduce the signal model adopted in this work and the definition of a set of channel-based acoustic indicators used in the proposed method.

Signal Model. Let us consider a sampled audio signal $x(n)$ acquired with sampling frequency F_s in a reverberant and noisy environment with a single microphone. We can express $x(n)$ as

$$x(n) = s(n) * h(n) + w(n), \tag{1}$$

where $s(n)$ is the source signal, $h(n)$ is the RIR between the source and the receiver, and $w(n)$ is an additive background noise term. In this work the source $s(n)$ is assumed to be a speech signal produced by a source randomly positioned in the considered room. Also the position of the used microphone is randomly selected. With these definitions at hand, the SNR between the main source $s(n)$ and the additive noise $w(n)$ can be written as

$$\mathrm{SNR} = 10 \log_{10} \frac{\sum_n x(n)^2}{\sum_n w(n)^2}. \tag{2}$$

Acoustic Indicators. The method proposed in this work makes use of a set of acoustic descriptors to estimate similarity among different recording environments. This set includes three different channel-based objective measures that depend on the RIR of the considered setup and that aim at describing the reverberation behaviour [10]. In the following we list the definitions of the considered parameters.

- The T_{60} is defined as the time in seconds the energy decay curve (i.e., the tail integral of the squared RIR) takes to drop by 60 dB [10].
- The DRR measures the ratio between the energy contained in the direct arrival and that of the rest of the reverberant tail. It is defined as [10]

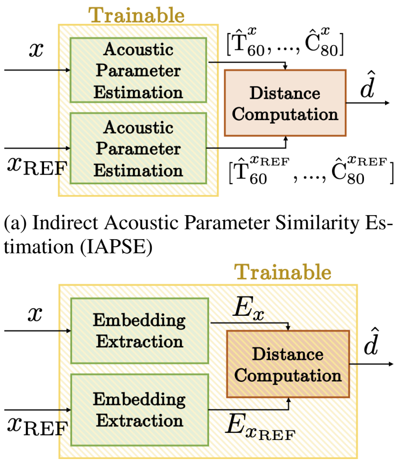
$$\mathrm{DRR} = 10 \log_{10} \frac{\sum_{n=0}^{n=n_d} h(n)^2}{\sum_{n=n_d+1}^{\infty} h(n)^2}, \tag{3}$$

  where the samples $h(n)$ of the impulse response from index 0 up to n_d correspond to the direct source-receiver propagation, whereas samples from $n_d + 1$ correspond to the reflection paths. Usually n_d corresponds to a time interval of 10 ms starting from the arrival time of the direct sound.

- The Clarity Index (CI) is a compact descriptor that can be linked to the way speech signal can be well perceived and understood. It is defined as [10]

$$\mathrm{CI} = 10 \log_{10} \frac{\sum_{n=0}^{n=n_e} h(n)^2}{\sum_{n=n_e+1}^{\infty} h(n)^2}, \tag{4}$$

  where n_e/F_s usually corresponds circa to either 50 ms or 80 ms, leading to two different indices, C_{50} and C_{80}, respectively.


Trainable
x
Acoustic Parameter Estimation
$[\hat{\mathrm{T}}_{60}^{x}, \ldots, \hat{\mathrm{C}}_{80}^{x}]$
Distance Computation
$\hat{d}$
x_{REF}
Acoustic Parameter Estimation
$[\hat{\mathrm{T}}_{60}^{x_{\mathrm{REF}}}, \ldots, \hat{\mathrm{C}}_{80}^{x_{\mathrm{REF}}}]$

(a) Indirect Acoustic Parameter Similarity Estimation (IAPSE)

Trainable
x
Embedding Extraction
E_x
Distance Computation
$\hat{d}$
x_{REF}
Embedding Extraction
$E_{x_{\mathrm{REF}}}$

(b) Direct Acoustic Parameter Similarity Estimation (DAPSE)

Fig. 1: (a) IAPSE pipeline: first acoustic parameters are estimated, then acoustic parameter similarity is computed. (b) DAPSE pipeline: similarity is directly estimated bypassing acoustic parameters.

3. PROPOSED METHOD

In this work we propose a method to blindly estimate acoustic parameter similarity between a reference audio signal and a signal under analysis by means of a distance measure. For each acoustic parameter, the distance measure is defined to have low values if the considered signals have been recorded in two environments with similar values for the considered acoustic indicator, e.g., similar reverberation behaviour or noise level, high values otherwise.

Formally, let us consider an audio speech signal under analysis x associated to the acoustic parameters $\mathrm{SNR}^x, \mathrm{T}_{60}^x, \mathrm{DRR}^x, \mathrm{C}_{50}^x$ and C_{80}^x. Let us consider a reference signal x_{REF} associated to the acoustic parameters $\mathrm{SNR}^{x_{\mathrm{REF}}}, \mathrm{T}_{60}^{x_{\mathrm{REF}}}, \mathrm{DRR}^{x_{\mathrm{REF}}}, \mathrm{C}_{50}^{x_{\mathrm{REF}}}$ and $\mathrm{C}_{80}^{x_{\mathrm{REF}}}$. The goal of our work is to propose a method that takes x and x_{REF} as input, and returns the euclidean distance d between each pair of acoustic parameters (i.e., $d_{\mathrm{T}_{60}} = \sqrt{(\mathrm{T}_{60}^x - \mathrm{T}_{60}^{x_{\mathrm{REF}}})^2}$ if T_{60} is considered).

To do so, we explore two possible data driven-strategies that are detailed in the following.

Indirect Acoustic Parameter Similarity Estimation (IAPSE). Figure 1a shows the pipeline of the first method, named Indirect Acoustic Parameter Similarity Estimation (IAPSE). This method first estimates the acoustic parameters, and then estimates similarity based on them. Each input (i.e., x. and x_{REF}) is separately processed by the acoustic parameter estimation block consisting of a CNN that estimates acoustic parameters. Parameters are then compared in the distance computation block that returns the estimated euclidean distances $\hat{d}$ for each parameter. Notice that in this method, only the acoustic parameters estimation block is data-driven, thus trainable.

The acoustic parameter estimation block predicts the set of acoustic parameters associated to its input signal. Following the approach proposed in [15], the parameters are jointly estimated using a CNN fed with a time-frequency representation of the input. In particular, considering the input x, we compute its log-mel spectrogram. To do so, Short-Time Fourier Transform (STFT) is applied to x, its magnitude is integrated over mel-spaced bins and a logarithmic function is applied to the magnitude of each bin as explained in [16]. This transformation produces a 2D representation

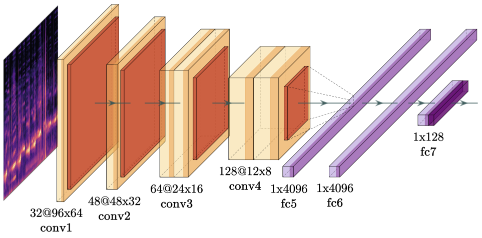


Fig. 2: VGGish architecture for embedding extraction.

of the input $X(m,k)$ where the index k indicates the frequency bin while the index m indicates the correspondent time window. This processed input is fed to a CNN. We decided to use the popular VGGish network [16], whose architecture is reported in Figure 2 and detailed in [16]. Additionally, a dense layer of five neurons is concatenated to the final VGGish layer to estimate the five parameters $\hat{\text{SNR}}^x, \hat{\text{T}}_{60}^x, \hat{\text{DRR}}^x, \hat{\text{C}}_{50}^x$ and $\hat{\text{C}}_{80}^x$. The loss function used for training the model is the Mean Squared Error (MSE) between the predicted and ground-truth parameters. The same process is applied independently to the second input x_{REF}.

Once the network is trained, the distance computation block takes all estimated acoustic parameters as input, and returns the euclidean distance $\hat{d}$ for each parameters pair. For instance, if T_{60} is considered, we obtain $\hat{d}_{\text{T}_{60}} = \sqrt{(\hat{\text{T}}_{60}^x - \hat{\text{T}}_{60}^{x_{\text{REF}}})^2}$. The same applies to the other parameters.

Direct Acoustic Parameter Similarity Estimation (DAPSE). Figure 1b reports the pipeline for the second method, named Direct Acoustic Parameter Similarity Estimation (DAPSE). This method directly estimates similarity bypassing acoustic parameters. Differently from IAPSE method, the inputs x and x_{REF} are jointly processed by a siamese CNN [17], which extracts audio embeddings and learns the desired distance measure $\hat{d}$ for the considered parameter. This is an end-to-end trainable method that is completely data-driven compared to IAPSE approach. It focuses on learning a distance measure on one single acoustic parameter at a time, rather than learning jointly an estimate for all parameters and then separately computing their distances.

The siamese CNN is composed of two twin networks that share the weights, jointly process two distinct inputs (i.e., x and x_{ref}) and are joined at the top by a specific function. In our scenario, each one of the twin networks is a VGGish model [16] presented in Figure 2. Each VGGish acts as an embedding extractor that maps the inputs x and x_{ref} into the embedding vectors E_x and $E_{x_{\text{REF}}}$, respectively. Embeddings are fed to a layer that computes the euclidean distance $\hat{d}$ between them. As the entire system is trained at once, the loss function is built such that the network minimizes the difference between the learnt distance $\hat{d}$ and the ground truth distance d by means of MSE.

It is worth noticing that in DAPSE method the configuration of the embedding space and the subsequent euclidean distance is learnt through training, while in IAPSE the learning aims only at estimating directly the acoustic parameters. Moreover, to the best of our knowledge, the choice of deep metric learning in acoustic parameter similarity estimation is a novel aspect of this work.

4. EXPERIMENTAL SETUP

In this section we present the dataset and methodologies used for evaluating the proposed techniques.

Dataset. For evaluating the proposed method we created a dataset ad-hoc. This consists of several speech signals corrupted by noise and convolved with RIRs obtained simulating a large amount of rooms with different acoustic properties.

Following the signal model introduced in Section 2, we used the TIMIT dataset [18], consisting of 6300 different utterances from different speakers, as clean speech signals $s(n)$ sampled with $Fs = 16000$ Hz. As additive noise $w(n)$, we considered both white noise and babble noise. The considered SNR levels are $\text{SNR} \in \{10, 15, 25, 35\}$ dB. The RIRs $h(n)$ have been simulated using Pyroomacoustic toolbox [19], allowing the direct control of reverberation parameters. In particular, we defined a set of shoebox rooms with volumes spanning between 27 m^3 and 256 m^3 and T_{60} values between 200 ms and 1200 ms. From the simulated RIRs we computed C_{50}, C_{80} and DRR values following (3) and (4). We obtained $\text{DRR} \in [-21.67, 15.37]$ dB, $\text{C}_{50} \in [-12.40, 20.66]$ dB and $\text{C}_{80} \in [-8.66, 25.59]$ dB. The total number of room configurations is approximately 100. The final dataset considering all speeches, noises and RIRs counts 104000 tracks, which is suitable for our data driven approach.

Training and metrics. The CNNs proposed in Section 3 are trained using the proposed dataset. The training set is composed of 90000 audio tracks while the test set is composed of 14000 audio tracks. To ensure the generalization capability of the networks, we divided training and test sets such that the subset of rooms considered during the training phase is disjointed from the subset used for testing.

For logmelspectrogram computation we considered a window of 0.96 s for each track and STFT is applied using Hanning window of length $N_w = 0.025$ s and hop size $N_h = 0.010$ s. The magnitude of the result is mapped in mel scale, using 64 bins spanning from $F_{min} = 125$ Hz up to $F_{max} = 7500$ Hz. Finally the natural logarithm function is applied. The final 2D matrix has dimension 96x64 samples, as required by VGGish.

To improve the overall performances of the proposed system, the actual training consists in fine tuning the original VGGish network pre-trained for audio classification task on a very large dataset [20]. For both configurations, the selected optimizer is Adam with learning rate set to 0.001. Training has been performed for 100 epochs using the early-stopping mechanism with patience 10 (i.e., if validation loss does not improve for 10 epochs, the training is stopped and the best validation model is saved).

To evaluate the proposed systems we choose as evaluation metric the Pearson correlation coefficient ρ as in [15]. This is used to compare estimated acoustic parameters or acoustic similarity distribution against the ground truth.

5. RESULTS

In this section we present the achieved results. First, we analyse the performances of the acoustic parameter estimation block of IAPSE method. Then, we present the results on the task of acoustic parameter similarity estimation obtained using IAPSE and DAPSE methods.

Acoustic Parameter Estimation. These experiments validate the use of VGGish to estimate five acoustic parameters jointly in the acoustic parameter estimation block of IAPSE.

First, we compared the proposed acoustic parameter estimation method exploiting VGGish against a baseline. As baseline, we use an extended version of the work presented in [15] adapted to jointly estimate the five acoustic parameters used in this work rather than the three parameters used in [15]. To this purpose, Table 1 shows a comparison between the baseline and the proposed acoustic parame-

Table 1: ρ values for the proposed acoustic parameter estimation method and the baseline.

	DRR	T60	SNR	C50	C80
Proposed	0.854	0.932	0.986	0.899	0.894
Baseline	0.841	0.923	0.979	0.901	0.912

Table 2: ρ values using single-output and multi-output acoustic parameter estimation configurations.

	DRR	T60	SNR	C50	C80
Multi-output	0.857	0.932	0.990	0.900	0.896
Single-output	0.856	0.923	0.992	0.895	0.889

ter estimation method. The results are expressed in terms of ρ computed between the predicted acoustic parameters and the ground-truth values, for each one of the five parameters. We can observe that performances are satisfactory for all the estimated parameters. Some parameters, like DRR and C_{50} are more challenging for both networks, while the noise level is often easily predicted, reaching almost $\rho = 0.98$. The proposed method has on average a slight improvement over the baseline, which motivates us in keeping VGGish as backbone for the proposed system.

As second experiment, we investigated the effect of jointly estimating all five parameters compared to the estimate of each parameter separately. In Table 2 we report the prediction results for each parameter estimated with two different configurations. In the first one, indicated as single-output, each parameter is estimated separately with a specific network trained for one output. In the second configuration, indicated with multi-output, all parameters are jointly estimated from a single network as proposed in Section 3. As shown in [15], this strategy allows to exploit the information shared between the different acoustic parameters and helps the learning phase in achieving higher generalization capacity. The experiment confirms that the multi-output strategy improves over the single-output one, apart for SNR estimation, for which there are no sensible differences.

Finally, we wanted to evaluate the robustness of the method to different noise conditions. To this purpose, in Figure 3 we present the results for the multi-output proposed architecture for different SNRs. As expected, ρ values decrease for lower SNR values, but also in the worst scenario the values are acceptable and the system is effective.

Acoustic Parameter Similarity Estimation. In this section we compare the IAPSE and DAPSE approaches. In the first experiment we compare the performances between the two methods in terms of ρ between predicted and real distance for each acoustic indicator. For IAPSE approach, we test a single network performing a joint estimation of parameters, for maximizing prediction accuracy. About DAPSE method, one siamese network is trained for each parameter. As evident from Table 3, the distance estimation is more accurate using DAPSE strategy, i.e., when the network is specifically trained for the task of distance estimation rather than for parameter estimation. We believe that the use of deep metric learning and siamese configuration helps in learning a meaningful embedding space, strictly related to the considered acoustic properties. Two audio tracks that corresponds to acoustically similar environments correspond to two close points in the embedding space. For this reason, distance prediction reaches higher accuracy. On the other side, IAPSE system provides an intermediate direct estimation of all the acoustic indicators, that can be easily interpreted by an analyst.

Table 3: ρ values for IAPSE and DAPSE methods.

	DRR	T60	SNR	C50	C80
IAPSE	0.706	0.837	0.964	0.785	0.784
DAPSE	0.735	0.871	0.979	0.834	0.840

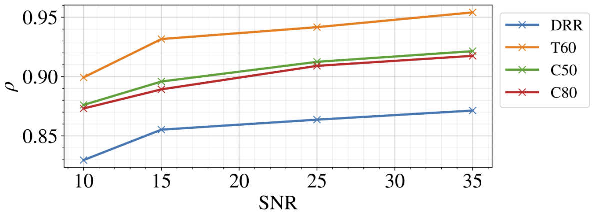


Fig. 3: ρ values for proposed Acoustic Parameter Estimation method varying SNR values.

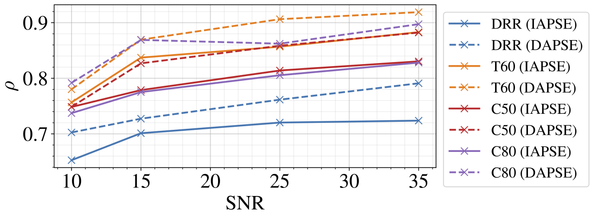


Fig. 4: ρ values for IAPSE and DAPSE methods for different SNR values.

In Figure 4, we present the performances of the two systems showing ρ for different SNR values. We can observe that, for both methods, the approximation accuracy increases for higher SNR values. We can also observe that DAPSE method outperforms IAPSE one for any SNR value, showing a good prediction robustness even when dealing with noisy recordings.

6. CONCLUSIONS

In this paper we presented two data-driven methods to estimate acoustic parameter similarity between two speech recordings. Both methods are based on CNN architectures fed with a time-frequency representation of the audio signal. The first method preliminary estimates a set of acoustic parameters on which a distance measure is defined. The second method learns an embedding space where the distance is highly correlated to the acoustic parameter similarity. We evaluated both methods using a large dataset of reverberant noisy speech signals. The second method outperforms the first one in terms of Pearson correlation coefficient in the acoustic parameter similarity estimation task. The first method, on the other side, reaches performances comparable with the state of the art in the approximation of the acoustic parameter. Moreover, it offers an easier interpretation of the acoustic conditions of the analysed setups. As possible future works we plan to investigate the use of different networks for the embedding extraction phase and test the method on data generated with real RIR.

7. REFERENCES

[1] Patrick A. Naylor and Nikolay D. Gaubitch, "Acoustic signal processing in noise: It's not getting any quieter," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–6.

[2] Armin Sehr, Emanuël AP Habets, Roland Maas, and Walter Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2010.

[3] Takahiro Fukumori, Masanori Morise, and Takanobu Nishiura, "Performance estimation of reverberant speech recognition based on reverberant criteria rsr-dn with acoustic parameters," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2010.

[4] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[5] Mary Harper, "The automatic speech recogition in reverberant environments (aspire) challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 547–554.

[6] Feifei Xiong, Stefan Goetze, Birger Kollmeier, and Bernd T Meyer, "Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 255–267, 2018.

[7] Alastair H. Moore, Mike Brookes, and Patrick A. Naylor, "Roomprints for forensic audio applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[8] Hong Zhao, Yifan Chen, Rui Wang, and Hafiz Malik, "Audio splicing detection and localization using environmental signature," *Multimedia Tools and Applications*, vol. 76, no. 12, pp. 13897–13927, 2017.

[9] Davide Capoferri, Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Speech audio splicing detection and localization exploiting reverberation cues," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020, pp. 1–6.

[10] Patrick A Naylor and Nikolay D Gaubitch, *Speech dereverberation*, Springer Science & Business Media, 2010.

[11] Thiago de M. Prego, Amaro A. de Lima, Rafael Zambrano-López, and Sergio L. Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015, pp. 1–5.

[12] Hannes Gamper and Ivan J. Tashev, "Blind reverberation time estimation using a convolutional neural network," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 136–140.

[13] Pablo Peso Parada, Dushyant Sharma, and Patrick A. Naylor, "Non-intrusive estimation of the level of reverberation in speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4718–4722.

[14] Pablo Peso Parada, Dushyant Sharma, Toon van Waterschoot, and Patrick A Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ace challenge," *Proceedings of the ACE Challenge Workshop - a satellite event of IEEE-WASPAA*, 2015.

[15] David Looney and Nikolay D Gaubitch, "Joint estimation of acoustic parameters from single-microphone speech observations," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 431–435.

[16] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "Cnn architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[17] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015, vol. 2.

[18] John Garofolo, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallett, Nancy Dahlgren, and Victor Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," 1993.

[19] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 351–355.

[20] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.