Predicting the perceived level of reverberation using machine learning

Saeid Safavi, Andy Pearce, Wenwu Wang, and Mark Plumbley
centre vision speech signal processing (CVSSP)
University of Surrey
UK, GU2 7XH.
Email: {S.safavi, Andy.pearce, W.wang, M.plumbley}@surrey.ac.uk

Abstract—**Perceptual measures are usually considered more reliable than instrumental measures for evaluating the perceived level of reverberation. However, such measures are time consuming and expensive, and, due to variations in stimuli or assessors, the resulting data is not always statistically significant. Therefore, an (objective) measure of the perceived level of reverberation becomes desirable. In this paper, we develop a new method to predict the level of reverberation from audio signals by relating the perceptual listening test results with those obtained from a machine learned model. More specifically, we compare the use of a multiple stimuli test for within and between class architectures to evaluate the perceived level of reverberation. An expert set of 16 human listeners rated the perceived level of reverberation for a same set of files from different audio source types. We then train a machine learning model using the training data gathered for the same set of files and a variety of reverberation related features extracted from the data such as reverberation time, and direct to reverberation ratio. The results suggest that the machine learned model offers an accurate prediction of the perceptual scores.**

Index Terms—**Reverberation time, Human subject test, Machine learning, MLP.**

I. Introduction

Perceived level of reverberation was identified as an important perceptual attribute that users of sound effect repositories search for frequently [1]. Being able to limit search results to sound effects that have no apparent reverberation would be of great benefit to users. Yet even after decades of research, models for predicting the apparent reverberation have yet to be developed.

Research has shown that the perceived level of reverberation depends considerably on the source signal and the shape of the reverberation decay [2]. Various studies, including an IEEE challenge, have resulted in different methods for extracting specific measurable reverberation features from audio signal, such as the reverberation time and direct-to-reverberant ratio [3], [4], [5], [6]. However, these measures do not always directly relate to the perceived level of reverberation. For example the reverberation time (RT) is an important parameter for characterizing the quality of an auditory space. Sounds in reverberant environments are subject to coloration. This affects speech intelligibility and sound localization. Many state-of-the-art audio signal processing algorithms, for example in Automatic Speech Recognition (ASR), speaker recognition [7], [8], [9], hearing-aids and telephony, are expected to have the ability to characterize the listening environment, and turn on an appropriate processing strategy accordingly [10].

This paper proposes a new method for predicting the perceived level of reverberation from audio files using machine learning approaches.

II. Human Perception

Human ratings of reverberance were collected to develop a model of perceived reverberation. Two types of ratings were collected in listening tests: (i) within-source, making direct comparisons of a single source type; and (ii) between-sources, comparisons across multiple different sources.

A. Stimuli selection

For the within-source evaluation, stimuli were selected that are likely to benefit from a model of reverberation. Source types were selected by examining a 1-month search history of freesound [1], identifying source types that were commonly searched for along with the terms *reverberant*, *dry*, or *dead* (common terms relating to reverberation [1]). This analysis leads to the identification of five commonly searched source types: snare, atmosphere, thunder, hit, and vocal.

A keyword search was performed using the freesound API. Each source type was searched in isolation, as well as with the reverberant, dry, and dead additional search terms (e.g. searching for 'snare', 'reverberent snare', 'dry snare', and 'dead snare'). For each search, 50 random sound effects were downloaded and converted to wav files with 44.1kHz sample rate.

A manual filtering was conducted, removing sounds not of the desired source type. Five sound effects were selected from each search term, and presented to an independent expert, who selected five sounds from each source type that demonstrated a range of apparent reverberation.

For the between-sources evaluation, the same stimuli selection method was conducted for five other timbral attributes of hardness, depth, brightness, roughness, and metallic-nature; five sources selected for each attribute, each source with five stimuli. After a pilot study, the median rated stimuli rated for each attribute/source combination was selected for the between-sources experiment; a total of 30 stimuli.

[1] https://freesound.org/

Each stimulus was loudness matched using the Nugen Audio LM Correct to -35.2 LUFS, the lowest loudness level of all normalised stimuli.

B. Listening tests

All listening tests were conducted in an acoustically treated editing room, using Neumann KH120A active studio monitors. The playback system was aligned to produce a level of 74 dB_{SPL} at the listening position with -14 dBFS pink noise. This produced a comfortable listening level when the stimuli were reproduced.

Both listening tests were conducted with a multiple stimulus comparison test interface, presenting a number of stimuli simultaneously, allowing participants to audition each stimulus as many times as desired, rate each stimulus on a scale from 0 to 100, and rearrange the ordering of the stimuli into ascending order based on their ratings.

For classification purposes and for labeling the audio files based on the human rating, each of the scores converted to three-class and two-class reverberation. For two-class reverberation setup any audio files which were scored from 0 to 50 are labeled as a file with low level of reverberation and the ones scored from 51 to 100 are labelled as a file with the high level of reverberation. Similarly for the three-class reverberation setup files with the score from 0 to 30 are labelled as low, 31 to 60 as medium and 61 to 100 as high.

1) Within-source ratings: For the within-source listening tests, each page contained all five stimuli of a single source type. Listeners were instructed to rate the relative perceived level of reverberation, using the full range of the scale in each page.

Sixteen listeners completed this listening test, all of whom were undergraduate students on the Tonmeister Sound Recording course at the University of Surrey, all having technical ear training and experience in listening tests.

2) Between-sources ratings: Prior to listening tests, an independent expert was asked to identify the most and least reverberent stimuli for use as hidden anchors.

Before each test, participants were presented all 30 stimuli on a single familiarisation interface to become accustomed to the range of reverberation. Each test page comprised nine sliders, two of which were the hidden anchors, the other seven being a randomised order for the remaining stimuli. Participants were ask to make ratings of perceived reverberation relative to the full range heard during the familiarisation stage.

III. Experimental setup

A. Feature extraction

Following features have been extracted form each of the audio files and per each channel.

- Reverberation time (RT60)
- Direct to reverberation ratio (DRR)
- Early decay time (EDT)
- Early to late index (CTE)

Each of the recordings was analyzed individually. Figure 1 shows the impulse responses and the decay curves for a sounds in our database.

The function that we have used to extract these four features from audio file is based on the usage of reverse cumulative trapezoidal integration to estimate the decay curve.

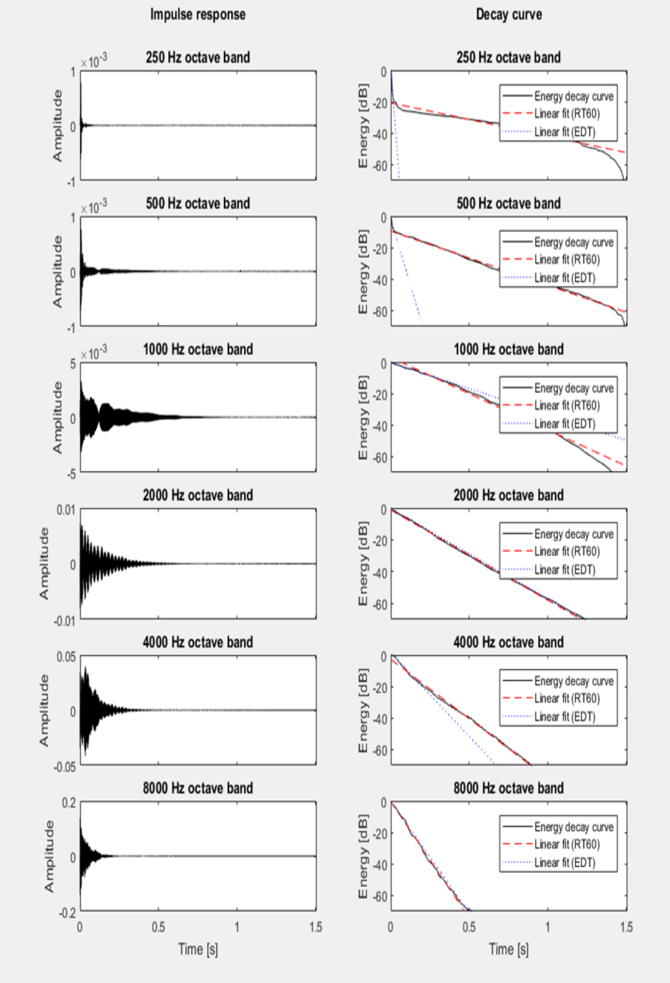

Fig. 1. Example of impulse response and the decay curve for the audio file with the low level of reverberation.

1) Reverberation time: The function which was used to estimate the reverberation time from the audio files uses reverse cumulative trapezoidal integration to estimate the decay curve, and a linear least-square fit to estimate the slope between 0 dB and -60 dB. Estimates are taken in octave bands and the overall figure is an average of the 500 Hz and 1 kHz bands. The function determines the direct sound as the peak of the squared audio wave files.

2) Direct to reverberation ratio: DRR is calculated in the following way:

$$DRR = 10 * log10(X(T_0 - C : T_0 + C)^2 / X(T_0 + C + 1 : end)^2)$$

where X is the approximated integral of the impulse, T_0 is the time of the direct impulse, and $C = 2.5ms$. The direct-

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

Features	Setup	Number of classes	Logistic Classifier	Decision Tree	MLP
RT, DRR, CTE, EDT	Within source type	3 reverberation classes	60.00 %	**68.00 %**	56.00 %
RT, DRR, CTE, EDT	Within source type	2 reverberation classes	64.00 %	**76.00 %**	64.00 %
RT, DRR, CTE, EDT	Between source type	3 reverberation classes	50.00 %	46.66 %	**60.00 %**
RT, DRR, CTE, EDT	Between source type	2 reverberation classes	66.66 %	73.33 %	**83.33 %**
RT	Between source type	2 reverberation classes	– %	– %	76.67 %
DRR	Between source type	2 reverberation classes	– %	– %	73.33 %
CTE	Between source type	2 reverberation classes	– %	– %	63.33 %
EDT	Between source type	2 reverberation classes	– %	– %	66.67 %

to-reverberant energy ratio cue, DRR, results primarily from the diffuse reverberant sound-field present in environments with sound reflecting surfaces. This sound-field is a collection of, perhaps, thousands of complex reflections: degraded, delayed, and attenuated copies of the original waveform. As source distance increases, reverberant energy remains roughly constant, although direct-path energy decreases by 6 dB per doubling of source distance; hence, the direct-to-reverberant energy ratio, DRR, decreases. The precise amount that DRR changes with distance depends critically on the amount of reverberant energy present, which is determined by properties of the acoustic environment. For room environments, reverberant energy as a function of time is determined principally by the size of the room and the acoustic properties of the reflecting surfaces of the room. Many outdoor environments also produce reverberation, therefore a direct-to-reverberant energy ratio cue varies with distance [11]. In simpler but related acoustic situations involving a sound source with a single simple reflection, or echo, detection thresholds for the echo are known to be stimulus dependent. The lowest thresholds result from brief, impulsive signals [12] and higher thresholds result from longer duration signals with slow onsets [13].

3) Early to late index: Early reflections arriving whitin 50ms after the direct sound are not perceived separately but are rather integrated for directional cues. A measure to characterize a reverbrant room situation with respect to speech intelligibility is the early to late energy ratio. CTE is the same size as RT. This is calculated in the following way: $CTE = 10 * log10(X(T_0 - C : T_0 + T_E)^2 / X(T_0 + T_E + 1 : end)^2)$ where TE is $50ms$.

4) Early decay time: EDT is also the same size as the RT. The slope of the decay curve is determined from the fit between 0 and -10 dB. The decay time is calculated from the slope as the time required for a 60 dB decay.

B. Modeling methods

For classification purposes we have used multiple functions provided by the Weka toolkit [2]. We have tried logistic classifier, decision tree, and multilayer perceptron (MLP) [14], [15], [16], [17].

All the machine learning approaches use the same set of features, which are: RT, DRR, EDT and CTE extracted from audio files and each channel. For the training part the class labels are obtained from the human experiments. Each audio file is scored by 16 human listeners and the class label for each file is the Median of scores provided by the listeners.

Because of the availability of small number of samples per each class we have carried out 3 folds cross validation setup. Per each round two folds have been used for training and the third fold kept for testing, and this setup repeated three times. So every instance has been used for both training and testing.

1) Logistic classifier: For k classes for n instances and m attributes, the parameter matrix B to be calculated will be an $m * (k - 1)$ matrix. The probability for class j with the exception of the last class is:

$$P_j(X_i) = exp(X_i B_j) / ((\sum_{j=1}^{(k-1)} exp(X_i * B_j)) + 1) \qquad (1)$$

The last class has the probability of:

$$1 - (\sum_{j=1}^{(k-1)} P_j(X_i)) \qquad (2)$$

2) Decision tree: Among decision tree algorithms, J. Ross Quinlan's ID3 and its successor, C4.5, are probably the most popular in the machine learning community [18]. In this research C4.5 algorithm is used.

Decision tree algorithms begin with a set of examples and create a tree data structure that can be used to classify new examples. Each example is described by a set of features which can have numeric or symbolic values. Associated with each training case is a label representing the name of a class. Each internal node of a decision tree contains a test, the result of which is used to decide what branch to follow from that node. The leaf nodes contain class labels instead of tests. In classification mode, when a test example (which has no label) reaches a leaf node, C4.5 classifies it using the label stored there [19].

3) Multilayer perceptron: Weka uses a classifier that relies on backpropagation to learn a MLP to classify instances. The network can be built by hand or set up using a simple heuristic. The network parameters can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric, in which case the output

[2] https://www.cs.waikato.ac.nz/ml/weka/

nodes become unthresholded linear units). For full details about the implimentation of MLP and how to use it in WEKA please refer to [14].

IV. Experimental results

In this section obtained results from different setups and using different number of classes have been compared. Table I summarizes the performance of three machine learning approaches in automatic prediction of a perceived level of reverberation from audio sounds. Machine learning approaches which have been used are logistic classifier, decision tree and MLP. Column 1 in the Table I shows what features/feature were used as an input. Initially all four extracted features are used for different experimental setups.

Table I shows that the best performances for within and between source types architectures, when using four extracted features as an input, are 76.00 % and 83.33 %, when using two reverberation classes (dry and high reverberation), and by Decision tree and MLP algorithms, respectively.

We expected to get better performance when using the within source type architecture compared with the between source type. This finding is in agreement with our initial expectation as based on the human listening test it is a simpler task for human listeners as well. The MLP method outperform the decision tree and logistic classifier for between source type experiments. This can be due to the fact that MLP needs large amount of data for training and for the between source type experiments we had larger number of audio files available for training.

In order to find the most effective extracted feature we repeated the experiment with the best obtained performance, which was the between source type architecture, using two reverberation classes and MLP. The results of these experiments are placed in the last four column of the Table I. These experiments show the most effective stand alone feature is the reverberation time followed by direct to reverberation ratio, early to late index and early decay time.

V. Conclusion

The perceived level of reverberation has been a subject for many research. Different methods for extracting specific measurable reverberation features from audio signals have been proposed, e.g. different methods for calculating reverberation time and direct to reverberation ratio. None of these features on their own could always relate to the perceived level of reverberation.

In this paper, we develop a new method to predict the level of reverberation from audio signals by relating the perceptual listening test results with those obtained from a machine learned model. More specifically, we compare the use of a multiple stimuli test for within and between class architectures to evaluate the perceived level of reverberation from the humans opinion. We train a machine learning model using the training data gathered for the same set of files and a variety of reverberation related features extracted from the data, RT60, DRR, EDT and CTE. The best result of 83.33 % is obtained when using all four features as an input for between source type setup and 2 reverberation class.

Acknowledgment

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

References

[1] Andy Pearce, Tim Brookes, and Russell Mason. Timbral attributes for sound effect library searching. In *Audio Engineering Conference on Semantic Audio*, pages 629–633. IEEE, 2016.

[2] Jouni Paulus, Christian Uhle, and Jrgen Herre. Perceived level of late reverberation in speech and music. In *Audio Engineering Society Convention 130*, May 2011.

[3] Pavel Zahorik. Direct to reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5):2110–2117, 2002.

[4] Benjamin Cauchi, Hamza Javed, Timo Gerkmann, Simon Doclo, Stefan Goetze, and Patrick Naylor. Perceptual and instrumental evaluation of the perceived level of reverberation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 629–633. IEEE, 2016.

[5] Gaubitch N. Moore A. Eaton, J. and P. Naylor. Estimation of room acoustic parameters: The ace challenge. *IEEE Transaction on Audio, Speech and Language Processing*, 24(10):1681–1693, 2016.

[6] Lima A. Netto S. Lee B. Said A. Schafer R. Prego, T. and Kalker T. A blind algorithm for reverberation-time estimation using subband decomposition of speech signals. *The Journal of the Acoustical Society of America*, 131(4):2811–2816, 2016.

[7] Saeid Safavi. *Speaker characterization using adult and childrens speech*. PhD thesis, University of Birmingham, 2015.

[8] S. Safavi, A. Hanani, M. Russell, P. Jancovic, and M. J. Carey. Contrasting the effects of different frequency bands on speaker and accent identification. *IEEE Signal Processing Letters*, 19(12):829–832, Dec 2012.

[9] Saeid Safavi, Martin Russell, and Peter Jancovic. Automatic speaker, age-group and gender identification from children's speech. *Computer Speech & Language*, 50:141 – 156, 2018.

[10] Rama Ratnam, Douglas L. Jones, Bruce C. Wheeler, William D. OBrien, Charissa R. Lansing, and Albert S. Feng. Blind estimation of reverberation time. *The Journal of the Acoustical Society of America*, 114(5):2877–2892, 2003.

[11] Douglas G Richards and R Haven Wiley. Reverberations and amplitude fluctuations in the propagation of sound in a forest: implications for animal communication. *The American Naturalist*, 115(3):381–399, 1980.

[12] Pavel Zahorik. Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5):2110–2117, 2002.

[13] Earl D Schubert and Joel Wernick. Envelope versus microstructure in the fusion of dichotic signals. *The Journal of the Acoustical Society of America*, 45(6):1525–1531, 1969.

[14] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

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

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

[17] Saeid Safavi and Iosif Mporas. Improving performance of speaker identification systems using score level fusion of two modes of operation. In *International Conference on Speech and Computer*, pages 438–444. Springer, 2017.

[18] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[19] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.