High-Level Analysis of Audio Features for Identifying Emotional Valence in Human Singing

Stuart Cunningham
Manchester Metropolitan University
Manchester, UK
s.cunningham@mmu.ac.uk

Jonathan Weinel
Aalborg University
Aalborg, Denmark
jonweinel@gmail.com

Richard Picking
Wrexham Glyndŵr University
Wrexham, UK
r.picking@glyndwr.ac.uk

ABSTRACT

Emotional analysis continues to be a topic that receives much attention in the audio and music community. The potential to link together human affective state and the emotional content or intention of musical audio has a variety of application areas in fields such as improving user experience of digital music libraries and music therapy. Less work has been directed into the emotional analysis of human acapella singing. Recently, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was released, which includes emotionally validated human singing samples. In this work, we apply established audio analysis features to determine if these can be used to detect underlying emotional valence in human singing. Results indicate that the short-term audio features of: energy; spectral centroid (mean); spectral centroid (spread); spectral entropy; spectral flux; spectral rolloff; and fundamental frequency can be useful predictors of emotion, although their efficacy is not consistent across positive and negative emotions.

CCS CONCEPTS

Applied computing~Sound and music computing • Human-centered computing~Interaction techniques • Human-centered computing~Interactive systems and tools

KEYWORDS

Affect, valence, music, audio features, singing.

ACM Reference format:

Stuart Cunningham, Jonathan Weinel and Richard Picking. 2018. Analysis of Audio Features for Identifying Emotional Valence in a Singing Dataset. In *Proceedings of Audio Mostly 2018: Sound in Immersion and Emotion conference, Wrexham, Wales UK, September 2018 (AM'18), 4 pages.*

1 Introduction

The field of affective computing [1] has had rapid expansions into the world of sound and music, with studies involving the automatic analysis and emotional classification of music [2,3] having applications in a variety of fields, including navigation of music libraries [4,5] as well as fields such as health and music therapy [6]. In this work, our focus moves away from that of produced music that is polyphonic and characterised by the presence of multiple instruments, which can be a challenging and complex acoustical domain. Instead, we present some initial findings from an exploration of how certain short-term audio features might be used to predict emotional valence (the negative or positive direction or emotional 'state') being articulated in another form of music - that of unaccompanied human singing. Identification of useful audio features in differentiating emotional states would be a useful and valuable step in working towards more complex emotion recognition systems in singing.

2 Related Work

In a study comparing the expression of emotion between speaking and singing, Scherer *et al.* [7] undertook an analysis of two categories of audio features: the distribution of energy across the spectrum and measures of signal variability in the frequency and amplitude domains. This study utilised a series of audio files, produced by recording multiple singers, which were statistically analysed to identify significant differences in expression of each emotional state. Notably, the authors found that in the expression of arousal, the singers made use of a set of perturbation techniques, specifically vibrato, to influence the emotional intention of their recital. Their findings suggest that the expression of emotion through singing utilizes many of the same techniques as speech. The authors suggest that the expression of emotion via the human voice need not be dependent upon the meaning of the words being sung or spoken, since, in their study, the lyrics were nonsensical. In addition, the authors indicate that there is a lack of other studies and datasets available for comparison.

Other studies have taken similar approaches, utilizing feature analysis and statistical measures, whilst also highlighting the challenge that the recognition and prediction of arousal is simpler than that of valence [8, 9].

3 Valence and Feature Analysis in RAVDESS

3.1 The RAVDESS Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a recently released set of human emotional expressions, which has been externally validated, and consists of audio, video and audio-visual materials [10]. A total of 24 actors were used in the construction of the data set, each being asked to produce expressions of a range of discrete emotional states with two levels of intensity. Actors produced two versions of each utterance or song. The discrete emotions in the complete data set are: neutral; calm; happy; sad; angry; fearful; disgust; and surprised. The actors produced these materials through scripted speech and singing samples. In the case of singing, the emotional states of disgust and surprised were not present [10, 11].

Data was missing for one of the singers in the data set[i]. This led to the total number of actors being reduced to 23. The singing samples in the dataset are split into two levels of emotional intensity: normal and strong, which we equated with the notion of arousal in terms of the view of a dimensional model of emotion, leaving the named emotions as mapping to valence. This is a simplification introduced for early experiments with the dataset. As the focus of this work was to investigate valence, we used only the normal emotional intensity samples from the dataset. This resulted in the materials totaling 552 unique audio files (23 actors, reciting two statements, with 6 emotional intentions, and two repetitions of each). The mean duration of these files is 4.65 seconds ($\sigma = 0.43$).

3.2 Analysis Method

A series of short-term audio features were computed upon the singing samples selected from all twenty-four actors and incorporating the six emotional states, described in the previous section. The audio features were extracted using the *Matlab 2018a* software and the *Matlab Audio Analysis Library* devised by Giannakopoulos and Pikrakis [12, 13]. Features were calculated using a 50 ms window with 50% overlap. The features used were:

- Zero-crossing rate (ZCR);
- Energy;
- Energy Entropy
- Spectral Centroid (mean);
- Spectral Centroid (spread);
- Spectral Entropy;
- Spectral Flux;
- Spectral Rolloff;
- Fundamental frequency (F0).

This produced a time-bound set of features for each of the singing samples. Given the short nature of each sample and the validated response of each sample representing a single emotion, a mean value was subsequently calculated for each of the listed features per actor and per emotional valence state. Finally, a grand mean and standard deviation were calculated for the six discrete emotional states using means from each actor-feature pair.

[i] For an unknown reason, the folder that should contain the singing samples of Actor 18 in the RAVDESS dataset was empty at the time when the clips were to be analysed.

3.3 Results

To provide an initial analysis of how each feature might be used as a predictor of valence, we conducted multinomial logistic regression analysis, using emotional state as the dependent variable and the nine continuous audio features as co-variates. The neutral emotional state was used as a reference category, allowing us to determine how each feature might be used to indicate transition from the neutral to any of the five remaining states. The overall final model produced demonstrated significant performance in predicting emotional state $\chi^2(45) = 150.30$, Nagelkerke $R^2 = 0.682$, $p < 0.001$. This model fitting statistic stands in contrast to the goodness-of-fit outcome: Pearson $\chi^2 2(640) = 1100.62$, $p < 0.001$.

Table 1 demonstrates that: energy; spectral centroid (mean); spectral centroid (spread); spectral entropy; spectral flux; spectral rolloff; and fundamental frequency were statistically significant predictors of emotional state in the sample singing recordings.

Table 1: Contribution of Audio Features as Emotional Valence Predictors in Multinomial Logistic Regression (* = p < 0.05)

Feature	χ^2	df	p-value
Zero-Crossing Rate	6.61	5	0.251
Energy	23.35	5	<0.001*
Energy Entropy	4.63	5	0.463
Spectral Centroid (mean)	12.23	5	0.032*
Spectral Centroid (spread)	16.52	5	0.005*
Spectral Entropy	39.82	5	<0.001*
Spectral Flux	23.40	5	<0.001*
Spectral Rolloff	37.32	5	<0.001*
Fundamental Frequency (F0)	32.53	5	<0.001*

Fig. 1. through Fig. 7. show each significant audio feature's mean and standard deviation in order to provide a descriptive illustration of how each emotional state is represented by the respective audio feature. As such, the values shown correspond to each one of the six emotional valence states under investigation.

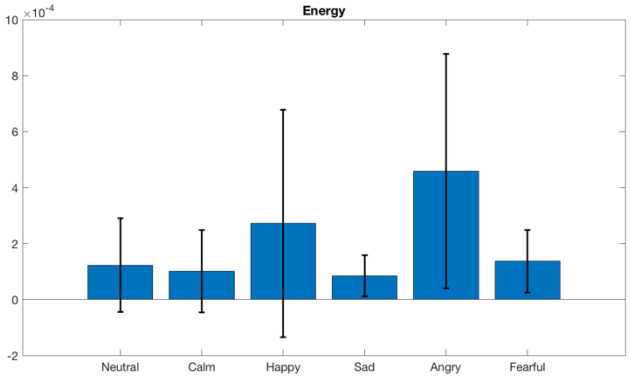

Figure 1. Energy

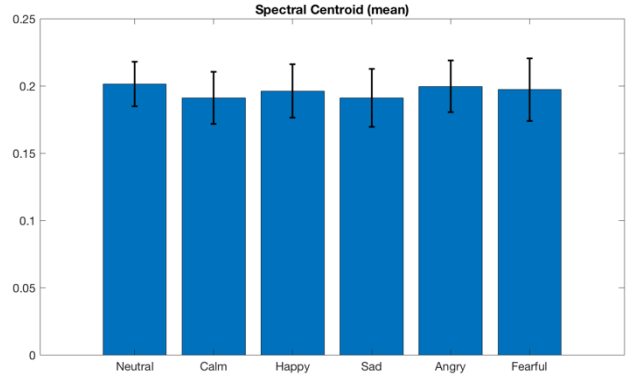

Figure 2. Spectral Centroid (mean)

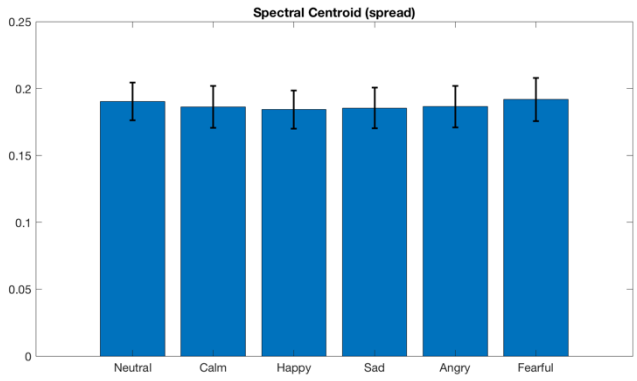

Figure 3. Spectral Centroid (spread)

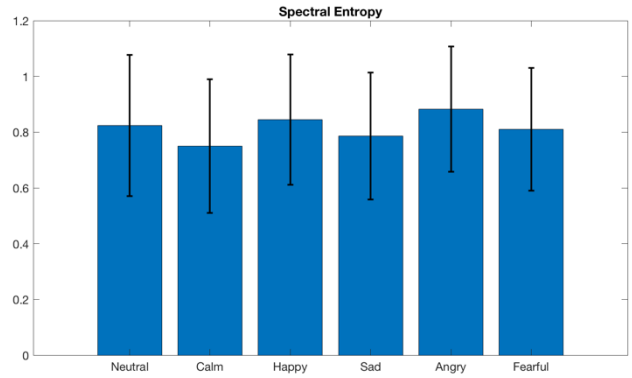

Figure 4. Spectral Entropy

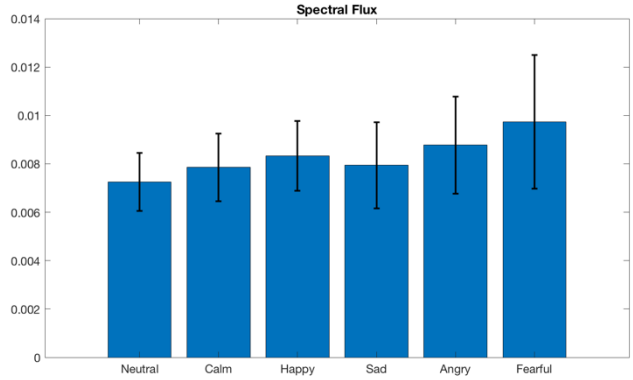

Figure 5. Spectral Flux

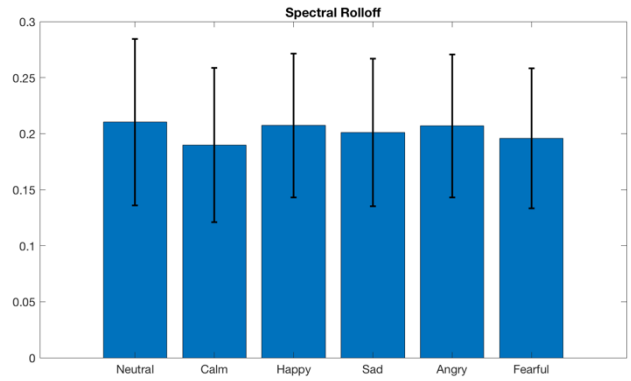

Figure 6. Spectral Rolloff

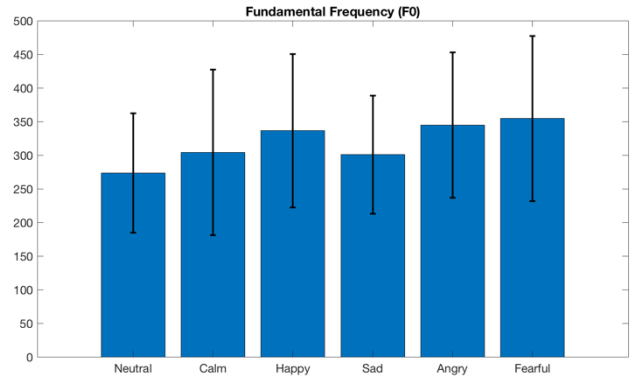

Figure 7. Fundamental Frequency (F0)

Significant predictors for comparing the neutral emotional state with the remaining five states are shown in Table 2, showing that the fearful and angry emotional states were the most frequently occurring states that could be predicted from the neutral state. It is an interesting observation that some features are estimates of more than one emotional state and that the majority of these states could be considered as negative emotions.

Table 2: Statistically Significant ($p < 0.05$) Parameter Estimates Contrasting the Neutral State with the other Emotional States

Predictor	Neutral vs.	p-value
Spectral Centroid (mean)	Fearful	0.022
Spectral Centroid (spread)	Fearful	0.006
Spectral Entropy	Angry	<0.001
	Fearful	0.006
Spectral Flux	Angry	0.009
	Fearful	<0.000
Spectral Rolloff	Angry	<0.001
Fundamental Frequency (F0)	Angry	0.002
	Calm	0.021

The quality of these predictors is evident in Table 3, which shows the classification of singing samples into the corresponding emotional states using the resultant regression model.

Table 3: Classification of Emotion using the Regression Model

Observed	*Angry*	*Calm*	*Fearful*	*Happy*	*Neutral*	*Sad*	**Correct**
Angry	17	0	0	4	2	0	73.9%
Calm	1	7	3	2	4	6	30.4%
Fearful	3	2	13	0	4	1	56.5%
Happy	5	3	3	8	4	0	34.8%
Neutral	0	2	2	5	12	2	52.2%
Sad	0	5	3	2	4	9	39.1%
Overall	18.8%	13.8%	17.4%	15.2%	21.7%	13.0%	**47.8%**

4 Discussion

The results provide a useful initial insight into the potential that short-term audio features might have in developing automated mechanisms for discrete emotional valence recognition in human singing. It was shown that frequency-domain features were most useful in being able to indicate emotional valence in the employed subset of samples from the RAVDESS dataset and that these features performed better in their prediction of negative valanced emotions, most notably the states of angry and fearful.

However, the overall performance of the regression model produced remains uncertain, especially since its classification capacity remains less than 50% overall and issues arose around its overall goodness-of-fit.

The static nature of emotional arousal in this study was a deliberate control choice made to reduce the time required to undertake the study. However, a natural expansion of this work would be to examine the complete RAVDESS dataset, including the strong intensity samples, to determine which audio features may be fruitful in providing estimates of this emotional component.

It is expected that regression models might be refined to a greater extent by incorporating such a more complete picture of the range of emotional states being conveyed and that the inclusion of mid-term audio features may also be helpful in expanding this work. For larger singing samples, it may be necessary to consider such analysis on a temporal level and consider options, such as windowing the signal and providing and time-domain emotional categorization. Similarly, comparing the regression approach to more complex machine learning approaches, such as neural networks also appear to be valid avenues for expansion.

In future work, the characteristics of the RAVDESS dataset should be more closely integrated into feature analysis. For example, the discrete emotional states should be attributed specific arousal and valence values, to be used as dependent variables, which could be accomplished by using a preexisting source, such as the Affective Norms for English Words (ANEW) dataset [14]. This would then allow intensity to be explored in addition, rather than its simplified equivalence with arousal, as in this early exploration of the RAVDESS data.

REFERENCES

[1] R. Picard, 1997. *Affective Computing*, MIT Press. Cambridge, MA.

[2] Y.H. Yang, Y.C. Lin, Y.F. Su, H.H. and Chen, H.H., 2008. A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, *16*(2), pp.448-457.

[3] S. Beveridge and D. Knox, D., 2012. A feature survey for emotion classification of western popular music. In *Proceedings of the 9th international symposium on computer music modeling and retrieval,(CMMR): Music and emotions, June* (pp. 19-22).

[4] D. Griffiths, S. Cunningham, and J. Weinel, 2013, September. A discussion of musical features for automatic music playlist generation using affective technologies. In *Proceedings of the 8th Audio Mostly Conference* (p. 13). ACM.

[5] C.Y. Chi, R.T.H. Tsai, J.Y. Lai, and J.Y.J. Hsu, 2010, November. A reinforcement learning approach to emotion-based automatic playlist generation. In *Technologies and Applications of Artificial Intelligence (TAAI), 2010 International Conference on* (pp. 60-65). IEEE.

[6] D. Knox, S. Beveridge, L.A. Mitchell, and R.A. MacDonald, 2011. Acoustic analysis and mood classification of pain-relieving music. The Journal of the Acoustical Society of America, 130(3), pp.1673-1682.

[7] K.R. Scherer, J. Sundberg, L. Tamarit, and G.L. Salomão, 2015. Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, *29*(1), pp.218-235.

[8] F. Eyben, G.L. Salomão, J. Sundberg, K.R. Scherer, K.R., and B.W. Schuller, 2015. Emotion in the singing voice—a deeper look at acoustic features in the light of automatic classification. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1), p.19.

[9] B. Zhang, E.M. Provost, R. Swedberg, and G. Essl, G., 2015, January. Predicting Emotion Perception Across Domains: A Study of Singing and Speaking. In *AAAI* (pp. 1328-1335).

[10] S.T. Livingstone and F.A. Russo, 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Version 1.0.0) [Data set]. PLoS ONE. Zenodo. http://doi.org/10.5281/zenodo.1188976

[11] S.R. Livingstone and F.A. Russo, 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), p.e0196391.

[12] T. Giannakopoulos and A. Pikrakis., 2014. *Introduction to audio analysis: a MATLAB® approach*. Academic Press.

[13] T. Giannakopoulos, 2014. Matlab Audio Analysis Library. Mathworks File Exchange. Available at: https://www.mathworks.com/matlabcentral/fileexchange/45831-matlab-audio-analysis-library [Last accessed: 10th June 2018].

[14] M.M. Bradley and P.J. Lang, 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings* (Vol. 30, No. 1, pp. 25-36). Technical report C-1, the center for research in psychophysiology, University of Florida.