Acoustic Markers of Emotions Based on Voice Physiology

Sona Patel[1], *Klaus R. Scherer*[1], *Johan Sundberg*[2], *Eva Björkner*[2]

[1] Swiss Center for Affective Sciences (CISA), University of Geneva, Switzerland
[2] KTH Dept. of Speech, Music and Hearing, Royal Institute of Technology, Sweden
Sona.Patel@unige.ch

Abstract

Acoustic models of emotions may benefit from considering the underlying voice production mechanism. This study sought to describe emotional expressions according to physiological variations measured from the inverse-filtered glottal waveform in addition to standard parameter extraction. An acoustic analysis was performed on a subset of the /a/ vowels within the GEMEP database (10 speakers, 5 emotions). Of the 12 acoustic features computed, repeated measures ANOVA showed significant main effects for 11 parameters. Subsequent principal components analysis revealed the three components that explain acoustic variations due to emotion, including "tension" (CQ, H1-H2, MFDR, LTAS) "perturbation" (jitter, shimmer, HNR), and "voicing" (fundamental frequency).
Index Terms: emotion, vocal expression, acoustic cues, voice quality, physiology, glottal waveform, affect bursts

1. Introduction

Much of the past research to identify the acoustic correlates to vocally expressed emotions has examined acoustic features such as fundamental frequency (*f*0), intensity, and duration that are easily accessible through standard speech analysis software. While a number of these parameters were found useful in differentiating among some emotions, such as the mean and variability of *f*0, emotion recognition algorithms have been mostly unsuccessful at classifying emotions into categories based on these basic measures alone. It has been suggested that these parameters may be useful in distinguishing among emotions on the arousal dimension (ranging from highly alert and excited to relaxed and calm) [1], but not necessarily other emotions, such as those differing in valence (divides positive and negative emotions).

Although the classification results have been relatively poor, perceptual accuracy in identifying high arousal emotions, such as happy and angry, that differ according to valence have been generally high. Hence, it is likely that a unique acoustic pattern for the valence dimension exists, but the appropriate acoustic measures to describe this dimension have not been identified. This hypothesis motivated recent investigations using more complex acoustic measurements. Many of these parameters have been aimed at quantifying voice quality, such as the level difference between the first two harmonics, the Hammarberg Index, the harmonics-to-noise ratio (HNR), and various measures of spectral slope [2].

A number of studies have suggested that the valence dimension may represent changes in voice quality [3]. Voice quality refers to the perception of physical changes during vocal fold vibration and in vocal tract shape outside of the percepts of pitch, loudness, and phonetic category [4]. Considering that research has shown a high correspondence of pitch and loudness with other dimensions such as arousal and potency, we predict that the inclusion of voice quality measures will improve the acoustic descriptions of emotions and hence classification accuracy.

Certain vocalizations may be optimal for the study of affective variations in voice quality. According to the Component Process Model proposed by Scherer [5], vocal expressions are influenced by socio-culturally determined norms or "pull" factors in addition to physiologically instigated "push" factors. Most emotional expressions are shaped by social display rules—pull effects—and as a result, have specific prosodic patterns. Although expressions arising from a purely physiological response do occur, i.e., push effects, these are infrequent and often brief. Push effects are often expressed as "affect bursts" or sudden and spontaneous vocalizations.

While most emotion research has focused on understanding the acoustic characteristics of sentence- or phrase-length samples, possibly to increase the ecological validity of the results, these are highly contextualized and affected by pull factors. These expressions may have a strong pull towards a specific acoustic pattern as dictated by social expectations that may potentially override the vocal behavior pushed by a physiological response to an emotional state. Therefore, measurements of voice quality resulting from the physiological response may be overridden by other acoustic factors. In addition, the use of sentence length stimuli does not easily lend itself to a global analysis of voice quality because many voice quality measurements cannot be accurately performed on time-varying segments of speech. It is also difficult to isolate the paralinguistic acoustic variations from linguistic variations.

To understand the physiological effects on voice quality, expressions resulting largely from push effects may be examined. Further study of affect bursts may help reveal how physiological changes in the speech production mechanism result in a variety of emotional expressions. An approximation of affect bursts may be achieved using isolated vowels, as in interjections. Vowels expressed in isolation are ideal as these are minimally affected by contextual factors such as coarticulation due to the location of the vowel within a word and changes in prominence due to the location of the word within a sentence. Produced in the form of affect bursts, vowels represent ecologically valid vocalizations.

Hence, the purpose of the present study was to investigate the extent to which the expression of emotions is mediated by voice quality. This was achieved using physiologically based measures obtained from the glottal flow waveform, since recent research has suggested that glottal information may be useful in describing differences in emotional expressions. To obtain suitable samples that more closely resembled physiologically inspired affect bursts, samples of the vowel /a/ were portrayed by professional actors for five emotional contexts. Acoustic differences across emotions were examined using a repeated measures design of speaker-normalized measures. Then a principal components analysis was

performed to identify the groups of variables underlying the emotional expressions driven by physiological push effects.

2. Methods

The stimulus materials were previously collected as part of the Geneva Multimodal Emotional Portrayal (GEMEP) database, a multimodal database consisting of facial, vocal, and gestural-postural samples for each recording across speakers [6]. An acoustic analysis of a subset of these samples was performed using measures from the time waveform, spectrum, and glottal waveform. The following sections provide a brief overview of the samples used in this study, a description of the acoustic features examined, and an explanation of the speaker normalization technique.

2.1. Speech stimuli

As part of the GEMEP database, 10 professional, French-speaking actors (5 male, 5 female) expressed the vowel /a/ in 12 emotional contexts. The open vowel /a/ was selected since it is minimally affected by the articulators. A subset of these emotions were selected for use in the present study, including "relief," "sadness," "joy," "panic fear," and "hot anger." These emotions were chosen to represent strong differences along the arousal and valence dimensions (see Table 1). In addition, they also differ on the power or potency dimension, particularly between hot anger and panic fear.

Table 1. *Emotion differences in arousal and valence.*

		Valence	
		Positive	Negative
Arousal	High	Joy (Elation)	Panic fear Hot anger
	Low	Relief	Sadness

The actors were given written scenarios to help invoke the emotion during an interaction with a professional stage director. They were asked to express each emotion using the /a/ vowel in addition to sentence length material. While this study used acted speech, the use of an interjection-like /a/ sound was intended to facilitate the expression of more natural affect bursts originating from push effects, rather than the stereotyped expressions biased by socio-cultural display rules. In addition, the use of vowel-length stimuli avoids contextual biases due to coarticulation and variations in stress patterns that may occur if a vowel is extracted from a sentence.

Each actor expressed each emotion twice, resulting in 100 samples (10 speakers X 5 emotions X 2 repetitions). All samples were digitally recorded using a head-mounted microphone (Sennheiser) at a sampling rate of 41 kHz.

2.2. Acoustic features

Twelve acoustic parameters were extracted from the speech samples. The Corr autocorrelation software was used to compute the equivalent sound level (Leq), the mean *f*0 (mf0), and the long-term average spectrum between 0 and 6700 Hz. From the long-term average spectrum the alpha ratio (alpha), the difference between the amplitudes of the first and second harmonics (H1-H2), and the level difference between H1 and H2 based on the long-term average spectrum (LTAS) were measured. The alpha ratio, a measure of the rate of glottal adduction, is the ratio between the summed energy in the 50-1000 kHz to the summed energy in the 1-5 kHz band. Higher values would indicate a shallower spectral slope and a faster closing speed. The H1-H2 and LTAS parameters were measures of the spectral tilt or open quotient.

Then, an inverse filtering algorithm was applied using Decap software (Svante Granqvist) in order to obtain the glottal waveform. Four parameters were measured from the glottal waveform including the pulse amplitude (PAmp), the maximum flow declination rate (MFDR), the normalized amplitude quotient (NAQ), and the closed quotient (CQ). CQ is the proportion of the closed phase of the glottal cycle relative to the total cycle time. This value increases with faster glottal adduction. MFDR is the absolute value of the most negative point of the derivative of the glottogram. Therefore, this parameter measures the rate of vocal fold adduction. NAQ is computed as the pulse amplitude divided by the product of the fundamental period and the MFDR. This measure also quantifies characteristics of glottal adduction [7].

Finally, Paul Boersma's Praat software was used to obtain the perturbations measures jitter and shimmer, as well as the HNR. Jitter and shimmer are measures of the cycle-to-cycle variations in frequency and amplitude and are commonly used to measure voice quality in disordered voices. Similarly, the HNR is a measure of the presence of turbulent noise energy as occurs with increased breathiness.

2.3. Normalization

A normalization was performed to compare the acoustic variations across emotions for each speaker. Previous research typically examines the difference in acoustic measures between each emotion and a "neutral" emotion. Since a "neutral" emotional expression was not recorded (as forced neutrality tends to produce unnatural vocalizations), a set of "baselines" were calculated for each speaker instead. These baselines were defined for each parameter as the mean value across all expressions for each speaker. Hence, 12 baselines were formed for each speaker, one for each parameter. In most cases the average was computed across 10 samples (5 emotions X 2 repetitions); however, it was not possible to compute some of the parameters for all samples due to technical limitations of the recordings (essentially due to flat acoustic waveforms, especially for the low amplitude sadness portrayals, which did not allow the computation of an inverse filtered solution).

The normalized measures for each acoustic feature were computed as the difference between each speaker's baseline and the mean of the two raw samples. This resulted in 50 cases (1 mean sample X 10 speakers X 5 emotions) for each of the 12 parameters. These values represented the extent of deviation from the speaker's baseline for each emotion and for each acoustic parameter and enabled comparisons of the direction and degree of acoustic variation across speakers.

3. Results

Once the acoustic measures were computed and normalized by speaker, statistical analyses were performed using the normalized values. A repeated measures ANOVA was completed followed by a principal components analysis. Additional hypothesis testing was performed to determine whether the expressions acoustically differed according to valence. The results are described in the following sections.

3.1. Repeated measures ANOVA

To evaluate whether the mean deviation from baseline differed across emotions for each parameter, a separate repeated measures ANOVA was performed in SPSS (v.17; SPSS Inc., Chicago, IL) for each of the 12 acoustic parameters with "emotion" as the within subject factor. The results of Mauchly's test of sphericity was violated for the main effects of Leq ($\chi^2(9) = 28.776$ $p < .05$), mf0 ($\chi^2(9) = 20.683$, $p < .05$), PAmp ($\chi^2(5) = 14.763$, $p < .05$), and jitter ($\chi^2(9) = 27.560$, $p < .05$). Therefore, the Greenhouse-Geisser estimates of sphericity ($\varepsilon = .576, .480, .548$, and $.502$, respectively) were used to correct the degrees of freedom. Results revealed a significant main effect of emotion on all acoustic parameters at $p < .05$ (minimum *df*, $F = 3, 1.80$) except *NAQ*.

Bonferroni corrected pairwise comparisons showed that a number of acoustic parameters were useful in differentiating between two or more emotions. The acoustic cues for which each pair of emotions differed significantly are shown in Table 2. Some parameters were useful in differentiating among most emotions such as mf0, whereas others such as H1-H2 were useful in separating one pair of emotions. To interpret these results a principal components analysis was performed. This technique clusters similar parameters together and describes them as a single "factor."

Table 2. *Bonferroni adjusted pairwise comparisons between emotions. Acoustic cues listed for each pair of emotions were significantly different at p < 0.5.*

	Sadness	Joy	Panic fear	Hot anger
Relief	Alpha	Leq, Alpha, LTAS, mf0, MFDR, Jitter, Shimmer	Leq, CQ, Alpha, LTAS, mf0, HNR, MFDR, H1-H2, Shimmer	Leq, Alpha, LTAS, mf0, Pulse Amp, MFDR
Sadness	-	Leq, Alpha, mf0	Leq, Alpha, LTAS, mf0, Shimmer	Leq, Alpha, LTAS, mf0
Joy		-	None	Pulse Amp, Shimmer
Panic fear			-	mf0, HNR, Shimmer

3.2. Principal components analysis

A principal components analysis (PCA) was performed in SPSS using the speaker-normalized measures for the 11 variables. The measures for NAQ were not included, since no significant results were found for this parameter. As this analysis does not tolerate missing values, 12 cases were dropped from the analysis (due to the limitations in extracting the glottal source parameters mentioned above).

The results suggested that the 11 variables can be best described by three underlying components (accounting for 83.5% of the variance). The number of components was selected based on the location of the "elbow" of the scree plot, i.e., the point after which the decrease in the Eigenvalue levels is small. An orthogonal rotation was applied prior to computing the component loadings. An analysis of the rotated

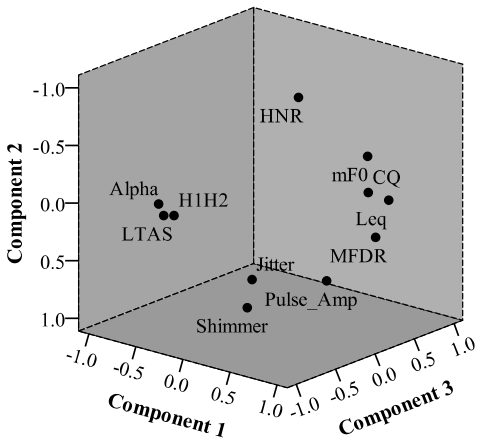

Figure 1: *Variable loadings in the component space.*

component loadings matrix showed that the first component is marked by the variables CQ, H1-H2, MFDR, and LTAS, all features that see related to vocal *tension*. The second component shows high loadings of shimmer, HNR, and jitter, and can thus be linked to phonation *perturbation*. Only mf0 loads highly on the third component, which can thus be seen as a frequency of *voicing* component. These loadings are shown in a three-dimensional component space in Figure 1.

In order to evaluate the three-component representation of the emotions, equations to describe each component or dimension were derived from the component coefficient matrix. The parameters that loaded highly on each component and their corresponding coefficients were averaged to provide a more stable acoustic representation of each component and allow replication across studies independent of sample-specific optimization. These are as follows:

$$C1 = .321 * mean\,(CQ, MFDR) - .245 * mean\,(H1H2, LTAS) \quad (1)$$

$$C2 = .293 * mean(Shimmer, Jitter) - .297 * HNR \quad (2)$$

$$C3 = .628 * mf0 \quad (3)$$

Using these equations, a composite score could be computed for each sample and graphically represented in the acoustic-component space (Figure 2).

The plot suggests that Component 1 represents the degree of underlying sympathetic arousal, separating relief from joy, hot anger, and panic fear. Component 2 mainly differentiates hot anger from the remaining emotions, and may thus represent the dimension of power or potency. Finally, Component 3 mainly separates relief and panic fear, which may reflect an "ability to control" dimension.

3.3. Controlling for arousal

One of the predictions of the present research is that the effects of valence are often not apparent because they are masked by the dominance of arousal. To overcome this limitation, it was necessary to keep the arousal level constant in order to identify any acoustic differences according to valence. Paired samples t-tests were conducted to separately compare the differences between the two low arousal emotions and the three high arousal emotions for the variables corresponding to the three components. Results showed a significant difference between joy and hot anger in shimmer and HNR ($t(9)=-6.051$, $p=0.000$;

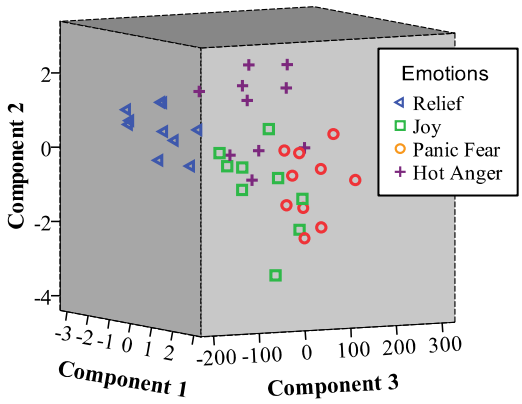

Figure 2: *Three-dimensional representation of the emotional samples in acoustic-component space.*

t(9)=3.129, p=0.012). This suggests that the perturbation measures of Component 2 may be important in indexing differences between high arousal emotions. However, it is not clear whether Component 2 separated the two high arousal emotions with respect to valence, since they also differ in potency, and Component 2, on the whole, seems to be linked to the potency dimension.

4. Discussion

To understand emotion expression in speech, it may be necessary to identify the acoustic effects of push and pull factors. Typically, past research has examined expressions that are dominated by pull factors. Many of these studies found acoustic features to describe emotions that differ in arousal. While some researchers have reported acoustic cues that may distinguish between emotions on the valence and power dimensions, these cues are inconsistent across studies. It has recently been suggested that the valence dimension might be described by voice quality cues. However, it has been generally difficult to determine the relevant voice quality cues because of the overpowering effects of arousal.

In this study we attempted to explore the effect of the vocal portrayals of different emotions on voice quality using vowel samples mostly driven by push effects. This approach was intended to minimize the influence of pull effects, thereby facilitating measurement of emotion-specific voice characteristics. An analysis of the glottal waveform was performed to enable the characterization of physiological aspects of vocal fold vibration without the influence of the vocal tract. Principal components analysis of the normalized acoustic measurements revealed that a subset of these features, those related to rate and duration of glottal adduction, jointly characterized high arousal emotions, suggesting that sympathetic arousal results in high vocal tension.

Another interesting result is the discrimination between the high arousal emotions of joy and hot anger using perturbation measures (Component 2). These emotions are psychologically differentiated on both the valence and potency dimensions. While the results of this study cannot disambiguate whether this difference corresponds mostly to valence or potency, there is reason to assume that measures related to the irregularity of voicing may be indicative of potency [5]. Further, we must consider the possibility that the psychological distinction of valence may not translate to a unique acoustic dimension that is regulated by a particular physiologic mechanism. Finally, the importance of fundamental frequency is noteworthy. This feature had the single highest loading on Component 3 and may relate to a psychological dimension of "ability to control." While the terms "control" and "power" have been used synonymously to refer to a single emotion dimension, these two factors are differentiated in the Component Process Model [5]. It is possible that the variability in findings for the power dimension may be partially due to the combination of two factors that can uniquely co-vary.

5. Conclusions

The research literature on vocally expressed emotions has not been able to conclusively identify the acoustic markers for all emotions. This has prompted the exploration of other factors, particularly voice quality features. However, attempts to describe vocal expressions according to physiological mechanisms have been met with mixed success. A number of factors may have influenced the findings, particularly the extent to which the expressions were influenced by conventional pull rather than physiologically driven push factors. In this work, using an inverse filtering approach for actor-produced /a/ vowels, the results of the statistical analyses of five emotions suggest that voice quality is indeed an important aspect of emotional speech. We have linked the extracted parameters to three components of the physiological mechanisms involved in vocal expression: tension, perturbation, and voicing. While these results are preliminary, they suggest that vocal source characteristics strongly influence the acoustic output, particularly in affect bursts or expressions inspired by push effects, and may thus account for the ease with which human judges recognize vocally expressed emotions. We suggest that such source parameters should have a major role in further research.

6. Acknowledgements

This research was supported by a grant to Klaus Scherer from the Swiss National Science Foundation (No. 101411-100367).

7. References

[1] Banse, R. and Scherer, K.R. "Acoustic profiles in vocal emotion expression," Journal of Personality and Social Psychology 70(3): 614-636, 1996.

[2] Tamarit, L., Goudbeek, M., and Scherer, K. R. "Spectral slope measurements in emotionally expressive speech," In SPKD-2008, Paper 007, 2008.

[3] Ladd, D., Silverman, K., Tolkmitt, F., Bergmann, G., and Scherer, K. "Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect," Journal of the Acoustical Society of America, 78:435-444, 1985.

[4] Titze, I. R. Principles of voice production, Englewood Cliffs, N.J.: Prentice Hall, 1994.

[5] Scherer, K. R. "Vocal affect signalling: A comparative approach," In J. Rosenblatt, C. Beer, M.-C. Busnel, & P. J. B. Slater (Eds.), Advances in the Study of Behavior (Vol. 15, pp. 189-244). New York: Academic Press, 1985.

[6] Banziger, T. and Scherer, K. R. "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," In A. Paiva, R. Prada, and R.W. Picard, editors, Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738, 476–487. Springer-Verlag Press, Berlin, Germany, 2007.

[7] Bjorkner, E., Sundberg, J., Cleveland, T., and Stone, E. "Voice Source Differences Between Registers in Female Musical Theater Singers," Journal of Voice, 20(2):187-196, 2006.