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

Bagus Tris Atmaja
Department of Engineering Physics
Sepuluh Nopember Institute of Technology
Surabaya, Indonesia
bagus@jaist.ac.jp

Masato Akagi
School of Information Science
Japan Adv. Inst. of Sci. & Tech.
Nomi, Japan
akagi@jaist.ac.jp

***Abstract*—In this paper, we argue that singing voice (song) is more emotional than speech. We evaluate different features sets, feature types, and classifiers on both song and speech emotion recognition. Three feature sets: GeMAPS, pyAudioAnalysis, and LibROSA; two feature types, low-level descriptors and high-level statistical functions; and four classifiers: multilayer perceptron, LSTM, GRU, and convolution neural networks; are examined on both songand speech data with the same parameter values. The results show no remarkable difference between song and speech data on using the same method. Comparisons of two results reveal that song is more emotional than speech. In addition, high-level statistical functions of acoustic features gained higher performance than low-level descriptors in this classification task. This result strengthens the previous finding on the regression task which reported the advantage use of high-level features.**

***Index Terms*—Song emotion recognition, speech emotion recognition, acoustic features, emotion classifiers, affective computing**

I. Introduction

Music emotion recognition is an attempt to recognize emotion, in either categories or dimensions, within pieces of music. Music expresses and induces emotion. Therefore, the universality of emotion within the music can be extracted regardless the origin of music [1]. Recognizing emotion in music is important, for instance, in the music application's recommender system.

A Song is part of music that is performed by the human voice. While research on music emotion recognition is well established, research on song emotion recognition is less developed. Since it is a part of music, recognizing song emotion recognition is essential for music emotion recognition. This research aims to evaluate song emotion recognition in parallel with speech emotion recognition.

In speech emotion recognition, several acoustic features and classifier have been developed. Moore et al. proved that using high-level features improves the performance of emotion recognition compared to using low-level features [2]. Moreover, Atmaja and Akagi [3] showed that using high-level statistical functions (HSF), i.e., Mean+Std, of low-level descriptors (LLDs) of pyAudioAnalysis feature set [3] gains higher performance than using LLD itself. The reported results are obtained in dimensional emotion recognition tasks. In categorical emotion recognition, we know no report showed the effectiveness of Mean+Std from categorical emotion recognition. We evaluated different types of features (LLDs vs HSFs) from three feature sets in categorical song and speech emotion recognition as well as evaluation of different classifiers.

The contribution of this paper, besides the feature types evaluation, is an evaluation of a feature set derived from LibROSA toolkit [4]. We compared both LLDs and HSF from selected LibROSA acoustic features to GeMAPS [5] and pyAudioAnalysis [6]. We expect an improvement of performance from those two feature sets by utilizing a larger number of acoustic features, particularly on HSF feature type. Although the data is relatively small, i.e., about 1000 utterances for both song and speech, we choose deep learning-based classifiers to evaluate those features due to its simplicity. Other machine learning method, e.g., support vector machine, may obtain higher performance due to its effectiveness on smaller data.

II. Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset from Ryerson University is used. This dataset contains multimodal recordings of emotional speech and song on both audio and video formats. Speech includes seven emotion categories: calm, happy, sad, angry, fearful, surprise, and disgust expressions; a neutral with a total of 1440 utterances. Song includes five emotion categories: calm, happy, sad, angry, and fearful; and a neutral with a total of 1012 utterances. Both speech and song are recorded at 48 kHz. The detail of the dataset can be found in [7].

III. Methods

Three main methods are evaluated for both emotional speech and song: three different feature sets, two feature types for each feature set, and four classifiers. The following three sections describe each of those methods.

A. Feature Sets

The first evaluated feature set for emotional speech and song is Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [5]. This feature set is a proposal to standardize acoustic features for voice and affective computing. Twenty-three acoustic features, known as low-level descriptor (LLD), are chosen as a minimalistic parameter set while the extended version contains added spectral and frequency related parameters and

its functional. The extended version (eGeMAPS) consist of 88 parameters. This research used the minimalistic GeMAPS feature set due to its effectiveness compared to other features sets [5], [8].

The openSMILE toolkit [9] was used to extract 23 LLDs GeMAPS feature set for each time frame. This frame-based processing is conducted with 25 ms window length and 10 ms hop length resulting (523, 23) feature size for speech data and (633, 23) feature size for song data.

The second evaluated feature set is pyAudioAnalysis (pAA). pyAudioAnalysis was designed for general-purpose Python library for audio signal analysis. The library provides a wide range of audio analysis procedures including: feature extraction, classification of audio signals, supervised and unsupervised segmentation, and content visualization [6]. Thirty-four LLDs are extracted on frame-based processing from this feature set. Those LLDs, along with the previous GeMAPS and next LibROSA feature sets, are shown in Table I.

TABLE I
ACOUSTIC FEATURES SETS USED TO EVALUATE SONG AND SPEECH EMOTION RECOGNITION.

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.

As the final feature set, we selected five features from LibROSA feature extractor including MFCCs, chroma, mel spectrogram, spectral contrast and tonnetz. The number of features for each LibROSA LLD is shown in Table I with a total of 193 features for a time frame. This number of features is chosen based on experiments. The detail of LibROSA version used in this experiment can be found in [4].

B. Feature Types

The traditional method to extract acoustic feature from a speech is done on a frame-based processing method, i.e., the aforementioned LLDs in each acoustic feature set. The higher level acoustic features can be extracted as statistical aggregation functions over LLDs on fixed-time processing, e.g., an average feature value of each 100 ms, 500 ms, 1 s, or per utterance. This high-level statistical functions (HSF) is intended to roughly describe the temporal variations and contours of the different LLDs over a fixed-time or an utterance [10]. In dimensional speech emotion recognition, this HSF feature performs better than LLDs, as reported in [2], [3], [11].

Schmitt and Schuller [11] evaluated mean and standard deviation (Mean+Std) of GeMAPS feature set and compared it with eGeMAPS and bag-of-audio-word (BoAW) representations of LLDs. The result showed that Mean+Std works best among the three. Based on this finding, we incorporated Mean+Std as HSFs from the previously explained three feature sets. The Mean+Std is calculated per utterance on each feature set resulting difference size/dimension for each feature set, i.e. 46-dimensional for GeMAPS, 68-dimensional for pyAudioAnalysis, and 386-dimensional for selected LibROSA features. Hence, two feature types are evaluated; LLD and HSF, from GeMAPS, pyAudioAnalysis and selected LibROSA features.

C. Classifiers

Four classifiers are evaluated: a dense network (or multi-layer perceptron, MLP), a long short-term memory (LSTM) network, a gated recurrent unit (GRU) network, and a convolution network. The brief explanations of those networks are described below.

1) MLP: Three dense layers are stacked with 256 units and ReLU activation function for each layer. The last dense layer is flattened, and a dropout rate with probability 0.4 is added after it. The final layer is a dense layer with eight units for speech and six units for song with a softmax activation function.
2) LSTM: Three LSTM layers are stacked with 256 units each and returned all values. The rest layers are the same as MLP classifier, i.e., a dropout layer with probability 0.4 and a dense layer with a softmax activation function.
3) GRU: The GRU classifiers similar to LSTM. The LSTM stack is replaced by GRU stack without changing other parameters.
4) Conv1D: Three 1-dimensional convolution networks are stacked with 256 units and a ReLU activation function for each layer. The filter lengths (strides) are 4, 8, 12 for first, second, and third convolution layers. The rest layers are similar to other classifiers.

IV. RESULTS AND DISCUSSION

We divided our results into two parts, analysis of different feature sets and feature types (i.e., difference features) and analysis of different classifiers. Both results are presented in terms of accuracy and unweighted average recall (UAR). Accuracy is widely used to measure the classification error rate in balanced/near-balanced data. It defines the number of correctly classified examples divided by the total number of examples, usually presented in % (here we used 0-1 scale). UAR is an average recall from all classes, i.e., the number of correctly classified positive examples divided by the total number of positive examples in each class. UAR is widely used to justify classification method from imbalanced data.

A. Effect of different feature types

Table II shows accuracy and UAR of different feature sets and feature types. On different feature types, LibROSA-based acoustic features perform best on both emotional song and speech data. On different feature sets, HSF features perform better than LLD, except on pyAudioAnalysis feature set for speech data. Only in that pyAudioAnalysis feature set LLDs of pyAudioAnalysis obtained better accuracy and UAR than its HSF. In overall evaluation, our proposal on using Mean+Std of LibROSA-based acoustic features perform best among six different features. This finding suggests that Mean+Std performs well not only on dimensional speech emotion recognition (regression task) but also on categorical song and speech emotion recognition.

Fig. 1. Confusion matrix of different evaluated acoustic features on *song* data.

To find in which emotion category our method performs best and worse, we performed confusion matrix presentation as shown in Figure 1 and 2. The results, however, show inconsistency among different features sets on both song and speech data. The highest and lowest recall scores vary among features. For instance, in song data the highest recall from GeMAPS feature set is angry while from LibROSA is calm. On both song and speech data, it was found the HSF improved recall scores of LLDs features. Since each feature has a different highest recall score, it is interesting to combine those features sets to improve the current result for future research direction.

Although we found no remarkable different on the performance trends from the same features sets between emotional song and speech data, a significant difference maybe found on the using of specific acoustic features, e.g., F0 contour, spectral features and amplitude envelope, as reported in [12] for emotional singing voice.

B. *Effect of different classifiers*

For different classifiers, the accuracy and recall scores are presented in Table IV-B. On the previous table, the results was obtained using LSTM classifiers. In overall evaluation, the LSTM classifier obtained the best result among other three

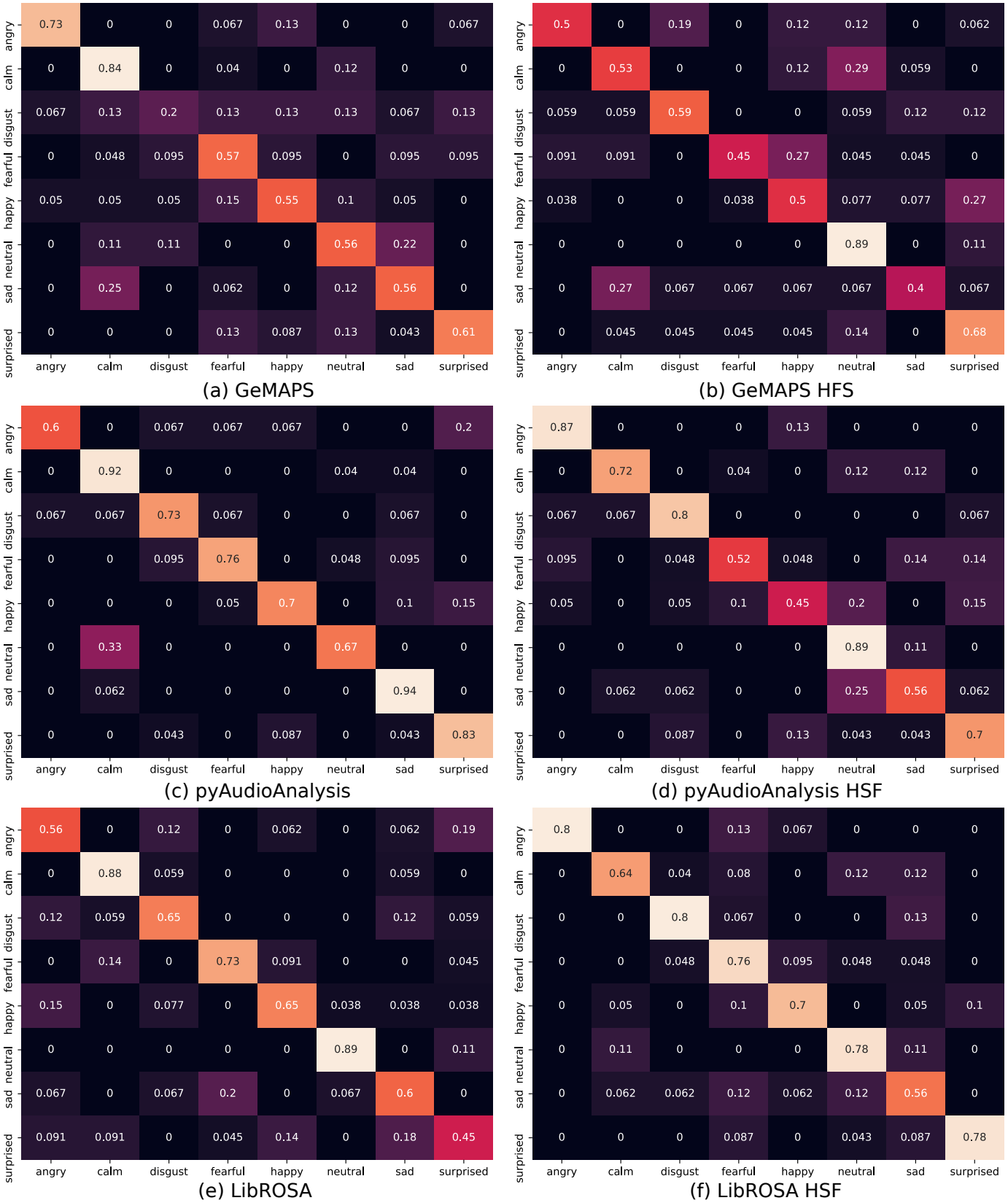

Fig. 2. Confusion matrix of different evaluated acoustic features on *speech* data.

TABLE II
ACCURACY AND UNWEIGHTED AVERAGE RECALL (UAR) OF DIFFERENT FEATURE SETS AND FEATURE TYPES ON EMOTIONAL SONG AND SPEECH BASED ON 10-FOLD VALIDATION; BOTH ARE IN 0-1 SCALE.

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**

TABLE III
ACCURACY AND UNWEIGHTED AVERAGE RECALL (UAR) OF EMOTIONAL SONG AND SPEECH ON DIFFERENT CLASSIFIERS USING LIBROSA HSF FEATURE BASED ON 10-FOLD VALIDATION; BOTH ARE IN 0-1 SCALE.

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

classifiers. However, GRU classifier shows better UAR on song data, i.e., on recognizing each emotion category. This result shows that recurrent-based classifiers (LSTM and GRU) performs better than MLP and 1-dimensional convolution network on classification of emotional song and speech. Similar to the previous Table II, the song data showed higher scores than speech data. Since the song data contains fewer data (samples) and fewer emotion categories than speech data, it can be concluded that song is more emotional than speech. The result on the same classifier, same feature set, and same feature type supports this finding.

V. CONCLUSIONS

We presented an evaluation of different feature sets, feature types, and classifiers for both song and speech emotion recognition. First, we conclude that there is no remarkable difference between song and speech emotion recognition on the same features and classifiers based on the evaluated methods. In other words, the features types/sets and classifiers which gain better performance on song data will also gain better performance on speech data. Second, song is more emotional than speech. On both accuracy and unweighted average recall, the scores obtained by song data always higher than speech data. Both song and speech data contain the same statements; hence, the different is the intonation/prosody and other acoustic information which is captured by acoustic features. Third, on different feature types, high-level statistical functions consistently performed better than low-level descriptors. For the future research direction, we planned to combine different acoustic features types and sets since the result showed differences among emotion categories.

REFERENCES

[1] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici, and S. Koelsch, "Universal Recognition of Three Basic Emotions in Music," *Curr. Biol.*, vol. 19, no. 7, pp. 573–576, apr 2009.

[2] J. D. Moore, L. Tian, and C. Lai, "Word-level emotion recognition using high-level features," in *Int. Conf. Intell. Text Process. Comput. Linguist.* Springer, 2014, pp. 17–31.

[3] B. Tris Atmaja and M. Akagi, "The Effect of Silence Feature in Dimensional Speech Emotion Recognition," in *10th Int. Conf. Speech Prosody 2020*, no. May. ISCA: ISCA, may 2020, pp. 26–30.

[4] B. McFee and Others, "librosa/librosa: 0.7.1," oct 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3478579

[5] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[6] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, pp. 1–17, 2015. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis/

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

[8] B. T. Atmaja and M. Akagi, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," in *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2020, pp. 4482—-4486. [Online]. Available: https://ieeexplore.ieee.org/document/9052916/

[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. 21st ACM Int. Conf. Multimed. - MM '13*. New York, New York, USA: ACM Press, 2013, pp. 835–838. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2502081.2502224

[10] S. Mirsamadi, E. Barsoum, C. Zhang, and M. S., "Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2017*, no. March, pp. 2227–2231, 2017.

[11] M. Schmitt and B. Schuller, "Deep Recurrent Neural Networks for Emotion Recognition in Speech," in *DAGA*, 2018, pp. 1537–1540.

[12] T. H. Nguyen, "A Study on Correlates of Acoustic Features to Emotional Singing Voice Synthesis Nguyen Thi Hao," Ph.D. dissertation, 2018.