Predicting the Perceived Level of Reverberation using Features from Nonlinear Auditory Model

Saeid Safavi, Wenwu Wang, and Mark Plumbley
University of Surrey
Guildford, United Kingdom
{s.safavi, w.wang, m.plumbley}@surrey.ac.uk

Ali Janalizadeh Choobbasti
Amirkabir University of Technology,
Tehran, Iran
alijanalizadeh@aut.ac.ir

George Fazekas
Queen Mary University of London,
London, United Kingdom
g.fazekas@qmul.ac.uk

***Abstract*—Perceptual measurements have typically been recognized as the most reliable measurements in assessing perceived levels of reverberation. In this paper, a combination of blind RT60 estimation method and a binaural, nonlinear auditory model is employed to derive signal-based measures (features) that are then utilized in predicting the perceived level of reverberation. Such measures lack the excess of effort necessary for calculating perceptual measures; not to mention the variations in either stimuli or assessors that may cause such measures to be statistically insignificant. As a result, the automatic extraction of objective measurements that can be applied to predict the perceived level of reverberation become of vital significance. Consequently, this work is aimed at discovering measurements such as clarity, reverberance, and RT60 which can automatically be derived directly from audio data. These measurements along with labels from human listening tests are then forwarded to a machine learning system seeking to build a model to estimate the perceived level of reverberation, which is labeled by an expert, autonomously. The data has been labeled by an expert human listener for a unilateral set of files from arbitrary audio source types. By examining the results, it can be observed that the automatically extracted features can aid in estimating the perceptual rates.**

I. Introduction

After close to a century of research, methods of acquiring and utilizing various attributes of room acoustics properties using different room acoustic measures is still a discussed topic. One of these aspects most recognized by end-users of sound effect repositories is the perceived level of reverberation. It would be of great value to these users if they were able to narrow their search results towards audio files with the desired amount of reverberation. Another application of automatic characteristics of reverberation time would be their utilization in music information retrieval tasks. [10] illustrates how the accuracy of automatic musical instrument recognition (MIR) models are affected by the amount of reverberation present. By inferring the level of reverberation of sounds as a priori, one could simply train a unique MIR model for different levels of reverberation. As a contribution, this paper suggests a new approach towards deriving the perceived level of reverberation directly from the recorded audio files.

This paper proposes a fresh manner of deriving auditory parameters, that have been proved to be relevant for the overall perception of acoustic quality. Most of the feature extraction methods that have been proposed in the context of characterizing reverberations up to now are based on the usage of room impulse response as an input. However, in this paper, all the approaches rely on the use of recorded audio files as an input. The parameters considered include four of the most significant attributes of auditory perception listed in the ISO 3382-1 standard [3]: Reverberance, Listener envelopment (i.e., LEV, the feeling encompassed by sound), apparent source width (ASW), and the clarity. These parameters, along with a number of other spatial features, including the level of both foreground and background streams, the interaural time differences (ITD) present in these streams, and the level of the low-frequency part of the spectrum (LLS) are estimated by a binaural, nonlinear auditory model [1]. Section II-B will cover further details on applying the binaural auditory model for feature extraction and the properties of such features.

Another parameter of considerable importance in characterizing the quality of acoustic space is the reverberation time (RT) [1] [6]. The model presented in [5] can predict the reverberation time directly from an audio signal. Combined with the features gathered from the auditory model, these features can be used to predict the perceived level of reverberation directly from the recorded audio signals, eliminating the need for the room impulse response.

This paper starts by describing the process in which the necessary features for predicting reverberations are extracted in the section II. In the section III, the setup in which the experiments were performed and the models applied for predicting the level of reverberations are described. Section IV defines the experimental setup used in this research. Finally, section V concludes this paper.

II. Perceptual Attributes and Acoustic Parameters

In the previous section, it was pointed out that this work aims to create a model that can predict the perceived level of reverberation given the raw recorded audio signal. The succeeding subsections elaborate on the models used to extract the necessary features, along with details on their corresponding attributes, that are required to be later fed into machine learning approaches.

A. Sound Decay Model

To make an effective estimate of the reverberation time directly from the audio signals an algorithm utilizing the Laplacian distribution based energy decay model has been proposed by [5].

The reverberant audio signal is first divided into a number of overlapping frames [11]. These frames are then preselected in order to identify any possible sound decays. The preselection process involves splitting each frame into many subframes, and examining whether the maximum or minimum energy values of each sub-frame deviates from its consecutive subframes according to [11]. If such a deviation is observed in a consecutive sequence of subframes, they are identified and marked as a possible sound decay. The detected frames are then used for calculating the reverberation time, to create a finite number of RT values.

A histogram with a fixed bin size of 10 containing the estimated RT rates is created in order to improve the estimation veracity. This histogram is updated with the inclusion of each additional RT value calculated. Since there are no significant number of outliers present in this histogram [14] due to the preselection, at every given time, the current RT estimate is then associated with the maximum value present in this histogram, instead of the first peak. The variance of the estimated RT value is then reduced via recursive smoothing.

B. The Binaural Auditory Model

A variety of different auditory models have been developed to imitate the human binaural auditory system. The binaural auditory model utilized in this research is a variation of the better known binaural model titled as the Room Acoustic Analyzer (RAA) and has been fully detailed out in [9]. A block diagram displaying an overview of this system is shown in Fig. 1. The model consists of a peripheral processor that is first applied separately to the left and right ear channels, followed by a central processing module.

In order to create an effective model of the human auditory system, one must take into account the non-linearity of the human auditory system [13]; it must accurately model the temporal and spectral masking [16] [18], and the binaural interactions made in the human auditory system. A model that can encompass all of the foregoing features is the binaural model proposed by [24] [26], which is a binaural extension of the monaural model proposed by [16]. This model has been further expanded by [4] in order to predict content specific measures aspects of room acoustic perception.

1) The peripheral processor: This stage imitates the outer and middle ear, the hair cells, neural firing, and the basilar membrane residing in the cochlea. As shown in Fig. 1, there are separate modules and processes carried out for each ear channels. To create a nonlinear binaural model, the input signals are first scaled to the correct level, so that an SPL of 0 dB resembles a root mean square (RMS) value of 1.
To begin, outer and middle ear filtering, which has been developed as a second-order band-pass IIR filter with cutoff frequencies between 1 and 4 kHz, is then applied; a fourth order gammatone (critical-band) filter bank consisting of 41 frequency bands with center frequencies ranging from 27 to 20—577 Hz is then applied to simulate the basilar membrane inside the cochlea [19]. To simulate phase locking at higher frequencies and to preserve the signal envelope, the signals are then half-wave rectified and then passed through a fifth order low-pass filter with a cut-off frequency of 770 Hz. According to the absolute threshold of hearing (ATH) curve from [21], a lower limit is then incorporated into the signals that are dependant on the center frequency of each band. Values that are below this frequency dependent threshold are set to zero. The reason behind this phase is the incorporation of the ATH. Adaptation loops are then applied in order to imitate the adaptive properties of the auditory periphery [16]. Neurons maintaining the human auditory system by transmitting electrical signals in the brain, adjust this sensitivity to the input they receive. The output is then smoothed so that a stationary input of 100 dB SPL for the mid-frequency range yields a steady state output of 100 model units. This is while silence at the input (i.e., input of 0 dB SPL) produces a steady state output of 0 model units.

2) The binaural processor: To simulate the binaural interaction in the human auditory system, an equalization-cancellation approach has been proposed in [26]. This approach applies so called excitation-inhibition (EI)-type elements, each with a characteristic ITD and interaural level difference (ILD), and incorporates a finite binaural temporal resolution to the left and right ear monaural model outputs. The produced output of the EI-type elements together shapes a pattern of the EI activity as a function of the characteristic ITD and ILD. Based on [20], it can be deduced that the frequency range 125-1000 Hz is prevalent concerning the perception of spaciousness. It was later found by [12] that ITD is the dominant localization cue for this frequency range.

3) The central processor: As the final stage, the central processor takes the output of the binaural processor (ITD values), along with both of the monaural outputs as its input. Note that a low-pass filter with a time constant of 20 ms is applied to the monaural stage outputs to extract the envelope as described by [16]. The human auditory system splits an input stream into a direct foreground stream, which corresponds with the input source, and a reverberant background stream, which corresponds with the environment (noise) around the source. The nonlinear behavior of this model is meant to mimic this behavior and perform the splitting.

Four of the auditory parameters used for predicting the perceived reverberances (i.e., reverberance, clarity, apparent source width, listener envelopment) are produced by the central processor. Combined with the previous stages, the following features can be derived from the binaural auditory model:

Level of foreground stream (LFS): An auditory parameters closely associated with the sound source is the level of the foreground (i.e., source) stream (LFS). To calculate this parameter, the mean level of the monaural output streams are

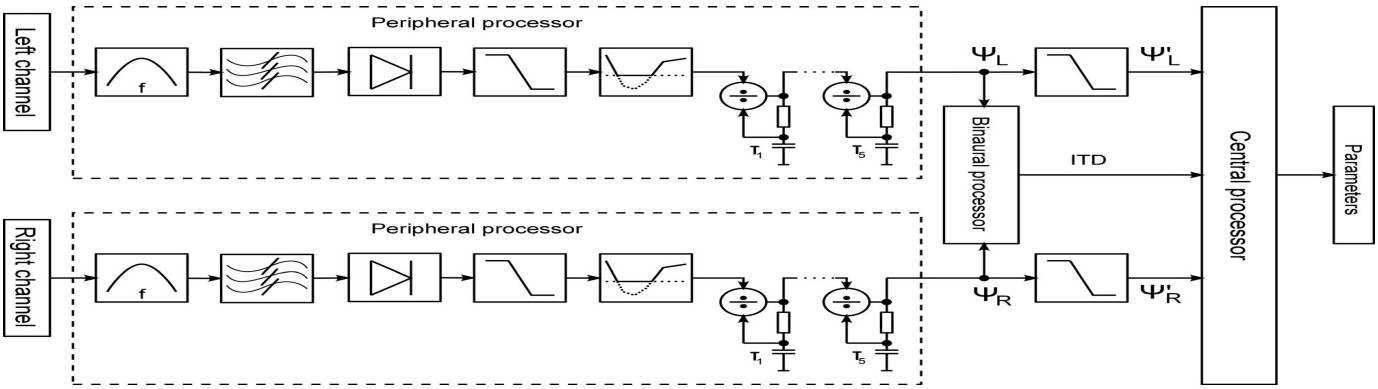

Fig. 1. Schematic of the binaural auditory model [4]. The full model consists of five adaptation loops, two of which are shown above (with τ_1 and τ_5 being their constants) [1].

measured.

Level of background stream (LBS): The acoustic parameter associated with the reverberance is the level of the background stream. Much like the LFS, this parameter can be estimated using the monaural outputs of the model. To measure the level of the background stream (LBS) the mean level of the reverberant sound stream output by the monaural model is calculated.

ITD fluctuation in the foreground (ITD_f): The interaural time differences present in the output of the binaural processor are split into two streams. The foreground stream that is closely related to the sound source is used in calculating the ITD_f. The mean standard deviation of the foreground stream is used to calculate ITD_f.

ITD fluctuation in the background (ITD_b): Another binaural parameter closely related to the reverberance is the fluctuations present in the background stream (ITD_b). Similar to the ITD_f, ITD_b is calculated via measuring the mean standard deviation of the background stream output by the binaural processor.

Level of the low-frequency part of the spectrum (LLS): Another factor relating to both the perceived reverberance and the apparent source width explained below is the absolute sound pressure levels present in lower frequency bands. By taking the mean of the output of gammatone filters applied to lower frequency bands in the peripheral processor, the level of the low-frequency part of the spectrum (LLS) can be calculated.

Reverberance (REV): The intensity of the reverberations perceived by listeners which is commonly regarded as being closely related to the physical reverberation time, is what is formally known as reverberance; in other words, the amount of time required from the moment the sound source stops until the sound pressure level deteriorates by 60 dB. As proposed by [4], a valid approach towards evaluating reverberance is to take the average level of the reverberant sound stream (i.e., LBS), which can be calculated using the outputs of the monaural processing units of the model.

Clarity: Another important aspect of sound is the extent to which discrete sounds in a signal stay distinct from each other subjectively in relation to time. The higher the clarity rate, the easier it is to recognize separate phonemes in a audio or to identify individual notes residing in a musical piece. To calculate the perceived clarity, the proportion of the average direct sound stream levels (LFS) with respect to the mean reverberant sound level (LBS), which are calculated from the peripheral processing unit outputs, is measured [4].

Apparent Source Width (ASW): One of the two most significant aspects of auditory spaciousness paired with listener envelopment is the discernable increase of a sound source with respect to early lateral reflections. ASW is frequently determined by the early interaural cross-correlation [23] [22]. As mentioned earlier in II-B2, fluctuations in both ILD and ITD with respect to time, create the notion of spaciousness, with ITD being the more dominant cue. Since the model outputs ITD as a function of time, it can be utilized in obtaining a parameter related to ASW.

Moreover, [25] demonstrated how the perceived source width is not only dependant on the interaural decorrelations but also relates to the absolute sound pressure level present in lower frequencies (i.e., the level of the low-frequency part of the spectrum).

Therefore, the perceived ASW can be calculated from the model by incorporating the output of the binaural processor ITD_f, and the level in the lower bands LLS [4].

Listener Envelopment (LEV): One could denote a sound field as an enveloping one when a feeling of being encompassed by the sound transpires. As mentioned above, the second critical perceptual parameter determining spaciousness that relates to the environment in contrast to the source is the LEV.

LEV includes two important elements: The interaural cross-correlation (i.e., the spacious aspect of the sound), and the level in the diffuse part of the impulse response (the absolute late SPL). A blind prediction of the LEV, which is closely related to the auditory impression, can be made since this concept is associated with the binaural and monaural model outputs. The mean level of the background stream (i.e.,

TABLE I
PREDICTION ACCURACY ACHIEVED IN TERMS OF CORRECTLY IDENTIFIED INSTANCES.

Features	Setup	Number of classes	Logistic Regression	Decision Tree	MLP
RT60, LFS, LBS, ITD_f, ITD_b, LLS, REV	Between source type	2 reverberation classes	67.75 %	72.75 %	**75.25 %**
RT60	Between source type	2 reverberation classes	–	–	63.75 %
LFS	Between source type	2 reverberation classes	–	–	54.00 %
LBS	Between source type	2 reverberation classes	–	–	49.50 %
ITD_f	Between source type	2 reverberation classes	–	–	50.25 %
ITD_b	Between source type	2 reverberation classes	–	–	51.25 %
LLS	Between source type	2 reverberation classes	–	–	57.00 %
REV	Between source type	2 reverberation classes	–	–	58.75 %

LFS), and the ITD fluctuations in the background stream can be utilized in making this prediction.

III. EXPERIMENTAL SETUP

Evaluations are performed using 400 audio files downloaded from freesound. Due to the small number of audio samples available and in order to make use of all the files in both the training and testing stages, a popular resampling strategy called three fold cross validation has been employed for the evaluation. Within each increment, one fold is reserved as the test set, and the two other folds are used for training the model. This process is repeated three times so that every fold is utilized in both training and testing. This problem has been modeled as a binary classification problem, with two classes named as a low and high class, which contains recordings with the low and high perceived level of reverberations, respectively. Three distinct approaches have been employed in addressing this problem. The models used are Multinomial Logistic Regression, Decision Tree, and Multilayer Perceptron [2] [8] [7]. The features applied in each experiment are listed in Table I. Each sample has been labeled as to either pertaining high perceived level of reverberation or low by an expert auditor.

A. Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a widely known machine learning approach for classifying a set of features that belong to one of two (or more) possible classes. Given some samples with distinct feature vectors, a set of functions is then constructed. Each function holds the probability of a feature vector belonging to each class. The class with the highest odds is selected as the predicted label for the specific feature vector. Each function has a number of parameters that are calculated during the training phase, where a number of sample feature vectors, along with their known labels are presented. These parameters adapted in such a way that whenever encountering a new set of features that are similar to a set of features seen previously, the function will output a high probability that the newly observed features belong to the same class as the similar feature vectors observed during the training process. Given n sample, m features, and k classes, the parameter matrix W is an $m * (k-1)$ matrix. The probability that each observation X_i belongs to each class j, except for the final class, is equal to:

$$P_j(X_i) = \frac{exp(X_i \cdot B_j)}{((\sum_{j=1}^{k-1} exp(X_i \cdot B_j)) + 1)}$$

The odds that the sample X_i belongs to the last class is equal to:

$$1 - (\sum_{j=1}^{k-1} P_j(X_i))$$

B. Decision Tree

Among the different decision tree (DT) algorithms, the C4.5 algorithm, which is a successor to J. Ross Quinlan's ID3, is probably the most popular ones in the DT family that are used in the machine learning community.
The approach taken in decision trees is creating a tree data structure in a recursive manner, in order to partition a data set into sub-divisions based on a number of tests that are defined at each node. The final tree consists of a root node, which serves as an entry point for every sample, internal nodes, that define the splits, and leaf nodes that represent the observations. Decision trees are particularly good at establishing the nonlinear relationships between feature vectors and their corresponding classes [17], and deriving content specific measures of room acoustic perception using binaural, nonlinear auditory model.

C. Multilayer Perceptron

The multilayer perceptron (MLP) uses an algorithm proposed by [15] named backpropagation in its learning procedure that helps make predictions in machine learning tasks. The network consists of multiple logistic units that act together in learning abstract representations of the input feature vectors in the middle layers. There are nonlinear activation units between each layer that help in modeling real-world phenomenon by introducing nonlinearity into the model. The ReLU function is probably the most commonly used activation function presently available, due to its robustness against the vanishing gradients phenomenon. For further details on the implementation and workflow of the MLP, please refer to [15].

IV. EXPERIMENTAL RESULTS

Let us now compare the results obtained from the different experimental setups. The results obtained from these experiments are summarized and displayed in Table I. These

results are comprised of the performance of different models on different sets of features, in predicting the perceived level of reverberation for the input audio signals.

Initially, seven of the features described above have been incorporated in predicting the perceived level of reverberation. This experiment also holds the highest level of accuracy observed in such experiments, as expected initially. Table I shows that the highest performance is obtained when using the multilayer perceptron classifier along with the seven features, resulting in an accuracy of **75.25**%.

Due to the low number of samples, it was assumed that the more simple machine learning models (i.e., logistic regression, decision tree) would outperform the MLP, given that the MLP typically requires a high number of samples to converge. But surprisingly MLP performs very well even with relatively small amount of available data for training the model.

In order to discover the most significant extracted feature, the experiment was repeated for each feature using the model with the highest accuracy (i.e., the MLP). The results of these experiments are shown in Table I. As evident in the results, the feature with the highest effect in predicting the perceived level of reverberation is the reverberation time, with an accuracy of **63.75**% followed by the room reverberance (REV) with an accuracy of **58.75%**.

V. Conclusion

After decades of research, a robust method of estimating the perceived level of reverberation from raw audio signals is yet to be discovered. Different approaches in extracting quantified measures which can be utilized in tackling this problem and coming up with specific measurable reverberation features from audio signals have been presented over the years (e.g., calculating reverberation time, the direct to reverberation ratio). Each of these features has had their own shortcomings in relating to the perceived level of reverberation.

In this paper, a new approach has been proposed towards addressing this problem. To be able to blindly extract features from the raw audio signals, a sound decay model and a binaural auditory model have been applied. The extracted features are then used in training various machine learning algorithms. The obtained results are promising and suggest that the features extracted using both feature extraction models can be applied in predicting the perceptual level of reverberation.

VI. Acknowledgment

This work was partially supported by the H2020 Project entitled AudioCommons funded by the European Commission with Grand Agreement number 688382.

References

[1] Osses Vecchi A., Kohlrausch A., Lachenmayr W., Mommertz E. Predicting the perceived reverberation in different room acoustic environments using a binaural auditory model. The Journal of the Acoustical Society of America. 2017 Apr;141(4):EL381-7.

[2] Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[3] ISO A. Measurement of room acoustic parameterspart 1. ISO Std. 2009.

[4] van Dorp Schuitman, J., de Vries, D. and Lindau, A. Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model. The Journal of the Acoustical Society of America, 133(3), 2013, pp. 1572-1585.

[5] Jan T, Wang W. Blind reverberation time estimation based on Laplace distribution. In the Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 2050-2054.

[6] Safavi, S., Pearce, A., Wang, W., Plumbley, M., Predicting the perceived level of reverberation using machine learning, Asilomar Conference on Signals, Systems, & Computers, 2018.

[7] Safavi, S., Gan, H. and Mporas, I., 2017, March. Improving speaker verification performance under spoofing attacks by fusion of different operational modes. In IEEE 13th International Colloquium on Signal Processing and its Applications (CSPA), 2017, pp. 219-223

[8] Safavi, S., Gan, H., Mporas, I. and Sotudeh, R., December. Fraud detection in voice-based identity authentication applications and services. In Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on, 2016, pp. 1074-1081.

[9] van Dorp Schuitman, J., Auditory modelling for assessing room acoustics, 2011.

[10] Barthet M, Sandler M. On the effect of reverberation on musical instrument automatic recognition. InAudio Engineering Society Convention 128 2010 May 1. Audio Engineering Society.

[11] Lllmann, H., Yilmaz, E., Jeub, M. and Vary, P., An improved algorithm for blind reverberation time estimation. In Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC), 2010, August, (pp. 1-4).

[12] Griesinger, D., October. Room impression, reverberance, and warmth in rooms and halls. In Audio Engineering Society Convention 93. Audio Engineering Society, 1992.

[13] Griesinger, D., The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces. Acta Acustica united with Acustica, 1997, 83(4), pp. 721-731.

[14] Ratnam, R., Jones, D.L., Wheeler, B.C., OBrien Jr, W.D., Lansing, C.R. and Feng, A.S., Blind estimation of reverberation time. The Journal of the Acoustical Society of America, 114(5), 2003, pp. 2877-2892.

[15] Rumelhart, D.E., Hinton, G.E. and Williams, R.J., Learning representations by back-propagating errors. nature, 1986, 323(6088), p. 533.

[16] Dau, T., Pschel, D. and Kohlrausch, A., A quantitative model of the effectivesignal processing in the auditory system. I. Model structure. The Journal of the Acoustical Society of America, 1996, 99(6), pp. 3615-3622.

[17] Friedl, M.A. and Brodley, C.E., Decision tree classification of land cover from remotely sensed data. Remote sensing of environment, 1997, 61(3), pp. 399-409.

[18] Lehmann, P. and Wilkens, H., Zusammenhang subjektiver Beurteilungen von Konzertslen mit raumakustischen Kriterien. Acta Acustica united with Acustica, 1980, 45(4), pp. 256-268.

[19] Patterson, R.D., Robinson, K.E.N., Holdsworth, J., McKeown, D., Zhang, C. and Allerhand, M., Complex sounds and auditory images. In Auditory physiology and perception, 1992, pp. 429-446.

[20] Barron, M. and Marshall, A.H., Spatial impression due to early lateral reflections in concert halls: the derivation of a physical measure. Journal of sound and Vibration, 1981, 77(2), pp. 211-232.

[21] Terhardt, E., Calculating virtual pitch. Hearing research, 1979, 1(2), pp. 155-182.

[22] Schroeder, M.R., Gottlob, D. and Siebrasse, K.F., Comparative study of European concert halls: correlation of subjective preference with geometric and acoustic parameters. The Journal of the Acoustical Society of America, 1974, 56(4), pp. 1195-1201.

[23] Bradley, J.S. and Soulodre, G.A., The influence of late arriving energy on spatial impression. The Journal of the Acoustical Society of America, 1995, 97(4), pp. 2263-2271.

[24] Breebaart, J., Modeling binaural signal detection. Technische Universiteit Eindhoven, 2001.

[25] Okano, T., Beranek, L.L. and Hidaka, T., Relations among interaural cross-correlation coefficient (IACC E), lateral fraction (LF E), and apparent source width (ASW) in concert halls. The Journal of the Acoustical Society of America, 1998, 104(1), pp. 255-265.

[26] Breebaart, J., Van De Par, S. and Kohlrausch, A., Binaural processing model based on contralateral inhibition. I. Model structure. The Journal of the Acoustical Society of America, 2001, 110(2), pp. 1074-1088.