On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers*

B.T Atmaja, M. Akagi
Japan Advanced Institute of Science and Technology

TENCON 2020

* paper, slide, and codes are available at
https://github.com/bagustris/ravdess_song_speech

Introduction

- Speech emotion recognition (SER) has been extensively studied over the years, it enters commercial market in recent years.
- Music is increasingly being used to understand cognitive and neural function in populations [1]; music itself has been created largely to express emotions.
- Understanding processing differences in speech and song is useful to implement different strategies to cope with their differences.
- This study evaluate the effect of different feature sets, region of analysis (feature types), and classifiers on emotional song and speech.

Dataset

- RAVDESS dataset [1] was used: the dataset contains lexically-matched emotional song and speech.
- Although the dataset is multimodal, the video data is not used (only speech and song)
- The dataset was created using induced emotional expressions.
- The speech data includes seven emotion categories: calm, happy, sad, angry, fearful, surprise, disgust and neutral (1440 samples)
- Song includes five emotion categories: calm, happy, sad, angry, and fearful; and a neutral (1012 utterances).

[1] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS),” PLoS One, pp. 1–35, 2018.

Acoustic Feature Sets

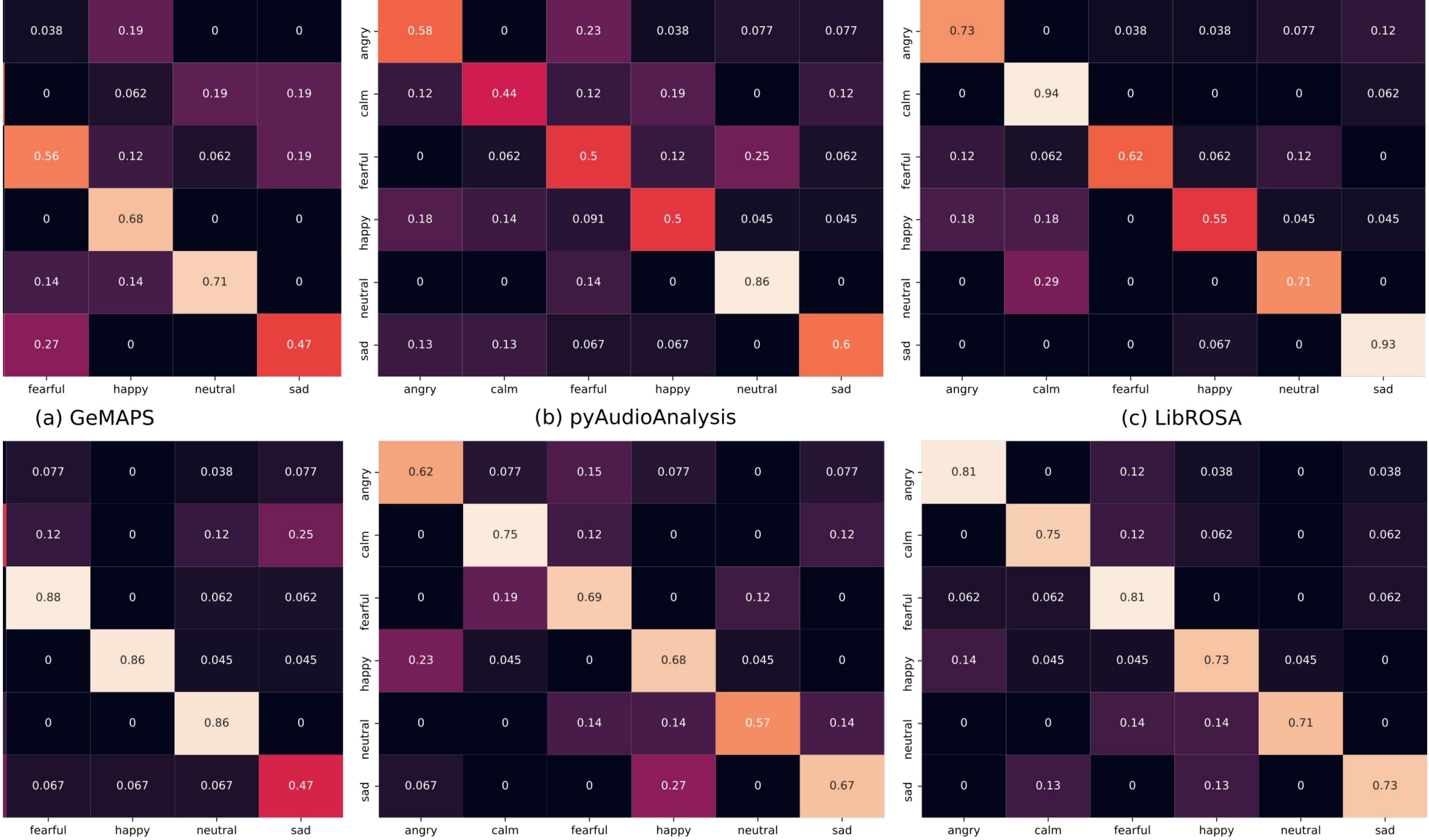
Feature set	LLDs
GeMAPS	intensity, alpha ratio, Hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, F0, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), harmonic difference H1-H2, harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.
pAudioAnalysis	zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectra flux, spectral roll-off,13 MFCCs, 12 chroma vectors, chroma deviation.
LibROSA	40 MFCCs, 12 chroma vectors, 128 mel-scaled spectrograms, 7 spectral contrast features, 6 tonal centroid features.

HSF were also extracted from those 3 feature sets

Classifiers

- MLP: 3 Dense layers @256 stacked
- LSTM: 3 LSTM layers @256 stacked
- GRU: 3 GRU layers @256 stacked
- CNN 1D: 3 Conv1D layers @256 stacked

Result: Song Data



(a) GeMAPS

(b) pyAudioAnalysis

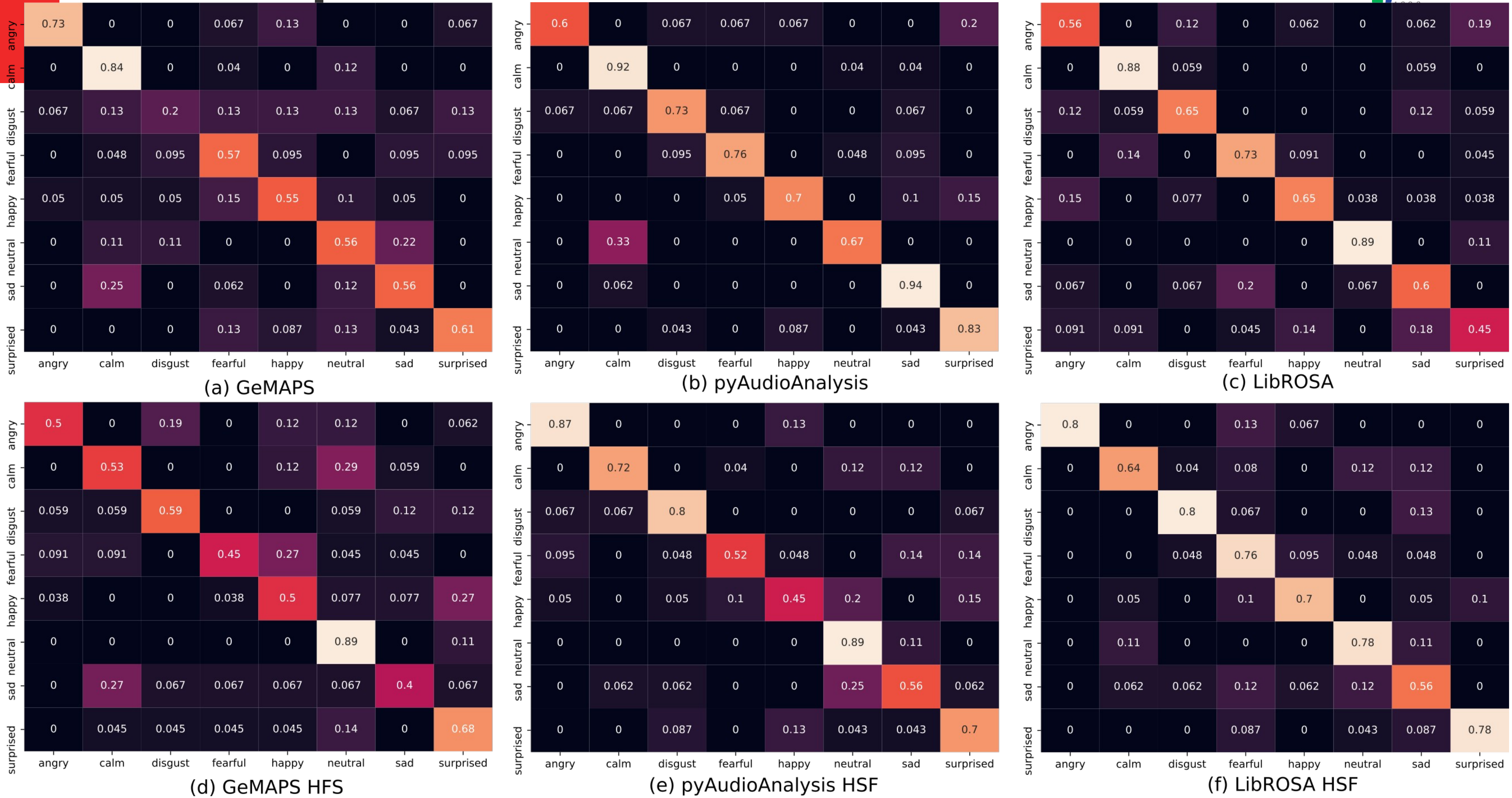
(c) LibROSA

(d) GeMAPS HSF

(e) pyAudioAnalysis HSF

(f) LibROSA HSF

Result: Speech Data



	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.73	0	0	0.067	0.13	0	0	0.067
calm	0	0.84	0	0.04	0	0.12	0	0
disgust	0.067	0.13	0.2	0.13	0.13	0.13	0.067	0.13
fearful	0	0.048	0.095	0.57	0.095	0	0.095	0.095
happy	0.05	0.05	0.05	0.15	0.55	0.1	0.05	0
neutral	0	0.11	0.11	0	0	0.56	0.22	0
sad	0	0.25	0	0.062	0	0.12	0.56	0
surprised	0	0	0	0.13	0.087	0.13	0.043	0.61

(a) GeMAPS

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.6	0	0.067	0.067	0.067	0	0	0.2
calm	0	0.92	0	0	0	0.04	0.04	0
disgust	0.067	0.067	0.73	0.067	0	0	0.067	0
fearful	0	0	0.095	0.76	0	0.048	0.095	0
happy	0	0	0	0.05	0.7	0	0.1	0.15
neutral	0	0.33	0	0	0	0.67	0	0
sad	0	0.062	0	0	0	0	0.94	0
surprised	0	0	0.043	0	0.087	0	0.043	0.83

(b) pyAudioAnalysis

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.56	0	0.12	0	0.062	0	0.062	0.19
calm	0	0.88	0.059	0	0	0	0.059	0
disgust	0.12	0.059	0.65	0	0	0	0.12	0.059
fearful	0	0.14	0	0.73	0.091	0	0	0.045
happy	0.15	0	0.077	0	0.65	0.038	0.038	0.038
neutral	0	0	0	0	0	0.89	0	0.11
sad	0.067	0	0.067	0.2	0	0.067	0.6	0
surprised	0.091	0.091	0	0.045	0.14	0	0.18	0.45

(c) LibROSA

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.5	0	0.19	0	0.12	0.12	0	0.062
calm	0	0.53	0	0	0.12	0.29	0.059	0
disgust	0.059	0.059	0.59	0	0	0.059	0.12	0.12
fearful	0.091	0.091	0	0.45	0.27	0.045	0.045	0
happy	0.038	0	0	0.038	0.5	0.077	0.077	0.27
neutral	0	0	0	0	0	0.89	0	0.11
sad	0	0.27	0.067	0.067	0.067	0.067	0.4	0.067
surprised	0	0.045	0.045	0.045	0.045	0.14	0	0.68

(d) GeMAPS HFS

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.87	0	0	0	0.13	0	0	0
calm	0	0.72	0	0.04	0	0.12	0.12	0
disgust	0.067	0.067	0.8	0	0	0	0	0.067
fearful	0.095	0	0.048	0.52	0.048	0	0.14	0.14
happy	0.05	0	0.05	0.1	0.45	0.2	0	0.15
neutral	0	0	0	0	0	0.89	0.11	0
sad	0	0.062	0.062	0	0	0.25	0.56	0.062
surprised	0	0	0.087	0	0.13	0.043	0.043	0.7

(e) pyAudioAnalysis HSF

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	0.8	0	0	0.13	0.067	0	0	0
calm	0	0.64	0.04	0.08	0	0.12	0.12	0
disgust	0	0	0.8	0.067	0	0	0.13	0
fearful	0	0	0.048	0.76	0.095	0.048	0.048	0
happy	0	0.05	0	0.1	0.7	0	0.05	0.1
neutral	0	0.11	0	0	0	0.78	0.11	0
sad	0	0.062	0.062	0.12	0.062	0.12	0.56	0
surprised	0	0	0	0.087	0	0.043	0.087	0.78

(f) LibROSA HSF

Result: Effect of Different Classifiers

Classifier	Song		Speech	
	Accuracy	UAR	Accuracy	UAR
MLP	0.794	0.804	0.729	0.755
LSTM	0.820	0.813	0.785	0.781
GRU	0.812	0.844	0.785	0.764
Conv1D	0.743	0.806	0.687	0.690

Result: Effect of Different Feature Sets

Feature	Song		Speech	
	Accuracy	UAR	Accuracy	UAR
GeMAPS	0.637	0.592	0.602	0.614
GeMAPS HSF	0.753	0.762	0.662	0.653
pyAudioAnalysis	0.592	0.619	0.731	0.701
pyAudioAnalysis HSF	0.736	0.761	0.658	0.620
LibROSA	0.751	0.780	0.732	0.676
LibROSA HSF	**0.820**	**0.813**	**0.774**	**0.781**

Conclusions

- An evaluation of speech and song emotion recognition across different feature sets, region of analysis (feature types), and classifiers has been performed.
- No remarkable difference between song and speech emotion recognition; the best features set, feature type, and classifier on speech also obtain the similar result on song.
- Song is more emotional than speech