ON LOSS FUNCTIONS FOR DEEP-LEARNING BASED T60 ESTIMATION

Yuying Li[1]; Yuchen Liu[2]; Donald S. Williamson[2]

[1]Department of Intelligent Systems Engineering, Indiana University; [2]Department of Computer Science, Indiana University

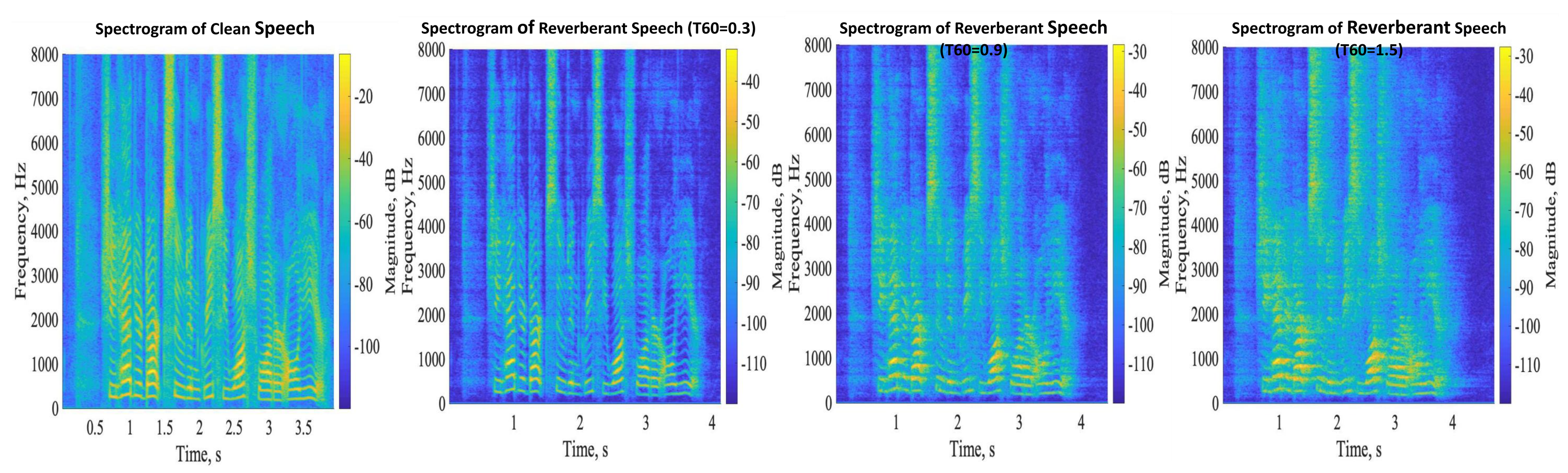
{liyuy, liu477, williads}@indiana.edu

Introduction

The goal of the current study is to estimate reverberation time, T_{60}, by using a deep-learning approach with appropriate loss terms. Previous studies traditionally use signal processing techniques or explore different input features for deep-learning based methods. We propose a composite classification- and regression-based cost function for training a deep neural network that predicts T_{60} for a variety of seen and unseen reverberant conditions. In particular, we explore a multi-task framework that uses magnitude and phase features of the signals, incorporates an additional convolutional-based feature extraction stage, and generates predictions using regression, classification, and classification-based regression training targets.

Motivation

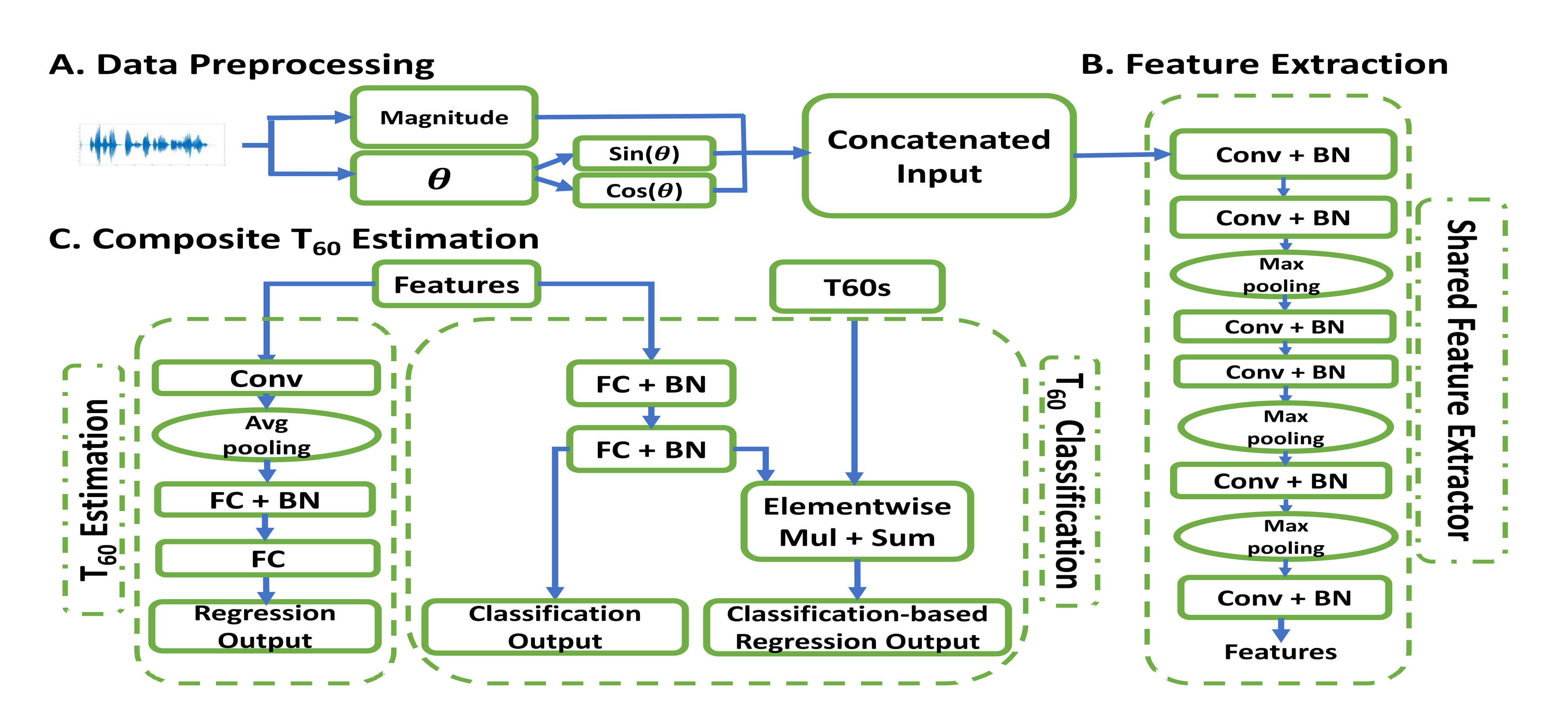
- Reverberation time, T_{60} influences the amount of reverberation in a signal
- T_{60} tells how long it takes a given signal to decay by 60 dB, higher T_{60} times indicate more reverberation
- It contains meaningful information about the room environment, and it also discloses information about the corresponding room impulse response
- By estimating T_{60} help with auditory scene analysis and dereverberation



Previous Studies

- Different Features
  - Mel-frequency cepstral coefficient(MFCC) [Gomez et al., 2010]
  - Gabor feature vector [Bryan 2020]
  - Short-term root-mean square(RMS) [Cox et al., 2001]
- Different model structures
  - Hidden Markov model(HMM) [Hirsch et al., 2008]
  - Multi-layer perceptron(MLP) [Xiong et al., 2013]
  - Convolutional Neural Network(CNN) [Gamper et al., 2018]
- Different loss function
  - Mean-square error (MSE) [Xiong et al., 2013] [Xiong et al., 2015] [Gamper et al., 2018] [Bryan 2020]

Approach



Composite T60 Estimation

- Pure Regression task: directly T_{60} estimation
- Classification task, decomposed into two sub-tasks:
  - Classification: probabilities of each T_{60}
  - Classification-based regression:

$$CReg_{T_{60}} = \sum_{i=1}^{H} (C_{out}^{i} \times T_i), i = 1, \cdots, H$$

Proposed Cost Functions

- Combination of cross-entropy loss L_{cel} and mean-squared error (MSE) L_{reg}:

$$L_{total}^{A} = \beta * L_{cel} + (1-\beta) * L_{reg}, \ \beta \in [0,1]$$

- Incorporated classification-based regression loss L_{creg}:

$$L_{total}^{B} = \beta * (\alpha * L_{cel} + (1-\alpha) * L_{creg}) + (1-\beta) * L_{reg}$$

- Incorporated evaluation scores Pearson's correlation coefficient (PCC) ρ and Spearman's rank correlation coefficient (SRCC) η:

$$L_{total}^{C} = L_{total}^{B} - |\rho_{reg}| - |\eta_{reg}| - |\rho_{cls}| - |\eta_{cls}|$$

- Mean absolute error (MAE) from regression task incorporated:

$$L_{total}^{D} = \beta * \left(\alpha * L_{cel} + (1-\alpha) * (L_{creg} + M_{creg})\right) + (1-\beta) * (L_{reg} + M_{reg}) - |\rho_{reg}| - |\eta_{reg}| - |\rho_{cls}| - |\eta_{cls}|$$

Speech Materials

- Dataset: TIMIT corpus [Garofolo et.al., 1993]
- Randomly select 5,000, 500, and 500 sentences to construct training, validation and testing datasets
- All 6,000 utterances downsampled to 8kHz
- Simulate RIRs from 11 different rooms via image method [Habets 2010].
- Select 13 different reverberation times from 0.3s to 1.5s with steps of 0.1s
- 65,000 reverberant utterances for training set, 6,500 reverberant utterances for validation set, and another 6500 reverberant utterances for testing set

Results

Table 1. Seen Rooms Comparison with different approaches

	MSE		MAE		ρ		η	
	Reg	Cls	Reg	Cls	Reg	Cls	Reg	Cls
MLP [6]	0.075	-	0.211	-	0.783	-	0.788	-
CNN [4]	**0.044**	-	**0.196**	-	0.931	-	0.940	-
$L_{total}^{A}(\beta = 0)$	0.057	0.145	0.208	0.329	0.929	-0.128	0.939	-0.107
$L_{total}^{A}(\beta = 0.4)$	0.270	0.033	0.425	0.147	-0.211	0.927	-0.165	0.940
$L_{total}^{A}(\beta = 1)$	0.448	0.198	0.566	0.365	0.092	0.120	0.101	-0.013
$L_{total}^{B}(\beta = 0.3, \alpha = 0.1)$	0.176	0.135	0.347	0.318	0.781	0.573	0.819	0.635
$L_{total}^{C}(\beta = 0.4, \alpha = 0.2)$	0.131	**0.022**	0.289	**0.116**	0.609	0.955	0.606	**0.973**
$L_{total}^{D}(\beta = 0.3, \alpha = 0)$	0.098	0.093	0.270	0.228	0.771	0.808	0.800	0.816
$L_{total}^{D}(\beta = 0.9, \alpha = 0.1)$	0.120	0.057	0.290	0.204	**0.955**	**0.963**	**0.958**	0.968
$L_{total}^{D}(\beta = 0.3, \alpha = 1)$	0.284	0.250	0.435	0.412	-0.013	0.428	0.003	0.430

Table 2. Unseen Rooms Comparison with different approaches

	MSE		MAE		ρ		η	
	Reg	Cls	Reg	Cls	Reg	Cls	Reg	Cls
MLP [6]	0.092	-	0.239	-	0.715	-	0.723	-
CNN [4]	0.096	-	0.212	-	0.856	-	0.860	-
$L_{total}^{A}(\beta = 0)$	**0.047**	0.145	**0.189**	0.329	0.942	-0.098	0.953	-0.084
$L_{total}^{A}(\beta = 0.4)$	0.298	0.056	0.449	0.171	-0.042	0.919	-0.198	0.942
$L_{total}^{A}(\beta = 1)$	0.457	0.201	0.577	0.368	0.040	0.069	0.050	-0.070
$L_{total}^{B}(\beta = 0.3, \alpha = 0.1)$	0.174	0.136	0.345	0.319	0.830	0.476	0.872	0.546
$L_{total}^{C}(\beta = 0.4, \alpha = 0.2)$	0.117	**0.023**	0.273	**0.114**	0.532	0.968	0.525	**0.984**
$L_{total}^{D}(\beta = 0.3, \alpha = 0)$	0.092	0.089	0.261	0.221	0.845	0.814	0.866	0.837
$L_{total}^{D}(\beta = 0.9, \alpha = 0.1)$	0.102	0.045	0.263	0.180	**0.962**	**0.973**	**0.962**	0.977
$L_{total}^{D}(\beta = 0.3, \alpha = 1)$	0.295	0.242	0.444	0.405	0.219	0.601	0.229	0.622

Conclusions

- Our approach incorporates composite classification and regression-based cost function for training a deep neural network that predicts T_{60}
- Our approach is different from recent methods and benefits from the two tasks
- Our approach benefits from dividing the classification tasks into two subtasks
- The results show that the tradeoff between weighting classification versus regression tasks does influence results

INDIANA UNIVERSITY