DISCRIMINATIVE FEATURE DOMAINS FOR REVERBERANT ACOUSTIC ENVIRONMENTS

Constantinos Papayiannis, Christine Evers and Patrick A. Naylor

Department of Electrical and Electronic Engineering, Imperial College London, UK
{papayiannis, c.evers, p.naylor}@imperial.ac.uk

ABSTRACT

Several speech processing and audio data-mining applications rely on a description of the acoustic environment as a feature vector for classification. The discriminative properties of the feature domain play a crucial role in the effectiveness of these methods. In this work, we consider three environment identification tasks and the task of acoustic model selection for speech recognition. A set of acoustic parameters and Machine Learning algorithms for feature selection are used and an analysis is performed on the resulting feature domains for each task. In our experiments, a classification accuracy of 100% is achieved for the majority of tasks and the Word Error Rate is reduced by 20.73 percentage points for Automatic Speech Recognition when using the resulting domains. Experimental results indicate a significant dissimilarity in the parameter choices for the composition of the domains, which highlights the importance of the feature selection process for individual applications.

Index Terms— Feature Selection, Machine Learning, Environment Identification, Reverberant speech recognition.

1. INTRODUCTION

Reverberation is observed in rooms due to the reflection of sound waves as they meet surfaces such as objects or walls. It can be perceived as prolonging of the original sound due to temporal smearing and in most real-life cases as an audible change in timbre. In the case of music, reverberation is to some extent desirable since it provides warmth to the sound [1]. In the case of speech, it can degrade intelligibility [2] and also affects the performance of Automatic Speech Recognition (ASR) systems [3].

Complete knowledge of the reverberation effect is given by the Acoustic Impulse Response (AIR) between the source and the receiver [4]. AIRs however typically involve thousands of taps, which prohibits their direct use for several tasks such as classification due to computational and memory constraints associated with high dimensionality. The motivation for this work is to find low-dimensional descriptors for the acoustic environment that are discriminative with regards to the reverberation effect. The primary aim is, by considering a set of tasks, to provide insight into the discriminative properties of acoustic parameters that can be extracted from reverberant speech and used to form a low-dimensional feature domain.

Considering the range of potential applications, we are interested in parameters that can be extracted both from the AIR and/or can also be estimated from reverberant speech. Using knowledge about the choices of such parameter in the literature we investigate the use of Mel-frequency Cepstral Coefficients (MFCC) [5, 6], Reverberation Time (RT) [7] and Direct-to-Reverberant Ratio (DRR) which is related to clarity [8, 9, 2]. Despite the effectiveness of the parameter sets listed above, there is a lack of a comparative study of their discriminative properties when considering a collection of tasks. This study will allow us to identify whether a set of parameters can be considered as task-independently discriminative, or whether a task specific domain should be used for future applications.

Parameter values estimated from the received signals include uncertainty due to an estimation error. Therefore, in order to provide a clear insight into their discriminative powers, parameters extracted from AIRs are used in this work. They are evaluated as feature domain dimensions using feature selection methods for classification. A Classification and Regression Tree (CART) and a Support Vector Machine (SVM) classifier are used for a set of tasks involving environment identification and ASR acoustic model selection. Classification accuracy and Word Error Rate (WER) for ASR are used as metrics for the suitability of the resulting feature domains.

The structure of the remainder of this paper is as follows: Section 2 provides information about the formulation of the problem and the methods used for its solution, Section 3 describes the experimental setup used, Section 4 offers a discussion on the experimental results and the conclusion is provided in Section 5.

2. METHOD

2.1. Signal Model

For time index, n, the reverberant speech signal $x(n)$ can be modeled as a convolution process between the anechoic

speech signal $s(n)$ and the AIR from the sound source to the microphone $h(l)$ for $l = 0, 1, \ldots, L-1$. Additive noise is denoted as $\nu(n)$. In vector notation, we define $\mathbf{s} = [s(0), \ldots, s(N-1)]^T$, $\mathbf{h} = [h(0), \ldots, h(L-1)]^T$ and $\boldsymbol{\nu} = [\nu(0), \ldots, \nu(N+L-1)]^T$. Given vector $\mathbf{h}$, the convolution matrix $\mathbf{H}$ of dimensions $N+L-1 \times N$ can be formed as

$$\mathbf{H} = \begin{bmatrix} h(0) & 0 & \ldots & 0 \\ h(1) & h(0) & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ h(L-1) & \ldots & \vdots & 0 \\ 0 & h(L-1) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & h(L-1) \end{bmatrix} \quad (1)$$

and the reverberant speech signal as

$$\mathbf{x} = \mathbf{H}\mathbf{s} + \boldsymbol{\nu}. \quad (2)$$

2.2. Acoustic Features

Considering the case of no additive noise, the AIR is a description of a stationary acoustic environment. We form the rows of matrix $\mathbf{Y}$ by stacking M direct-path-aligned AIRs, $\mathbf{h}_m = [h(0), ..., h(L_m - 1)]$, $m \in \{1, \ldots, M\}$. We define the feature extraction operators $f_k(\mathbf{Y})$, where $k \in \{1, \ldots, K\}$. When applied to $\mathbf{Y}$, it transforms its rows to D_k elements, each one corresponding to an acoustic parameter. With D acoustic parameters under consideration and considering all possible parameter combinations indicates that $K = 2^D$. The transformation can be summarized as

$$f_k : \mathbb{R}^{L_m} \rightarrow \mathbb{R}^{D_k} \quad (3)$$

$$\mathbf{Y}_k = f_k(\mathbf{Y}), \quad (4)$$

where the dimensions of the matrix $\mathbf{Y}_k$ are $M \times D_k$.

As feature dimensions, the following acoustic parameters have been evaluated that are linked to different aspects of reverberation: full-band RT [2], frequency-dependent RT [2], full-band DRR [2], frequency-dependent DRR [2] and MFCC [10]. The parameters are respectively denoted as τ_f, $\tau(\omega_\xi)$, λ_f, $\lambda(\omega_\psi)$ and $\mu(\omega_\zeta)$, where ξ, ψ and ζ are sub-band indices and ω indicates the angular frequency of their geometric center. We extract the RT using [11] and following from the results of [7] we consider $^1/_3$ octave-bands for $\omega > 150\ \frac{2\pi}{f_s}$, where f_s is the sampling frequency in Hz. We extract the DRR also in $^1/_3$ octave-bands for the same frequency region, leading to $\xi \in \{1, \ldots, 16\}$ and $\psi \in \{1, \ldots, 16\}$. We extract 25 MFCC spanning the range 100 Hz and 8 kHz, hence $\zeta \in \{1, \ldots, 25\}$. In total $D = 59$ acoustic parameters are evaluated. For clarity and readability, in the following sections the sub-band indices ξ, ψ and ζ are indicated as subscripts of the parameter symbols.

2.3. Feature Selection

Feature selection is performed for each considered task as a wrapper around supervised training of classifiers, the method for which is discussed in this section. Each row of $\mathbf{Y}_k$ is a feature vector representation of the acoustic environment which is described by the corresponding row of $\mathbf{Y}$. Therefore for supervised training, M rows need to be labeled to indicate their ground-truth value relative to the task. The vector of ground-truth values can be formed as $\mathbf{c} = [c_1, \ldots, c_M]^T$. The classifier g_i is defined as the function

$$\hat{\mathbf{c}} = g_i(\mathbf{Y}_k). \quad (5)$$

With $\hat{\mathbf{c}}$ being an estimate of $\mathbf{c}$, the misclassification rate $E_{k,i}$ is defined as the proportion of misclassified feature vectors for classifier i operating in feature domain k.

Thus the objective is to identify the best pair (i, k) by considering K candidate acoustic parameter combinations (feature domains) and I classifiers using

$$\underset{k,i}{\operatorname{argmin}}\ E_{k,i}, \quad (6)$$

for $i \in \{1, \ldots, I\}$ and $k \in \{1, \ldots, K\}$.

In the experiments in this paper we consider $I = 2$ classifiers, a SVM with a Gaussian kernel [12] and a CART [13, 14] although this case be straightforwardly extended. For each one of the two cases, the minimization search along the k dimension of (6) is performed differently. This operation is effectively the feature selection process for each classifier. For the case of SVM, Backwards Sequential Selection (BSS) [15] is used and for the case of CART the feature selection is performed by the tree growing algorithm. The number of candidate feature domains in this work is $K = 2^D = 2^{59}$. Rather than exhaustively searching the vast solution space in terms of K, the two feature selection methods used in this work rely on certain assumptions to simplify the problem. For BSS, the stopping criterion chosen empirically for our experiments is set as the reduction of the feature dimensions to 15 parameters. For CART, the split criterion used is Gini's diversity index [16].

3. EXPERIMENTAL SETUP

Four experiments are described in this section with the objective being to identify the most discriminative feature domain for each of the tasks studied. The four tasks are:

1. Room-type identification, for which practical applications involve context recognition and environment identification for data-mining [17, 6].
2. Room identification for forensic applications [7, 18, 5].
3. Classification of reverberant environments with regards to quantized receiver positions. [19, 20, 21].

4. Acoustic model selection for ASR [8].

For brevity in the following sections we refer to the tasks using indices 1,2,3 and 4, following the above order.

3.1. Evaluation Method

Cross-validation [12] is used for the evaluation of the misclassification rate of (6). The cross-validation grouping is performed in terms of rooms for Tasks 1 and 4, in terms of receiver position for Task 2 and a leave-one-out approach is used for Task 3.

3.2. Environment Identification

The objective of Tasks 1 to 3 is to identify properties of reverberant speech inputs that relate to the enclosure. For these experiments, AIRs provided by the Acoustic Characterization of Environments (ACE) Challenge [22] are used. AIRs for 7 rooms are included. For the case of Task 1, the ACE database allows us to perform classification of types office, meeting room and lecture theater. For the case of Task 2, 2 offices, 2 meetings rooms, 2 lecture theaters and a building lobby are considered. For the case of Task 3, 70 microphone locations in a total of 7 rooms are considered, with the objective being to identify the room and position.

3.3. Acoustic Model Classification for ASR

Given a reverberant speech input for ASR, the objective of this task is to minimize the WER by choosing the appropriate acoustic model [8]. The experiment is performed using AIRs that have been recorded using the crucifix microphone array for the ACE Challenge. Therefore, two 5-channel AIR recordings for 7 rooms are available. All AIR channels are used for the corruption of test data and the center channel of each AIR is used for the training of the acoustic models. This gives in total 14 acoustic models and 70 ASR conditions. The training process for this experiment requires an extra step as the ground-truth value vector $\mathbf{c}$ (discussed in Section 2) is not readily available. To determine $\mathbf{c}$, ASR must be performed on all 70 conditions using all available models. We restrict the choice of acoustic models in the sense that for each condition, models trained using AIRs from the same room are excluded. Therefore, 12 possible models are available for each one of the 70 conditions. The vector $\mathbf{c}$ is formed by the models that provide the lowest WER for each one. Performing this step turns the problem in the same form as the remaining tasks considered.

For the training of acoustic models and running the experiment, the Kaldi Toolkit[1] and the TIMIT[2] speech database were used.

[1] http://kaldi-asr.org

[2] The s5 example for TIMIT was used as provided with the Kaldi Toolkit.

4. RESULTS

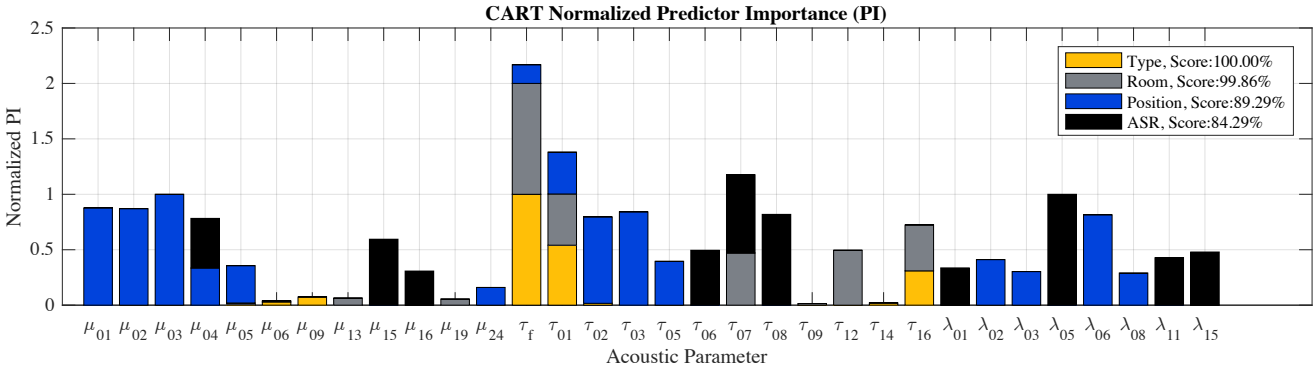
4.1. Feature Selection Results

The results of the experiments are shown in Fig. 1, which illustrates the choice of parameters both for the CART and SVM classifiers. The cross-validation accuracy for each one is shown as part of the legend. For the CART, in Fig. 1a we are able to visualize the level of importance of each parameter to the designed classifier as the Predictor Importance (PI) (also referred to as Attribute Importance) [14]. For the case of SVM with BSS, the selected features are shown as binary values on the polar plot of Fig. 1b, with a different radius for each task and the circle segmented to indicate the parameter sets. CART proved to be the best performing classifier for Task 1 with a cross-validation accuracy of 100%. The SVM gave the best results for Tasks 2, 3 and 4 with cross-validation accuracies respectively of 100%, 97.86% and 100%. Using (6) allows us to identify the feature domain Θ for each task as shown in Fig. 1c, with the domain subscript referring to the task index. The order of the acoustic parameters in this plot has been arranged to highlight the overlap and the differences between the choices for the acoustic parameters for each task.

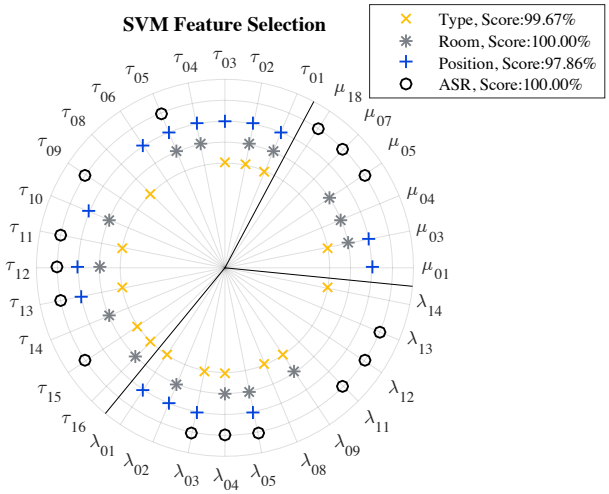
4.2. Discussion

Concerning CART, it can be observed from Fig. 1a that the choices of features are sparser for broader environment identification categories. Room type identification focuses on RT parameters and specifically on the extreme bands. This is attributed to the fact that rooms of the same type are likely to share similar shapes and dimensions. Under certain assumptions [23], the shape and dimensions of the enclosure determine properties of the room modes which are less overlapping in the lowest band [23]. The RT at the highest band can reveal attributes related to frequency dependent absorptions. With regards to room identification, significant additions to the PI involve further RT bands, while the extreme bands still remain highly important. The contribution of additional bands is to discriminate between rooms of the same type. Room position identification uses parameters from all three sets, with RT being able to provide information about the room itself, DRR being an indicator of the distance of the speaker to the receiver and MFCC being able to provide information about effects such as coloration [2]. The results show that all these indicators are substantially used to identify the speaker's room and position. Acoustic model selection similarly involves diverse parameters.

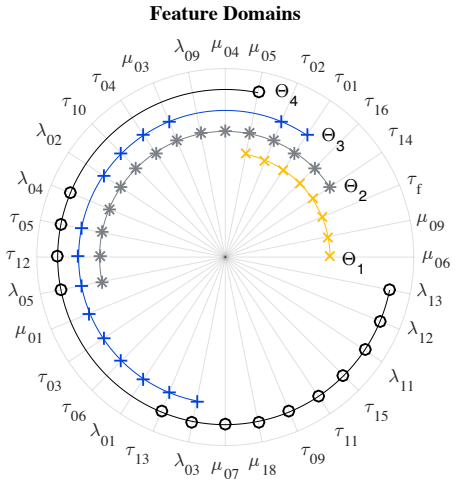
The analysis of the results for the SVM shown in Fig. 1b should be viewed under a different light compared to the CART results. Unlike a CART, a SVM assumes that the data are separable by hyperplanes in the D_k dimensional space. A Gaussian kernel is used to transform the space and increase its dimensionality [12], however the fundamental difference between the two classifiers is still present. Therefore the di-



(a) Predictor Importance to CART growing.



(b) Selected parameters for SVM using BSS.



(c) Resulting feature domains.

Fig. 1: Features selected and resulting feature domains.

verse parameters that are chosen for all four tasks are chosen as good predictors that also match the expected structure for class separability.

For the case of cross-validating the classifier for ASR acoustic model selection, in order to demonstrate the effectiveness of such an approach, a further experiment is performed. WER figures are compared for the case of performing ASR using a single acoustic model trained using anechoic speech and for the case of choosing between multiple models using the designed SVM classifier. The results show that using a single model yields a WER of 57.73% and using the Θ_4-domain SVM classifier with multiple models the WER reduces to 37.00%, a reduction of 20.73 percentage points.

An important observation is the dissimilarity in the parameter choices for each task. The significant difference in the domains for room-type identification and ASR acoustic model selection indicates that the usefulness of the descriptor of the acoustic environment is relative to the task. In this distinct example, all but one of the parameters that provided excellent cross-validation accuracy for the former task are disregarded for the latter. The resulting domain still provides a cross-validation accuracy of 100% for the corresponding task. This highlights the usefulness of targeted parameter-extraction as it would not only improve the computational and memory efficiency of relevant applications but would also not compromise performance.

5. CONCLUSION

This paper presented an analysis of the suitability of a set of acoustic parameters as feature dimensions of discriminative domains for reverberant acoustic environments. Environment identification and the task of acoustic model selection for ASR were considered. Using feature selection methods for classification led to the formation of a feature domain for each task studied. The results of this work provided clear insight for future applications in terms of which acoustic parameters are inherently relevant to each one.

6. REFERENCES

[1] M. A. Ermann, *Architectural Acoustics Illustrated*, Dec. 2014.

[2] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, Springer, 2010.

[3] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[4] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, *Theory and Applications of Spherical Microphone Array Processing*, Springer Topics in Signal Processing. Springer International Publishing, 2017.

[5] N. Peters, H. Lei, and G. Friedland, "Name that room: Room identification using acoustic features in a recording," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 841–844.

[6] M. Mascia, A. Canclini, and F. Antonacci, "Forensic and anti-forensic analysis of indoor/outdoor classifiers based on acoustic clues," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015, pp. 2072–2076.

[7] A. H. Moore, M. Brookes, and P. A. Naylor, "Room identification using roomprints," in *Proc. Audio Eng. Soc. (AES) Conf. on Audio Forensics*, Jun 2014.

[8] P. P. Parada, D. Sharma, P. A. Naylor, and T. van Waterschoot, "Reverberant speech recognition exploiting clarity index estimation," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 1–12, 2015.

[9] A. Tsilfidis, I. Mporas, J. Mourjopoulos, and N. Fakotakis, "Automatic speech recognition performance in different room acoustic environments with and without dereverberation preprocessing," *Computer Speech and Language*, vol. 27, no. 1, pp. 380–395, 2013.

[10] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, MacMillan, New York, 1993.

[11] M. Karjalainen, P. Antsalo, A. Mäkivirta, T. Peltonen, and V. Välimäki, "Estimation of modal decay parameters from noisy response measurements," *J. Audio Eng. Soc. (AES)*, vol. 11, pp. 867–878, 2002.

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[13] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 2015.

[14] D. Steinberg, *CART: classification and regression trees*, vol. 9, CRC Press, 2009.

[15] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from Data*, pp. 199–206. Springer, 1996.

[16] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and regression trees," in *Wadsworth, Monterey, CA, USA*, 1984.

[17] H. Malik and H. Mahmood, "Acoustic environment identification using unsupervised learning," *Security Informatics*, vol. 3, no. 1, pp. 1, 2014.

[18] A. H. Moore, M. Brookes, and P. A. Naylor, "Roomprints for forensic audio," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2013.

[19] J. N. Mourjopoulos, "Digital equalization of room acoustics," *J. Audio Eng. Soc. (AES)*, vol. 42, no. 11, pp. 884–900, Nov. 1994.

[20] S. Bharitkar and C. Kyriakakis, "Perceptual multiple location equalization with clustering," in *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*. IEEE, 2002, vol. 1, pp. 179–183.

[21] I. Omiciuolo, A. Carini, and G. L. Sicuranza, "Multiple position room response equalization with frequency domain fuzzy c-means prototype design," in *Proc. Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2008.

[22] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "Estimation of room acoustic parameters: The ACE challenge," *IEEE Trans. Audio, Speech, Lang. Process.*, June 2016.

[23] H. Kuttruff, *Room Acoustics, Fifth Edition*, CRC Press, June 2009.