The role of the singing acoustic cues in the perception of broad affect dimensions

Pauline Mouawad[1], Myriam Desainte-Catherine[1], Anne Gégout-Petit[2], and Catherine Semal[3]
University of Bordeaux and CNRS: [1]LaBRI, [2]IMB, [3]INCIA
{pauline.mouawad, myriam.desainte-catherine}@u-bordeaux1.fr
anne.petit@u-bordeaux2.fr
catherine.semal@ipb.fr

Abstract. This experiment investigated the role of acoustic correlates of the singing voice in the perception of broad affect dimensions using the two-dimensional model of affect. The dataset consisted of vocal and glottal recordings of a sung vowel interpreted in different singing expressions. Listeners were asked to rate the sounds according to four perceived affect dimensions. A cross-tabulation was done between the singing expressions and affect judgments. A one-way ANOVA was performed for 11 acoustic cues with the affect ratings. It was found that the singing power ratio (SPR), mean intensity, brightness, mean pitch, jitter, shimmer, mean harmonic-to-noise ratio (HNR), and mean autocorrelation discriminate broad affect dimensions. Principal component analysis (PCA) was performed on the acoustic correlates. Two components were retained that explained 78.1% of the total variance of vocal cues and 73.5% of that of the glottal cues.

Keywords: vocal, glottal, singing expressions, affect dimensions

1 Introduction

Voice “is a primary instrument for emotional expression” [14] and “emotional expression is an essential aspect in singing” [16]. Although substantial research has addressed affect perception in speech, it is still in its early stages for the singing voice. Previous experiments have asked professional singers to perform songs according to a set of discrete emotions achieving results varying in accuracy and different emotions not identified equally well [5], [14]. Furthermore, the acoustic cues that determined listeners’ judgments mediated the singers’ emotional state and therefore no judgment could be made as to the inherent faculty of the singing voice in conveying emotions independently of the singer’s affect expression.

This experiment has two aims: first, to learn whether listeners perceive broad affect dimensions in a singing voice that doesn’t portray a specific emotion. And second if affect is perceived, to reveal the intrinsic role of the acoustical characteristics of the singing voice in the perception of affect. To our knowledge, these questions have not been studied before.

2 Experiment

2.1 Stimuli

The stimuli was taken from the Singing Voice Database[1] and consisted of scale recordings of vocal as well as glottal sounds of a sung vowel ‘ah’ interpreted by professional singers (1 male and 1 female). The musical notes range from A2 to E4 and A3 to A5 for male and female voice respectively. The recordings are mono files in WAVE PCM format, at 16 bits and 44 kHz. The sound files were trimmed using MIRToolbox [6] to remove the silence at the beginning, and were segmented using R statistical software [12] so that only the first note of the scale is retained. The final dataset consisted of 44 sound samples, 22 vocal and 22 glottal of 1 second duration each in the following singing expressions[2]: bounce, hallow, light, soft, sweet, flat, mature, sharp, clear, husky and no expression. The female sound files don’t include the flat singing expression. The type of the stimuli was relevant as the singers didn’t perform specific emotions and there was no accompanying music or lyrics to influence the listener’s affect perceptions, hence their judgments were expected to relate to the voice alone.

2.2 Participants

The participants consisted of 9 males and 6 females, (age M = 26.1, SD = 9.6) of whom 1 is a professional singer and 4 have had some kind of formal singing training. 7 reported enjoying singing, 14 agreed that music expresses emotions and that the voice of the singer is important to their personal enjoyment of a song. All participants reported that the voice of the singer is important in expressing emotions in singing.

2.3 Procedure

Participants were asked to rate the perceived affect dimension of each voice sample on a 5-point Likert scale using the two-dimensional model of affect [13] represented by four broad affect terms: pleasant-unpleasant for valence, and awake-tired for arousal [15]. Considering that with today’s internet bandwidth and sound technologies it has become possible to conduct psychoacoustic tests over the internet [2], the experiment was distributed through email with instructions explaining its objectives. Participants could play the sound file more than once, and could at any time save their answers and come back to complete it later. Each sample occurred 3 times in the dataset and the order of the files was randomized. Duration of the experiment was 30 minutes.

[1] http://liliyatsirulnik.wix.com/liliyatsirulnik1411#!scale/cee5

[2] http://liliyatsirulnik.wix.com/liliyatsirulnik1411#!synopsis/cjg9

3 Results

3.1 Affect Ratings

Considering that the number of responses for each of the 5 categories on the Likert scale was slight therefore making it difficult to meet the assumptions of statistical validity for the ANOVA, the responses were grouped under ‘pleasant’, ‘unpleasant’, ‘neutral’ for valence, and ‘awake’, ‘tired’, ‘neutral’ for arousal. For example, responses for ‘awake’ and ‘extremely awake’ were grouped under ‘awake’. The mean of the ratings for the 3 occurrences of each file was computed, and then the file was classified according to the emotion that brought the highest total number of votes. On the valence dimension, 16 were rated as pleasant (13 vocal, 3 glottal), 22 were rated as unpleasant (7 vocal, 15 glottal) and 6 were rated as neutral (2 vocal, 4 glottal). On the arousal dimension, 22 were rated as awake (17 vocal, 5 glottal), 17 were rated as tired (3 vocal, 14 glottal) and 5 were rated as neutral (2 vocal, 3 glottal).

3.2 Acoustic Features

A total of 11 acoustic features were selected according to their perceptual validity as established in the relevant literature (see Table 1) and were extracted from the original sound files using Praat software [1]. Pitch information was retrieved using a cross-correlation method for voice research optimization, with pitch floor and ceiling set to 75 Hz and 300 Hz respectively for male voice and to 100 Hz and 500 Hz respectively for female voice. The spectrum was obtained from the waveform using Fast Fourier Transform method with a dynamic range of 70 dB, a window length of 5 ms and a view range from 0 to 5000 Hz for male and from 0 to 5500 Hz for female voice. The singer's formant [16], [9] was quantified by computing the singing power ratio (SPR) [8]. To this end the two highest spectrum peaks between 2 and 4 kHz and between 0 and 2 kHz were identified and SPR was obtained by computing the ‘amplitude difference in dB between the highest spectral peak within the 2 – 4 kHz range and that within the 0 – 2 kHz range’ [7]. Since the ‘perceptual singer’s formant’ is ‘contributed by the underlying acoustic formants F2, F3 and F4’ [9], the means of F2, F3 and F4 were measured individually for each sound file. The mean intensity of the sound was measured using energy averaging method. Measures of jitter, shimmer mean autocorrelation and mean harmonics-to-noise ratio (HNR) were extracted from the voice report. Brightness was extracted using MIRToolbox [6].

3.3 Analysis

The entire analysis was carried out in the R statistical software environment [12].

Singing Expressions and Affect. To our knowledge, the relationship of various singing expressions to affect dimensions is not established in the literature. A cross-tabulation was done for the affect judgments and the singing expressions to determine the counts of the combination of each factor level.

Table 1. List of acoustic features and relevant literatures

Features	Literature
Singing Formants: F2 to F4	Sundberg et al. 1994, Ishi and Campbell 2012, Millhouse and Clermont, 2006
Singing Power Ratio (SPR)	Grichkovtsova et al. 2011, Laukkanen et al. 1997, Sundberg et al. 1994, Omori et al., 1996, Watts et al., 2004, Lundy et al., 2000
Mean Intensity, Mean Pitch	Jansens et al. 1997, Patel et al. 2011
Jitter, Shimmer, Harmonicity: Mean HNR, Mean Autocorrelation	Lundy et al., 2000, Scherer K., 1995
Brightness	Ishi and Campbell, 2012
General voice acoustic attributes	http://www.speech-therapy-information-and-resources.com

On the valence dimension, all vocals in light, soft and sweet expressions as well as 67% of those with no specific expression were perceived as pleasant; mature and sharp vocals were perceived as unpleasant. All glottals in bounce, husky, mature, sharp and no expression were perceived as unpleasant, and soft glottals were perceived as pleasant. On the arousal dimension, all vocals in bounce, clear, mature, sharp and no expression were perceived as awake, and those in soft and sweet expressions were perceived as tired (low energy). All glottals in clear, hallow, soft and sweet as well as 67% of those having no specific singing expression were perceived as tired.

Acoustic Cues and Affect. A one-way ANOVA was performed for each acoustic measure with valence and arousal as factors with three levels each. The analysis results were verified using Tukey's multiple comparisons of means and were Bonferroni corrected.

On the valence dimension, acoustic cues whose means were statistically different for the pleasant-unpleasant factors are SPR, mean intensity, jitter, shimmer, mean autocorrelation and mean HNR for vocal files (see Table 2), and brightness, mean intensity, shimmer and mean autocorrelation for glottal files, with brightness being significant for the pleasant-neutral factors as well (see Table 3). Comparing mean values between vocal and glottal sound files, it is noticed that on the pleasant dimension the shimmer's mean value is higher in glottal sounds and lower in vocal sounds, and mean autocorrelation's mean value is lower in glottal sounds and higher in vocal sounds.

On the arousal dimension, acoustic cues whose means were statistically different for the awake and tired factors are SPR and mean intensity for vocal files (see Table 4), and mean intensity, jitter, shimmer, mean HNR and mean pitch for glottal files, with mean HNR being also significant for the neutral-awake factors (see Table 5). On the pleasant dimension, the mean intensity's mean value is higher for both sound types.

Table 2. *p* values, mean and standard deviation of vocal cues for pleasant-unpleasant

		Pleasant		Unpleasant	
Acoustic features	P	M	SD	M	SD
SPR	0.002	28.580	7.475	15.900	5.998
Mean intensity	0.001	53.940	5.026	62.670	1.834
Jitter	0.023	0.653	0.334	1.166	0.471
Shimmer	0.032	3.591	0.857	5.123	1.710
Mean Autocorrelation	0.004	0.985	0.006	0.952	0.031
Mean HNR	0.001	21.440	2.131	16.200	3.305

Table 3. *p* values, mean and standard deviation of glottal cues for pleasant-unpleasant

		Pleasant		Unpleasant	
Acoustic features	P	M	SD	M	SD
Brightness	0.032* 0.003	0.163	0.105	0.069	0.018
Mean intensity	0.027	59.420	7.172	65.890	2.601
Shimmer	0.011	7.583	3.865	3.750	1.528
Mean Autocorrelation	0.042	0.970	0.026	0.988	0.007

*pleasant-neutral

Table 4. *p* values, mean and standard deviation of vocal cues for awake-tired

		Awake		Tired	
Acoustic features	P	M	SD	M	SD
SPR	0.002	19.130	6.689	32.430	7.275
Mean intensity	0.002	60.310	3.898	52.500	5.394

Table 5. *p* values, mean and standard deviation of glottal cues for awake-tired

		Awake		Tired	
Acoustic features	P	M	SD	M	SD
Mean intensity	0.020	68.840	2.335	63.510	3.911
Jitter	0.007	0.560	0.136	1.285	0.423
Shimmer	0.023	2.314	0.704	5.257	2.263
Mean HNR	0.006 0.031*	29.320	3.019	23.570	3.328
Mean Pitch	0.014	218.700	3.937	138.100	51.543

*awake-neutral

Principal Component Analysis. The motivation for performing a PCA on the acoustic correlates is duple: first, to determine what are the main components that best describe the stimuli knowing that these exist in the audio feature data, and second, to project the sounds on a factorial plane that illustrates graphically their distribution on the valence-arousal groups alongside the acoustic variables used for the PCA. Two components were retained that explained 78.1% of the total variance of vocal cues and 73.5% of that of the glottal cues.

For the vocal cues, the first component explains 57.7% of the original variance and accounts mainly for variations in SPR, F4, mean pitch, opposed to jitter and mean intensity; the second component explains a further 20.4% of the original variance and account mainly for variations in brightness. Figures 1 and 2 show that the vocal files projected onto the PC1-PC2 planes appear to cluster reasonably according to valence and arousal, although a bit weaker for arousal. For example, pleasant sounds are those having higher values of SPR, F4 and mean pitch and lower values for jitter and mean intensity and/or lower values for brightness. Unpleasant sounds are those having higher values for jitter and mean intensity and lower values for SPR, F4 and mean pitch, and/or higher values for brightness.

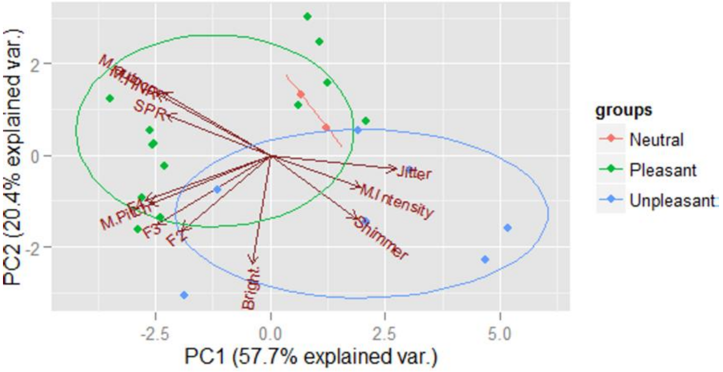

Fig. 1. Vocal files projected on PC1-PC2 plane and clustered by valence

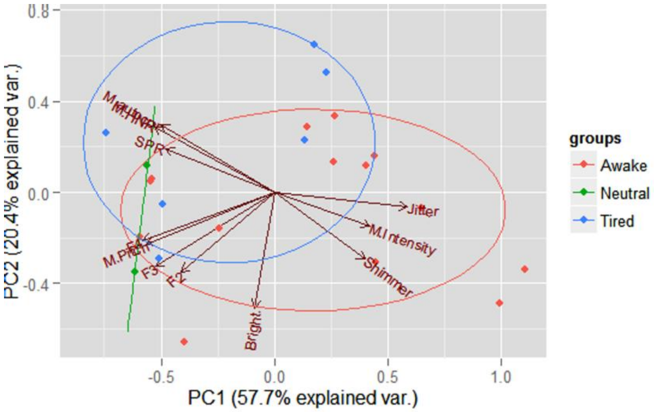

Fig. 2. Vocal files projected on PC1-PC2 plane and clustered by arousal

Sounds perceived as awake are those having higher values for jitter and mean intensity and rather lower values for brightness, and sounds perceived as tired are those having higher values for SPR, mean pitch and F4 and/or lower values for brightness. For the glottal cues, the first component explains 53.4% of the original variance and accounts for variations in shimmer, F2, F3, opposed to mean intensity, mean autocorrelation and mean HNR; the second component explains a further 20.1% of the original variance and accounts for variations in mean pitch and F4 (see Fig. 3 and Fig. 4).

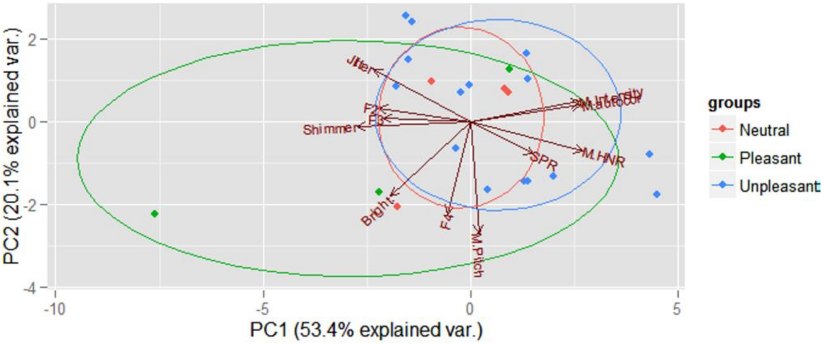

Fig. 3. Glottal files projected on PC1-PC2 plane and clustered by valence

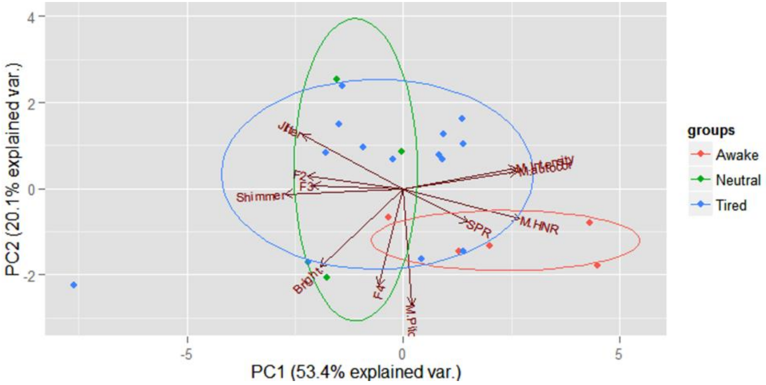

Fig. 4. Glottal files projected on PC1-PC2 plane and clustered by arousal

4 Conclusion and Future Work

This experiment reveals that broad affect dimensions are perceived in a singing voice independently of the singer’s emotional expression. The analysis revealed 8 features that explained the variance with respect to affect and are: SPR, mean intensity, brightness, jitter, shimmer, mean pitch, mean HNR and mean autocorrelation. This could have implications on the use of synthesized voices in the research on vocal expression of affect. PCA revealed 2 components that accounted for variations in 11 acoustic cues including the aforementioned 8 features. Further investigation is needed to assess the strength of the relationship between the components and the affect dimensions. Finally, this study will be replicated using singing voices expressing the four affect dimensions and the results will be contrasted with the present findings. We expect this would lead to conclusions of interest to affective voice synthesis.

References

1. Boersma, Paul.: Praat, a System for Doing Phonetics by Computer. Glot International 5:9/10, 341—345 (2001)
2. Cox, Trevor J.: Tutorial: Public Engagement through Audio Internet Experiments. University of Salford (2011)
3. Grichkovtsova I., Morel M. and Lacheret A.: The Role of Voice Quality and Prosodic Contour in Affective Speech Perception. Speech Communication (2011)
4. Ishi, C., and Campbell, N.: Analysis of Acoustic-prosodic Features of Spontaneous Expressive Speech. Revista de Estudos da Linguagem 12(2) (2012)
5. Jansens S., Bloothooft G. and De Krom, G.: Perception and Acoustics of Emotions in Singing. Proceedings of the 5th Eurospeech, 4, pp. 2155—2158 (1997)
6. Lartillot, O. and Toiviainen, P.: A Matlab Toolbox for Musical Feature Extraction from Audio. Proceedings of the 10th International Conference on Digital Audio Effects (DAFx-07) (2007)
7. Lin E., Jayakody D. and Looi V.: The Singing Power Ratio and Timbre-Related Acoustic Analysis of Singing Vowels and Musical Instruments. Voice Foundation's 38th Annual Symposium: Care of the Professional Voice (2009)
8. Lundy, D. S., Roy, S., Casiano R., Xue J. and Evans J.: Acoustic Analysis of the Singing and Speaking Voice in Singing Students. Journal of Voice, 14(4), pp. 490—493 (2000)
9. Millhouse, T. and Clermont, F.: Perceptual Characterisation of the Singer’s Formant Region: a Preliminary Study. Proceedings of the Eleventh Australian International Conference on Speech Science and Technology, pp. 253—258 (2006)
10. Omori, K., Kacker, A., Carroll, L. M., Riley, W. D. and Blaugrund, S. M.: Singing Power Ratio: Quantitative Evaluation of Singing Voice Quality. Journal of Voice, 10(3), pp. 228—235 (1996)
11. Patel S., Scherer K. R., Björkner E. and Sundberg J.: Mapping Emotions into Acoustic Space: the Role of Voice Production. Biological psychology, 87 (1), pp. 93—98 (2011)
12. R development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing (2008)
13. Russel, J.: A Circumplex Model of Affect. Journal of Personality and Social Psychology, 39(6), pp. 1161 1178 (1980)
14. Scherer K. R.: Expression of Emotion in Voice and Music. Journal of Voice, 9 (3), pp. 235—248 (1995)
15. Schimmack, U. and Grob, A.: Dimensional Models of Core Affect: A Quantitative Comparison by Means of Structural Equation Modeling. European Journal of Personality, 14, 325—345 (2000)
16. Sundberg, J., Iwarsson, J. and Hagegard, H.: A Singer’s Expression of Emotions in Sung Performance. Vocal fold physiology: Voice quality control pp. 217—229 (1994)