Proceedings of the Institute of Acoustics

SIMULATED VOCAL EMOTIONS: AN ACOUSTIC STUDY OF TWO PORTRAYALS

S. P. Whiteside

Department of Human Communication Sciences

1. INTRODUCTION

Speech and voice characteristics are shaped by a wide variety of complex and inter-related factors. These complex and inter-related factors include stress [1, 2] and human vocal emotion [3, 4]. It is widely documented that different emotions can be signalled and communicated through a speaker's vocal characteristics. See [4, 5 and 6] for reviews of the literature on vocal emotions.

Research into how naturalistic situations shape vocal emotions and the characteristics associated with them [7] would be preferred to the use of simulations by actors [3, 7, 8, 9, 10]. However, truly naturalistic data is difficult to obtain, because the collection of such data is often fraught with serious ethical and moral considerations.

The ecological validity of studying vocal emotion using actors has been questioned by some researchers [11]. However, some recent studies into the vocal expression of emotion [3, 9, 10, 12], have demonstrated that actors are able to simulate vocal emotions, which are on the whole, successfully decoded by listeners at better than chance levels. Banse and Scherer [3] for example, found that from a set of 14 emotions portrayed by actors (hot anger, cold anger, panic fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust and contempt), that the only two emotions which were not accurately identified, were shame and disgust. They suggest that these two emotions rely more on visual rather than vocal cues [3]. In the case of shame for example, people are likely to avoid speaking while in this emotional state and therefore there are few perceptually salient cues which listeners would associate with this emotion. With regard to disgust, the signal of this emotion often takes the form of brief affect bursts (e.g. "yuck!") rather than a speaker adopting a particular vocal setting and/or voice quality [3].

Anger, fear, sadness, joy and disgust are among the emotions that have been frequently studied and their acoustic characteristics have been documented in a number of sources [5, 6, 9, 10, 13, 14]. The main acoustic cues that have been documented as characterising these emotions can be summarised as follows.

- Anger: There are two types of anger that are described in the literature [e.g. 3]. These are hot anger and cold anger and although similar in quality, differ in intensity. The former, which is more intense in quality, is characterised by an increase in mean fundamental frequency (F0) and mean intensity; an increase in F0 variability and range; an increase in high-frequency energy, falling F0 contours and increased articulation rate. Acoustic characteristics which typify cold anger include: an increase in mean F0, an increase in mean intensity, an increase in high frequency energy and falling F0 contours.
- Fear: Fear is characterised by an increase both in mean F0 and F0 range and by an increase in both high frequency energy and articulation rate;
- Sadness: This emotion is characterised by a decrease both in mean F0 and F0 range; a decrease in mean intensity, falling F0 contours and decreases in high-frequency energy and articulation rate and lower levels of articulatory precision; and

- Joy or Elation: These emotions are characterised by an increase in the following: mean F0 and F0 range, F0 variability, mean intensity, and in some cases, an increase in high-frequency energy and articulation rate.

This study aims to examine a set of acoustic parameters for 70 sentences representing seven vocal emotions simulated by an actor and actress. The emotions simulated are neutral, hot anger, cold anger, happiness, sadness, interest and elation. The acoustic parameters investigated include: the overall power (dB) of long term average spectra of the sentences; overall mean energy; standard deviations of energy; mean fundamental frequency values; mean values of the first six formant frequencies and their corresponding frequency bandwidths; mean utterance durations and mean articulation rates. These acoustic parameters are presented and discussed within a framework that attempts to characterise the simulations. In addition, the similarities and differences between the portrayals of the actor and actress are profiled and discussed.

2. METHOD

2.1 Subjects

RT is a 27 year old female with a standard southern British English accent. RP is a 32 year old male, with a standard British English accent with some northern colouring. Both speakers are non-smokers and have normal hearing, speech and language. At the time of recording, RT and RP had a total of three and twelve years respectively, of amateur and professional acting experience.

2.2 Speech Material

The speech material used to elicit the emotional data consisted of five brief sentences (*Weigh your yellow ruler; Wheel your wallaroo away; Rule your row warily; Reel your wheel away; You will reel your wool*). The seven simulated emotions that were simulated were *neutral, cold anger, hot anger, happiness, sadness, interest* and *elation*. These emotions with the exception of *neutral,* were chosen from the set of emotions studied by Scherer and his colleagues [3, 9, 10] to represent a balance of different emotional strength, valence and activity.

2.3 Recording and digitisation

Several weeks prior to the recording session, both RT and RP were given the speech material to rehearse the portrayal of the seven emotions listed above. All recordings were carried out in a sound proof room using a SONY DAT recorder. All 168 sentences were digitised using a sampling rate of 10kHz onto a KAY Computerized Speech Lab (CSL) model 4300, which was used to carry out all the acoustic analyses and derive the acoustic parameters detailed below.

2.4 Acoustic analysis: acoustic parameters

A total of 18 parameters were derived for each of the 70 sentences (2 speakers x 5 sentences x 7 emotions) in the following way:

2.4.1 The power of long term average spectra (1 parameter). The mean power (dB) of long term average spectra was calculated using FFT analysis for the 70 sentences (frame length - 512 points, no pre-emphasis, Blackman window).

2.4.2 Mean energy and mean energy standard deviation (2 parameters). The mean energy (in decibels of sound pressure- dB SPL) of the speech samples was calculated using a 20ms frame length. The dB SPL are computed from the speech pressure waveform as 20 times the log of the square root of the energy which has been divided by the number of sampled data points in the frame. The standard deviation values calculated by this algorithm were subsequently used to derive the mean of the standard deviations to provide an indicator of the variability of energy.

2.4.3 Mean fundamental frequency (1 parameter). Mean fundamental frequencies were obtained for each phrase using an autocorrelation method and a frame size of 20 ms with a 20 ms frame advance.

2.4.4 Formant frequencies and formant frequency bandwidths (12 parameters). Mean formant frequencies (F1 to F6) and formant frequency bandwidths (B1 to B6) were obtained for each of the 168 sentences. This was done using the full formant histories of an LPC analysis (10 ms frame length, 12th order filter, with pre-emphasis - 0.9, with triangular window weighting, using the autocorrelation method, with a bandwidth cut off of 500 Hz).

2.4.5 Utterance duration and articulation rate (2 parameters). Utterance durations were measured in milliseconds. These durations were used to calculate articulation rate as the number of syllables articulated per second.

3. RESULTS

Mean and standard deviation values of the acoustic parameters for simulations are given in Tables 1 and 2 for actress RT (F) and actor RP (M), respectively. The data given in Tables 1 and 2 displayed both similarities and differences between the portrayals of the two actors. These similarities included:

- for LTA: an increase for hot anger and elation, when compared to neutral and cold anger; an increase for interest and elation compared to sadness; a decrease for happiness, sadness and interest compared to cold anger, and a decrease for sadness compared to happiness;
- for E: an increase for hot anger compared to neutral, cold anger and sadness; an increase in for hot anger compared to neutral; a decrease for sadness compared to neutral, cold anger, hot anger and happiness;
- for (Esd): an increase for interest and elation compared to sadness; a decrease for sadness compared to hot anger and happiness; and
- for F0: an increase for elation compared to sadness; a decrease for elation compared to neutral and cold anger.

The differences between the simulations of RT and RP on the other hand included:

- sadness being signalled by a significant increase in utterance duration and therefore a decrease in articulation rate when compared to hot anger and happiness for speaker RP, which was not found in the data of speaker RT; and
- happiness being signalled by a significant increase in F2 when compared to neutral for speaker RT, which was absent in the data of speaker RP.

SIMULATED VOCAL EMOTIONS

Parameter	Simulated Emotion						
Mean s.d.	Neutral	Cold Anger	Hot Anger	Happiness	Sadness	Interest	Elation
LTA (dB)	-7.2 2.4	-2.9 2.1	11.3 2.3	3.0 2.7	-8.2 1.6	3.1 2.3	5.8 1.9
Mean energy (dB)	53.3 1.6	55.5 1.6	65.0 .9	59.9 1.3	47.7 1.4	59.3 1.2	59.5 1.2
Energy s.d. (dB)	5.8 2.1	6.3 1.2	10.2 .4	8.1 .8	4.3 .5	7.4 1.0	7.6 .7
F0 (Hz)	188.8 4.2	188.0 5.0	171.7 9.2	172.9 12.9	183.3 6.7	167.9 13.2	156.5 14.2
F1 (Hz)	462.6 26.9	456.9 24.9	516.3 21.9	482.4 18.5	482.5 16.4	470.5 5.1	516.1 8.1
B1 (Hz)	76.8 13.7	89.4 11.3	80.7 12.9	77.6 18.7	80.5 8.8	103.7 27.2	80.7 6.1
F2 (Hz)	943.9 86.8	1207.0 73.3	1183.6 38.4	1183.0 86.6	1295.8 98.4	1116.1 65.6	1291.3 118.1
B2 (Hz)	195.8 25.9	250.4 23.3	251.2 22.0	279.9 35.1	271.4 43.5	244.8 36.0	289.0 16.9
F3 (Hz)	2107.6 33.2	2319.7 99.2	2166.8 69.2	2179.1 125.0	2347.8 61.7	2181.1 50.7	2258.9 111.6
B3 (Hz)	263.6 31.1	231.8 34.5	245.7 36.6	221.9 39.3	298.3 17.4	219.3 25.6	212.8 25.1
F4 (Hz)	3018.6 29.5	3189.7 32.2	3044.0 90.8	3119.9 99.8	3298.3 51.8	3034.8 56.0	3105.6 99.7
B4 (Hz)	197.3 20.1	225.7 30.1	216.8 22.6	214.1 25.8	259.6 16.2	220.1 36.4	227.5 16.1
F5 (Hz)	3880.2 21.3	3925.3 24.1	3805.9 62.2	3894.5 82.7	4053.5 50.6	3737.6 68.4	3801.0 91.6
B5 (Hz)	243.7 22.4	261.2 11.1	272.4 18.4	282.9 19.4	254.2 19.8	258.8 34.5	226.6 9.7
F6 (Hz)	4360.4 53.9	4288.4 23.7	4269.2 36.9	4323.9 46.1	4389.0 27.0	4293.5 32.3	4222.2 42.4
B6 (Hz)	288.3 25.2	249.3 29.0	226.6 28.7	214.9 36.05	266.0 54.3	270.2 78.5	226.3 34.2
Duration (ms)	1126.4 38.4	1184.6 122.0	1416.6 105.3	1221.0 104.5	1245.8 181.4	1156.0 116.1	1196.8 100.4
AR (sylls./s)	5.1 .6	4.9 .5	4.1 .6	4.8 .5	4.7 .3	5.0 .7	4.8 .5

Table 1. Mean and standard deviation values of acoustic parameters: speaker RT (F).

Parameter	Simulated Emotion						
Mean s.d.	Neutral	Cold Anger	Hot Anger	Happiness	Sadness	Interest	Elation
LTA (dB)	0.3 1.9	3.0 1.4	9.3 2.0	4.4 2.3	-8.6 1.7	2.5 2.0	9.3 1.3
Mean energy (dB)	56.2 1.8	57.7 1.6	60.8 .71	59.9 1.39	46.9 2.3	58.1 1.2	63.0 1.2
Energy s.d. (dB)	6.2 .9	7.1 .8	7.6 .8	7.1 .4	3.5 .9	6.5 .9	7.5 .7
F0 (Hz)	133.6 1.8	141.0 10.7	153.3 6.8	158.0 10.2	112.5 11.4	151.0 15.7	151.1 6.6
F1 (Hz)	411.9 30.5	404.9 28.1	443.2 31.0	415.2 30.7	385.5 31.8	419.1 33.5	492.4 30.1
B1 (Hz)	56.7 11.0	71.3 2.9	103.6 12.1	68.3 3.2	82.6 29.5	74.0 10.7	98.1 9.8
F2 (Hz)	1259.5 65.0	1209.9 75.2	1274.9 77.8	1252.4 88.2	1318.5 94.9	974.9 481.3	1291.4 91.3
B2 (Hz)	173.2 31.0	208.2 45.0	153.8 21.5	161.6 41.9	207.3 28.2	234.7 24.3	192.4 37.9
F3 (Hz)	2276.4 66.3	2207.5 19.6	2138.4 44.8	2172.8 82.7	2349.2 70.7	2210.6 30.9	2029.8 41.0
B3 (Hz)	250.6 21.4	243.7 23.5	228.1 14.41	210.4 20.1	247.8 22.5	261.9 47.5	230.6 37.4
F4 (Hz)	2940.3 60.2	2943.8 40.4	2851.5 52.0	2902.6 93.0	3085.4 98.9	2985.1 40.6	2842.2 63.4
B4 (Hz)	162.2 21.3	222.0 15.0	216.9 11.7	185.9 11.0	241.1 12.7	200.6 14.6	216.9 28.6
F5 (Hz)	3649.2 75.6	3673.9 97.6	3513.8 43.8	3499.7 58.6	3943.8 70.5	3731.0 36.5	3509.8 32.4
B5 (Hz)	266.6 32.9	215.9 15.0	225.8 20.5	206.1 13.3	279.9 27.7	253.6 20.4	198.5 28.0
F6 (Hz)	4250.9 84.2	3434.9 39.6	3997.4 94.9	4158.4 141.1	4356.0 27.7	4146.8 69.9	4003.4 103.4
B6 (Hz)	361.4 55.9	294.3 56.1	311.4 89.0	505.1 69.2	355.2 40.8	345.7 25.2	367.5 82.3
Duration (ms)	1616.2 187.0	1855.6 138.9	1492.4 111.7	1432.4 109.7	2040.0 246.9	1840.0 162.4	1734.0 252.1
AR (sylls./s)	3.6 .2	3.1 .5	3.9 .4	4.1 .5	2.9 .3	3.2 .3	3.4 .3

Table 2. Mean and standard deviation values of acoustic parameters: speaker RP (M).

In order to characterise the general pattern of these similarities and differences between emotion and the actors' simulations, the mean z-scores of all 18 parameters were plotted according to emotion. These plots can be found in Figure 1 and Figure 2 for speakers RT and RP, respectively. These plots suggest that for speaker RT (Figure 1), the most distinct emotions are sadness and hot anger in terms of the number, degree and excursion in z-scores representing the acoustic parameters of these simulations. For speaker RP (Figure 2), sadness, elation and happiness were the most salient.

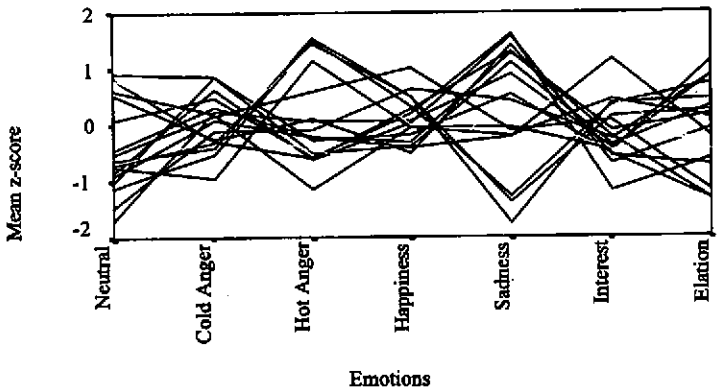

Figure 1: Mean z-scores for speaker RT (F) by emotion

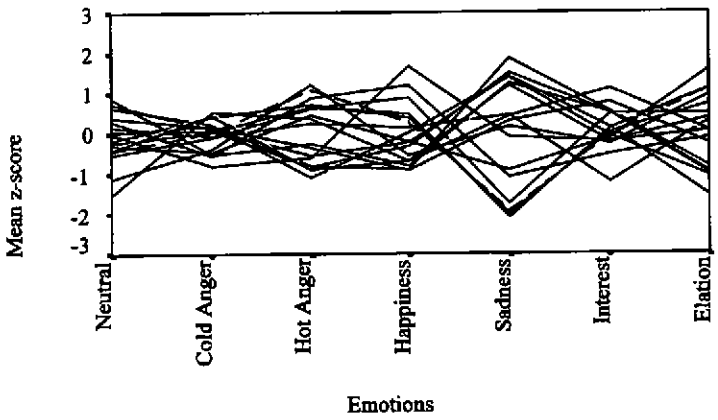

Figure 2: Mean z-scores for speaker RP (M) by emotion.

The 18 acoustic parameters for all 70 sentences and seven emotions were used to perform step-wise discriminant analysis for the simulations of both speakers. The results of these are given in Table 3 for both speakers RT and RP. Table 3 shows that with regard to the discriminant analysis, the acoustic measures for speaker RP were successful in giving a 100% correct classification for all seven emotions. For speaker RT the result is a little different, with cold anger being correctly classified 80% of the time with 20% being classified as interest, therefore giving an overall correct classification rate of 97.1%.

Actual Group	Predicted/ judged Group Membership (%)						
	Neutral	Cold Anger	Hot Anger	Happiness	Sadness	Interest	Elation
Neutral	100^a 100^b	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b
Cold Anger	0^a 0^b	100^a 80^b	0^a 0^b	0^a 0^b	0^a 0^b	0^a 20^b	0^a 0^b
Hot Anger	0^a 0^b	0^a 0^b	100^a 100^b	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b
Happiness	0^a 0^b	0^a 0^b	0^a 0^b	100^a 100^b	0^a 0^b	0^a 0^b	0^a 0^b
Sadness	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b	100^a 100^b	0^a 0^b	0^a 0^b
Interest	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b	0^a 0^b	100^a 100^b	0^a 0^b
Elation	0^a 0^b	0^a 0^b	0^a 0^b	20^a 0^b	0^a 0^b	0^a 0^b	100^a 100^b

Table 3. Results of discriminant analysis for speakers RT (F) and RP (M).

4. DISCUSSION

The data presented in this study clearly suggest that when simulating vocal emotion, actors adopted both similar and idiosyncratic strategies to produce a target emotion. Sadness for example, is associated with low levels of activity and energy. The result of this low level of activity is seen in some form within the portrayals of both actors with a general decrease in LTA, E and Esd and a general increase in formant frequencies and bandwidths when compared to other emotions. These data agree with Laukkanen et al. [8] who found low levels in sound pressure level (SPL) and small changes in SPL in the simulations of nonsense syllables representing sadness. The lower articulation rate associated with sadness [5, 6, 7, 9, 10] is portrayed in RP's data but not RT's data. This difference in portrayals replicates the findings of others that have also reported that individual differences exist in vocal parameters expressing emotion [3, 7, 8]. The distinct nature of sadness for both RT and RP however, are illustrated in Figures 1 and 2.

Another other basic emotion is anger which has a rather different degree of activity compared to sadness. It presents different acoustic profiles for both its forms. The more controlled version of cold (seething) anger is similar in quality to the explosive uncontrolled hot (raging) anger, but differs in intensity. In the simulations here, both actors had acoustic profiles that signalled high levels of activity and energy, but these were exaggerated in the portrayals of hot anger. This pattern replicates previous findings and

highlights the need for studies to distinguish between these two emotions and therefore specify what form of anger is being portrayed [3]. Figures 1 and 2 illustrate the differences in intensity.

These preliminary data suggest that the simulated portrayals of vocal emotions result in a varied and complex range of acoustic characteristics. While some of these acoustic characteristics are common to the simulation of some of the basic emotions such as sadness and hot anger, there are differences that may be the result of idiosyncrasies and the different subjective interpretation of how a specific emotion should be portrayed. What the study also highlights, is that the acoustic characteristics of vocal emotions, are determined by aspects of voice quality, pitch and timing.

5. ACKNOWLEDGMENTS

I wish to thank RT and RP for their co-operation and help with this study. The speech data were collected by the author within the EC TIDE project TP1174.

6. REFERENCES

1. Murray, I. R., Baber, C., & South, A. "Towards a definition and working model of stress and its effects on speech", *Speech Communication*, 20, 1996, 3-12.
2. Ruiz, R., Absil, E., Harmegnies, B., Legros, C., and Poch, D. "Time- and spectrum-related variabilities in stressed speech under laboratory and real conditions", *Speech Communication*, 20, 1996, 111-129.
3. Banse, R., & Scherer, K. R., "Acoustic profiles in vocal emotion expression", *Journal of Personality and Social Psychology*, 70, 1996, 614-636.
4. Murray, I. R., & Arnott, J. L. "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *Journal of the Acoustical Society of America*, 93, 1993, 1097-1108.
5. Scherer, K. R. "Vocal affect expression: a review and a model for future research", *Psychological Bulletin*, 99, 1986, 143-165.
6. Scherer, K. R. "Vocal correlates of emotional arousal and affective disturbance", In H. L. Wagner & A. S. R. Manstead (Eds.) *Handbook of Social Psychophysiology*, John Wiley, Chichester, 1989.
7. Williams, C. E., and Stevens, K. N. "Emotions and speech: some acoustic correlates", *Journal of the Acoustical Society of America*, 52, 1972, 1238-1250.
8. Laukkanen, A. M., Vilkman, E., Alku, P., & Oksanen, H. "Physical variations related to stress and emotional state: a preliminary study", *Journal of Phonetics*, 24, 1996, 313-335.
9. Scherer, K. R. "How emotion is expressed in speech and singing", *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 3, 1995, 90-96.
10. Scherer, K. R. "Expression of emotion in voice and music", *Journal of Voice*, 9, 1995, 235-248.
11. Greasley, P., Setter, J., Waterman, M., Sherrard, C., Roach, P., Arnfield, S., and Horton, D. "Representation of prosodic and emotional features in a spoken language database", *Proceedings of the XIIIth International Congress of Phonetic Sciences*, 1, 1995, 242-245.
12. Johnstone, T., Banse, R., & Scherer, K. R. "Acoustic profiles of prototypical vocal expressions of emotion", *Proceedings of the International XIIIth Congress of Phonetic Sciences*, 4, 1995, 2-5.
13. Pittam, J., & Scherer, K. R. "Vocal expression and communication of emotion", In M. Lewis & J. M. Haviland (Eds.), *Handbook of Emotions*, New York: Guilford Press, New York, 1993.
14. Scherer, K. R., Banse, R., Wallbott, H. G., and Goldbeck, T. "Vocal cues in emotion encoding and decoding", *Motivation and Emotion*, 15, 1991, 123-148.