Mixture Vocal Loudness Estimation and Genre Analysis

Kelian (Mike) Li
Music Informatics Lab
Center for Music Technology
kli421@gatech.edu

Introduction

With the isolated vocal and accompaniment tracks available for training, we can use machine learning to estimate the vocal loudness from the mixture signal.

The ground truth is the short-term loudness of isolated vocal loudness. The audio signals are sliced into overlapped segments of 3 seconds length with the hop size of 100 ms following the EBU recommendation

The input features to the SVR model are the first 20 Mel-Frequency Cepstrum Coefficients (MFCC), 128 VGGish embeddings, and the short-term loudness of the mixture signal.

Because the mixture loudness can be measured directly from the audio signal, estimating the relative vocal loudness (vocal-to-mixture ratio) is equivalent to estimating the absolute vocal loudness. The target vocal loudness values are transformed from the absolute short-term loudness to the loudness difference compared with the mixture loudness,
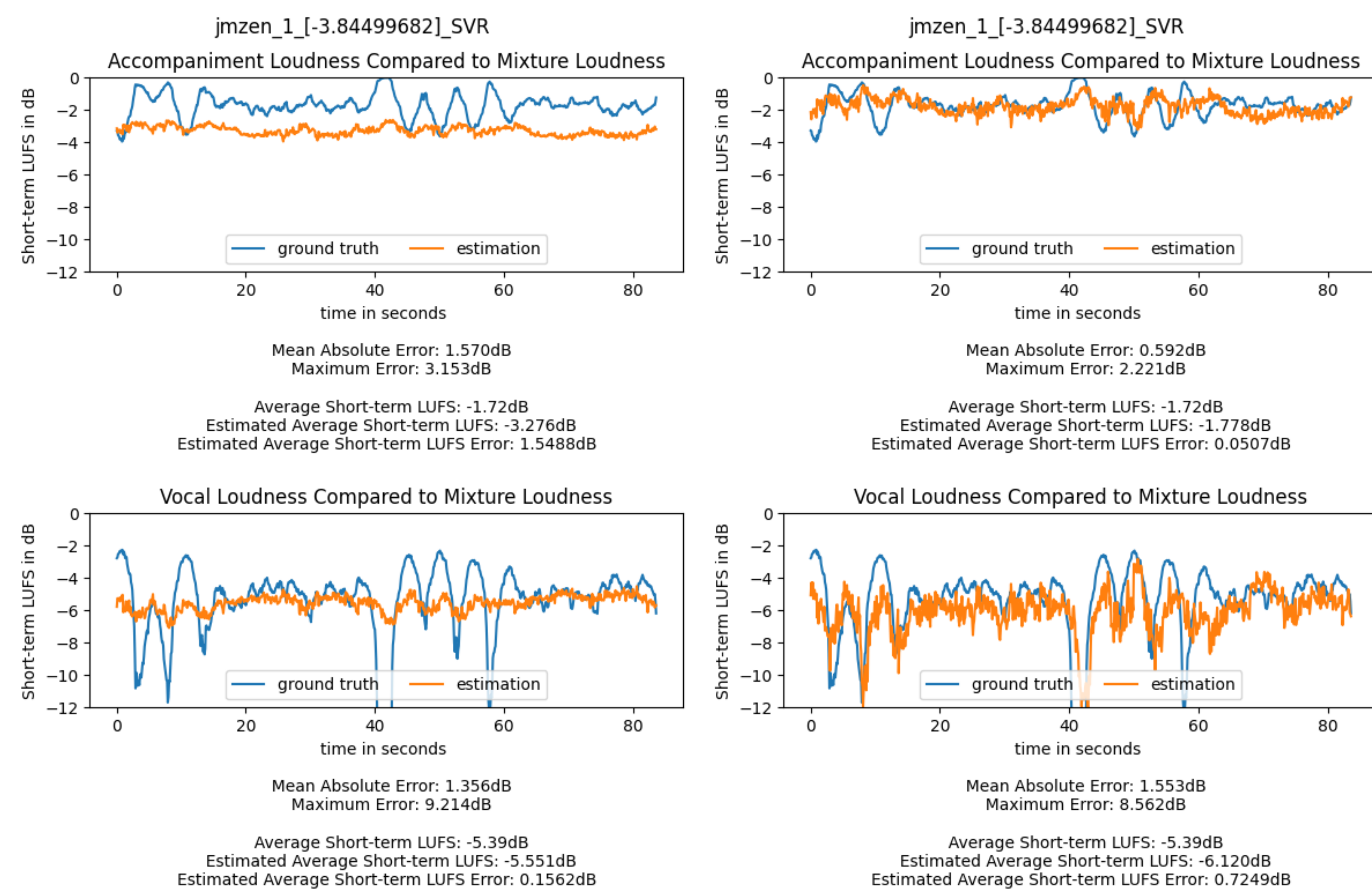
$$STL_{\Delta vox} = STL_{vox} - STL_{mixture} \quad (1)$$

where $STL_{\Delta vox}$ is the relative vocal loudness, STL_{vox} is the short-term loudness of the vocal signal, and $STL_{mixture}$ is the short-term loudness of the mixture signal.

Error Histogram
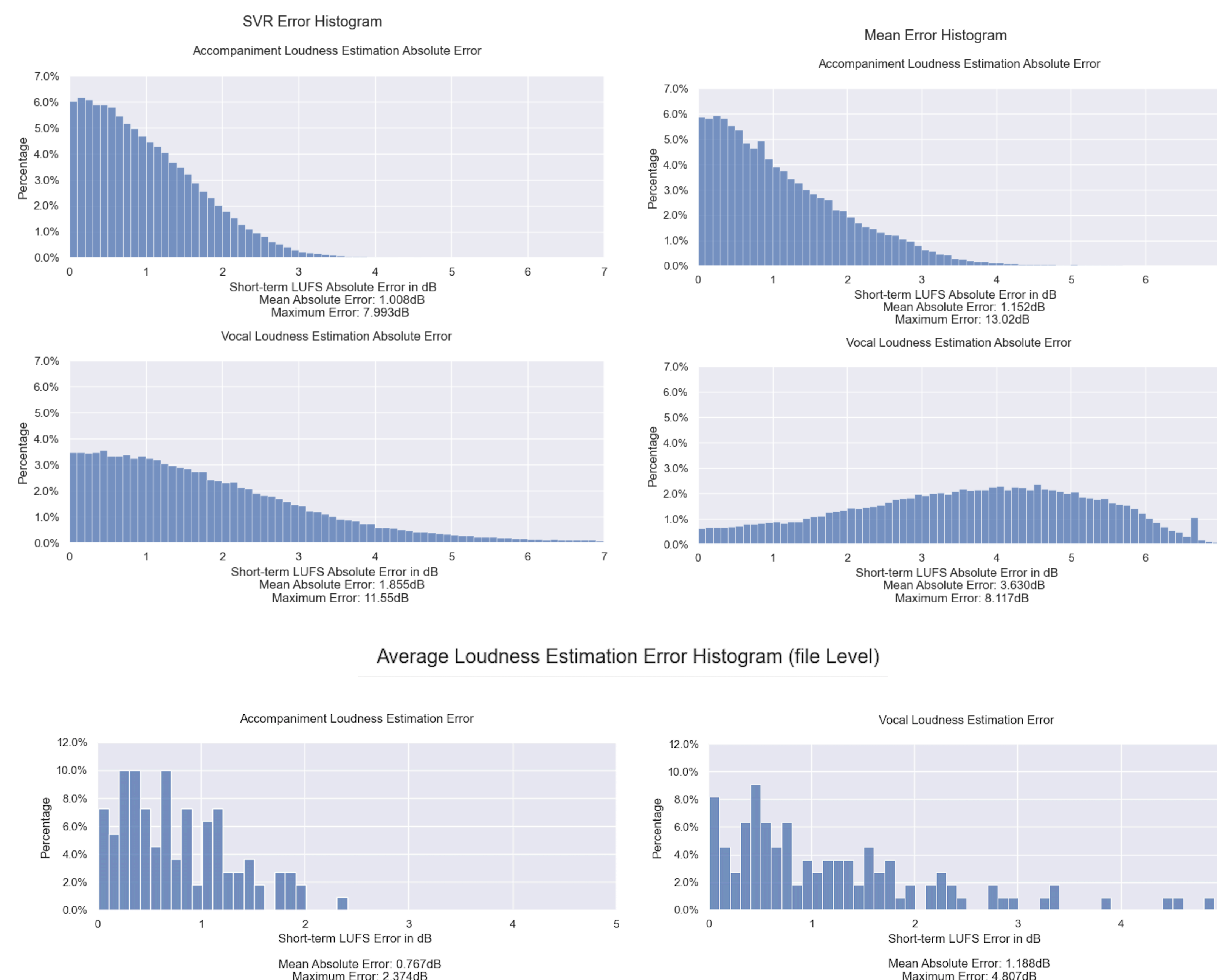
Estimation

- The left figure uses epsilon of 0.3, and the right figure uses epsilon of 0.1.
- Epsilon = 0.3 yields lower MAE, but unable to predict the loudness outside of a small range.
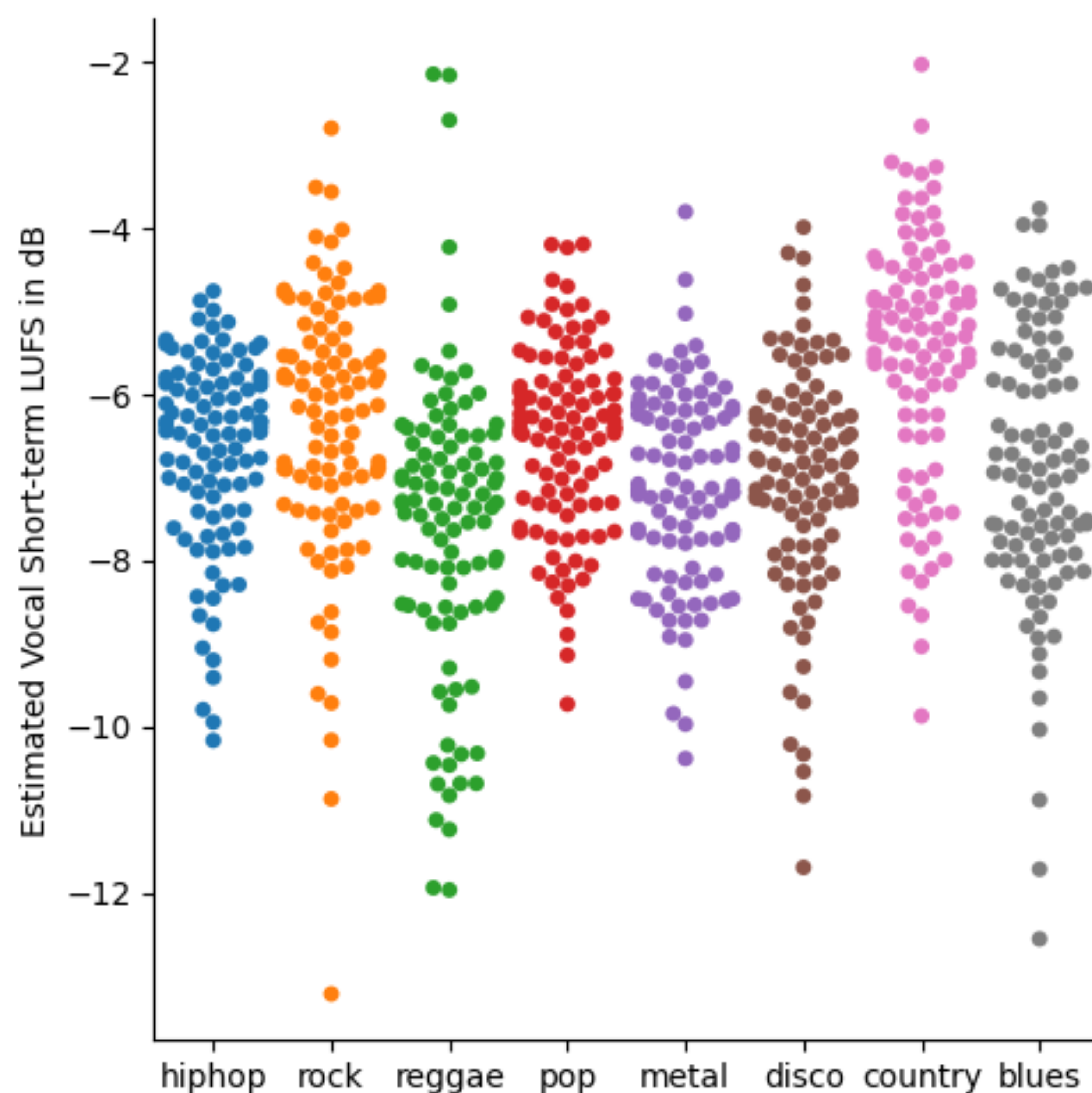
Results

TABLE I
RESULTS OF THE VOCAL AND ACCOMPANIMENT LOUDNESS ESTIMATION

	VOX MAE(dB)	VOX ME	ACC MAE	ACC ME
Mean*	3.65	**6.35**	1.17	3.65
SVR	**1.86**	7.12	**1.00**	**3.12**

*The baseline system.

Genre Analysis

Genre	hiphop	rock	reggae	pop	metal	disco	country	blues
Relative Vocal Loudness (dB)	-6.69	-6.36	-7.58	-6.52	-7.13	-7.02	-5.51	-6.90

CONTACT

Kelian (Mike) Li
Music Informatics Lab
Center for Music Technology
kli421@gatech.edu