A Preliminary Model of Emotional Prosody using Multidimensional Scaling

Sona Patel[1], *Rahul Shrivastav*[2]

[1] NCCR Center for Affective Sciences (CISA), University of Geneva, Switzerland
[2] Department of Speech, Language and Hearing Sciences, University of Florida, USA
Sona.Patel@unige.ch, Rahul@ufl.edu

Abstract

Models of emotional prosody based on perception have typically required listeners to rate emotional expressions according to the psychological dimensions (arousal, valence, and power). We propose a perception-based model without assuming that the psychological dimensions are those used by listeners to differentiate emotional prosody. Instead, multidimensional scaling is used to identify three perceptual dimensions, which are then regressed onto a dynamic feature set that does not require a training set or normalization to a speaker's "neutral" expression. The model predictions for Dimensions 1 and 3 closely matched the perceptual model; however, a moderately close match observed for Dimension 2.
Index Terms: emotional speech, affective prosody, vocal expression, computational model, voice quality, perception

1. Introduction

Evidence suggests that emotions can be described according to one or more underlying dimensions, where each dimension describes a continuous variation in a set of acoustic features. An emotion can then be described by its magnitude on each of these dimensions. A number of studies suggest that only two to three dimensions are required to capture the most relevant properties of vocally expressed emotion categories [1-2]. However, there is still considerable disagreement in both the number and nature of these dimensions. Nevertheless, some trends are apparent. The acoustic correlates to the "arousal" dimension are the most robust. This dimension has been described as the level of physical arousal or emotional energy in the expression. The most common features reported to distinguish emotions by arousal are mean and variability of fundamental frequency (*f*0), mean intensity, and speaking rate [3].

Many experiments suggest the need for a second dimension, typically referred to as "valence" or pleasantness. This dimension separates the negative emotions from positive emotions. Emotions with positive valence may be described as having less high frequency energy and more *f*0 variability [4]. However, the acoustic cues to this dimension are inconsistent. In fact, a number of studies have failed to determine the acoustic characteristics of the valence dimension [3, 5]. Due to the uncertainty in the number of dimensions needed to perceptually separate the emotions, it is not clear whether weak acoustic relations result from an excess of dimensions or an incomplete acoustic feature set.

Other dimensions reported include power, dominance, confidence, and intensity. This potency dimension separates the strong emotions from the weak emotions. Some evidence suggests that emotions high in power can be characterized by a high mean intensity and large range of intensity; however, a strong trend is not clear. Some have suggested that the acoustic characteristics of the power dimension may overlap with arousal, rendering this dimension redundant [6].

Most listener-based models of emotional prosody using the dimensional approach assume that the psychological dimensions of arousal, valence, power, etc. are the dimensions that listeners use to distinguish between emotions. These studies obtain perceptual ratings along a number of psychological dimensions and correlate or regress the acoustic parameters onto these dimensions [3, 7]. For example, in the study by Pereira [3], listeners were asked to rate five emotional states (happiness, sadness, hot anger, cold anger, and a neutral state) using two scales for the arousal, pleasure, and power dimensions. The mean dimension ratings were then correlated with four acoustic measures (mean and range of *f*0, mean RMS energy, and duration). However, there is no direct evidence that the psychological dimensions are equivalent to the perceptual dimensions.

We proposed that for emotions expressed for the purpose of communication (i.e., expressions governed by "pull effects" in the model by Scherer [8]), perceptual judgments are the gold standard for measuring differences in emotions. Therefore, predictive models of emotions should match listener perception. In this study, a computational model of emotions based on perception is presented. Since the nature of the perceptual dimensions was not assumed (i.e., these may or may not be identical to the psychological dimensions), it was first necessary to develop a perceptual model. For this, perceptual distances between emotions were obtained from a large-scale, same-different discrimination task of 19 emotions [4]. Multidimensional scaling (MDS) was applied to determine 1) the number of dimensions needed to accurately represent the differences between emotions and 2) the configuration of the emotions in a multidimensional space. This technique has been successfully used in the past to determine the underlying dimensions of other perceptual phenomena such as voice quality [9] and emotional responses to music [10].

To develop an acoustic model of emotions that mirrors listener perception, stepwise regressions were used to identify the acoustic parameters that account for the most variance in the perceptual data. While regression methods (multiple or stepwise) have been previously used to relate acoustics to impressionistic judgments of emotions [1], in this model a unique acoustic feature set is also presented. This set consisted of dynamic measures that did not require normalization to a "neutral" emotion. Using a neutral emotion as a comparison tool may ease the computational analyses; however the ability to detect the emotional state of a speaker without the use of each speaker's neutral emotion has broader applications such as real-time emotion detection from speech.

2. Methods

A brief overview of the speech stimuli and perceptual test is provided. However, complete details of the emotion selection and recording procedures, rationale for the perceptual task,

and signal detection analysis may be found in [4].

2.1. Speech stimuli

The speech samples were obtained from two individuals (one male, one female), who were recruited from the School of Theater and Dance at the University of Florida. These participants were asked to repeat two nonsense sentences in 19 emotional contexts including funny, content, love, respectful, happy, exhausted, confident, confused, bored, suspicious, embarrassed, sad, surprised, interested, annoyed, angry, jealous, anxious, and lonely. These terms were selected through a preliminary data reduction procedure of 70 emotion terms [4]. Nonsense sentences were used to retain only the suprasegmental and syntactic information in the speech signal while minimizing emotional information cued through semantics. This type of material has been used by others [11] to maintain naturalness in the prosodic structure of the sentence as opposed to numbers, dates, or alphabet recitations, which maybe unnatural to produce with a lot of emotion.

The recording equipment included a head-mounted microphone (Audiotechnica, ATM21a) and external sound card (Creative E-MU 0202) connected in series with a PC. Stimuli were recorded at a sampling rate of 44.1 kHz with 16-bit quantization. After all recordings were made, each speaker selected his or her "best of two" expressions of each emotion. The recordings were saved as 38 individual files (2 speakers X 19 emotions X 1 best sentence/emotion).

2.2. Perceptual data

The perceptual data were collected in a same-different discrimination test reported in [4]. In short, 16 listeners were presented with two speech samples in each trial. Each pair of stimuli consisted of either two sentences having the same emotion (a matched trial) or different emotions (an unmatched trial). The listeners' task was to determine whether the two sentences presented in each trial conveyed a matched or unmatched pair by selecting the "same" or "different" buttons on the computer screen. Each listener discriminated all possible combinations of speaker and emotion pairs at least twice (completed in 6 sessions of 1 hour and 45 minutes each). All stimulus pairs were randomly presented binaurally at a comfortable loudness level using headphones (Sennheiser HD280Pro) that were connected to an E-MU 0202 external sound card of a PC. Stimulus presentation and response collection was controlled using Matlab.

Accuracy across listeners was reported in terms of *d'*, the Theory of Signal Detection measure of discrimination ability [13; pp. 223, Eqn. 9.8]. The *d'* score is a better measure of listener performance because this parameter accounts for the false alarm rates (proportion incorrect of matched trials) in addition to the hit rate, unlike percent correct scores which are essentially the hit rate (proportion correct of unmatched trials). This parameter is a measure of the perceived difference or "perceptual distance" between emotions. The *d'* scores were calculated for each pair of emotions and were then entered into a 19 by 19 similarity matrix.

2.3. Perceptual model development

The perceptual model was developed using the *d'* matrix. An MDS analysis was performed using this proximity matrix to determine the number of dimensions needed to represent the perceptual distances among emotion categories. The ALSCAL algorithm was used to determine the locations of various emotion categories within a Euclidean space. The perceptual space was constructed through shrinking and stretching the original *d'* measures to obtain a nonmetric (ordinal) solution with minimum dimensionality that satisfied the requirement of monotonicity [14]. Thus, emotions separated by large *d'* values were farther apart on one or more dimensions.

The optimal number of dimensions was determined by computing the R^2 and stress coefficients for six solutions differing in dimensionality (one through six dimensions). The stress parameter specifies how well the MDS solution corresponds to the actual *d'* scores. Smaller stress values indicate a better mapping of the distances. The solution that maximized the R^2 with a low stress measure was selected as the solution that sufficiently explains the variance in the data. The resulting stimulus coordinates of the emotions form the perceptual model.

2.4. Acoustic features

The acoustic features examined here included a limited set of global measures of *f*0 and intensity (minimum and maximum normalized to the mean), dynamic or time-varying measures of *f*0 (gross trend, number of *f*0 contour peaks normalized to the utterance duration, and peak rise and fall times) and intensity (syllable attack time and duty cycle of the syllables, normalized syllable attack to the peak duration and duty cycle), measures of duration (speaking rate, vowel-to-consonant ratio or VCR, pause proportion, and speaking rate trend), and measures of voice quality (cepstral peak prominence, alpha ratio of a stressed and unstressed vowel, slope of the long-term averaged spectrum of a stressed and unstressed vowel). Most parameters were measured automatically using scripts developed in Matlab; however, the pauses and vowel and consonant durations were manually defined prior to measurements on these segments. In addition, the *f*0 measurements were manually corrected.

2.5. Acoustic model development

The acoustic model was developed based on the perceptual model. SPSS was used to compute stepwise linear regressions to systematically select the set of acoustic measures (dependent variables) that best explained the emotion properties for each dimension (independent variable). A mixture of the forward and backward selection models was used, in which the independent variable that explained the most variance in the dependent variable was selected first, followed by the independent variable that explained the most of the residual variance. At each step, the independent variables that were significant at the 0.05 level were included in the model (entry criteria $p \leq 0.09$), and the predictors that were no longer significant were removed (removal criteria $p \geq 0.10$). The optimal feature set for each dimension included the minimum set of acoustic features needed to explain the perceptual changes for each dimension. The final regression equations, describing the acoustic features that corresponded with the perceptual model, formed the acoustic model.

3. Results

Since the development of the acoustic model was dependent on the results of the perceptual model, the latter is discussed first. This is followed by a description of the acoustic predictors for the perceptual data.

3.1. Perceptual model

The optimum dimensionality of the perceptual space was

determined using the R^2 and stress curves. The three-dimensional (3D) model was selected, since it accounted for 90% of the variance in perceptual judgments and provided a tolerable stress level of 0.12. A 3D ALSCAL model indicates that three distinct perceptual dimensions were sufficient to describe the differences among the 19 emotion categories. Each dimension corresponds to a unique set of suprasegmental cues that listeners used to differentiate among the 19 emotion categories. The emotion categories that were perceptually separated according to each dimension can be identified by arranging the emotion categories in order of their ALSCAL coordinates on each dimension as shown in Table 1. Emotions that were farther apart on a dimension were easier for listeners to discriminate. By examining the emotions that were well-separated on each dimension, it was possible to hypothesize the acoustic changes that may describe each dimension. The acoustic features used in the model development were derived from the acoustic changes hypothesized here.

Table 1. *MDS coordinates of the emotions arranged in ascending order for each dimension.*

Dimension 1		Dimension 2		Dimension 3	
Anxious	-2.04	Love	-1.34	Anxious	-1.45
Surprised	-1.73	Happy	-1.29	Sad	-0.78
Happy	-1.70	Sad	-1.04	Annoyed	-0.76
Funny	-1.25	Funny	-0.74	Embarrassed	-0.75
Confident	-1.00	Interested	-0.68	Suspicious	-0.50
Interested	-0.79	Content	-0.63	Confused	-0.35
Angry	-0.73	Lonely	-0.49	Content	-0.23
Confused	-0.32	Surprised	-0.46	Interested	-0.20
Annoyed	-0.29	Anxious	-0.38	Respectful	-0.14
Respectful	-0.20	Confused	-0.15	Confident	0.04
Content	-0.19	Exhausted	-0.04	Bored	0.10
Jealous	0.16	Embarrassed	-0.01	Lonely	0.30
Suspicious	0.70	Respectful	0.35	Jealous	0.36
Love	0.91	Suspicious	0.58	Happy	0.49
Exhausted	1.18	Confident	0.62	Surprised	0.51
Sad	1.21	Jealous	0.71	Angry	0.52
Bored	1.78	Bored	1.08	Love	0.55
Embarrassed	2.01	Annoyed	1.20	Exhausted	0.88
Lonely	2.30	Angry	2.71	Funny	1.42

3.2. Acoustic model

Separate stepwise regressions were performed for each dimension to determine the acoustic features that corresponded best to each dimension of the perceptual model. Thus, the acoustic model was also a 3D model. The regression analysis revealed significant parameters for each of the three dimensions (i.e., *D1*, *D2*, and *D3*), as shown below:

$$D1 = 14.586 - 0.017 * pnorMIN - 0.002 * alpha_{UNST} - 1.359 * rate \quad (1)$$

$$D2 = -2.057 + 1.044 * normattack - 0.155 * normpnorMIN + 8.563 * dutycycle \quad (2)$$

$$D3 = -2.618 + 10.732 * gtrend + .003 * mpkrise - 1.444 * rate + 2.435 * VCR - .209 * iN\max + .001 * alpha_{ST} \quad (3)$$

where $alpha_{ST}$ and $alpha_{UNST}$ is the alpha ratio of a stressed and unstressed vowel, respectively, *pnorMIN* is the normalized minimum pitch, *normpnorMIN* is *pnorMIN* normalized to the speaking rate, *normattack* is the normalized attack time of the intensity contour, *mpkrise* is the mean slope of pitch contour peaks, *iNmax* is the normalized maximum intensity, and *gtrend* is the *f*0 contour trend. The *D1* features accounted for 90% of the variance in Dimension 1 of the perceptual model. These parameters were essentially measures of speaking rate and intensity. The features composing *D2* accounted for 65.5% of the variance and may correspond to the staccato-like quality that separated anger from the other emotions. The feature set for *D3* included dynamic aspects of the *f*0 contour (trend and peak change), duration measures, and measures of voice quality. These parameters accounted for 83.3% of the variance (R^2 = 0.833). Together, these equations formed the acoustic model.

To determine how well the acoustic model represented the perceptual model, the "predicted" value of each emotion was computed using the acoustic model (Equations 1-3). The acoustic values were averaged across both speakers and then converted into standard scores (z-scores). The predicted values were regressed onto the "perceived" MDS values (also in standard scores). Scatterplots of the predicted-perceived scores are shown for each dimension in Figure 1. Equal x- and y-axis scales are shown for easily identifying the emotions falling on the y = x (perfect prediction) line. These graphs show that the predicted values for Dimension 1 closely matched the perceived values (R^2 = 0.877). The predicted values for Dimension 3 were also close to the perceived values, with the exception of a few emotions (e.g., removal of lonely increased the R^2 from 0.576 to 0.711). While highly correlated, the most scatter was observed for Dimension 2 (R^2 = 0.395).

4. Discussion

Numerous models of emotional prosody have been developed by training pattern recognition algorithms on a large set of expressions with many acoustic features. Fewer studies have developed models based on perception. Most of these studies involve obtaining listeners ratings on specific scales such as the psychological dimensions of arousal, valence, and power. Since perceptual judgments are the gold standard for the measurement of emotions, it is essential to have an accurate perceptual model for developing predictive models of emotional prosody. Such a model was constructed by a multidimensional scaling of discrimination judgments of 19 emotions. A three-dimensional model was identified and used to develop an acoustic model by regressing the acoustic parameters to the dimensional coordinates.

The arrangement of emotions on each dimension did not reveal a clear pattern consistent with any of the psychological dimensions. An examination of the emotions separated on the ends of the dimensions provides insight on the emotions that were highly differentiated by the properties of each dimension. Here we see that the emotions of anxious, surprised, and happy were differentiated from lonely and embarrassed on Dimension 1. Still, this is not a clear separation of arousal or valence. The emotions at the high end of Dimension 1 (anxious and surprised) were louder and faster than the other end. These were also characterized by frequent and large pitch changes. Dimension 2 separated angry (low end) from the rest of the emotions, particularly love, happy, and sad (high end) primarily according to dynamic changes in the intensity contour, seemingly capturing the abrupt, staccato-like syllabic pattern of anger from the smooth, legato quality of the syllables in love, happy, and sad. Only the emotions anxious and funny were really well discriminated on Dimension 3. This resembles a separation based on

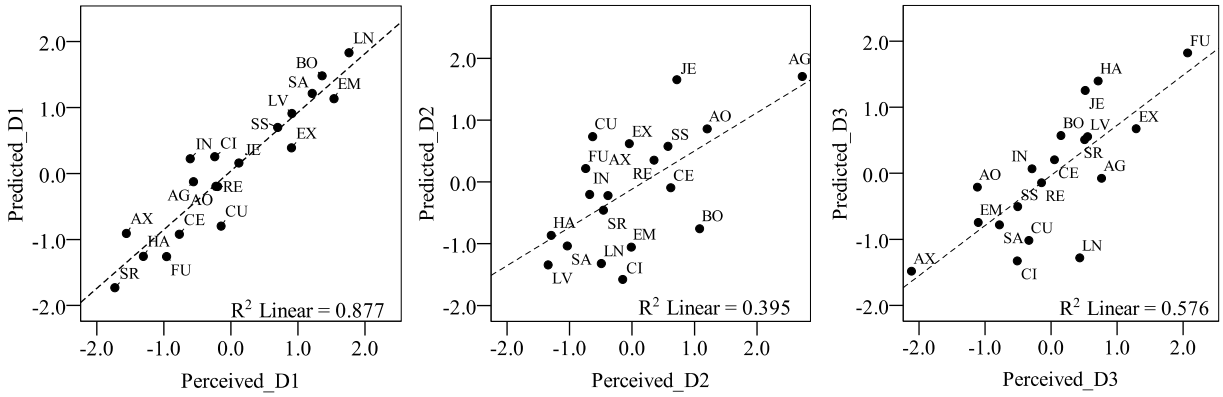

Figure 1: *Relation between the predicted acoustic and the perceived values for all emotions for Dimension 1 (left), Dimension 2 (center), and Dimension 3 (right), where AG = angry, AO = annoyed, AX = anxious, BO = bored, CI = confident, CU = confused, CE = content, EM = embarrassed, EX = exhausted, FU = funny, HA = happy, IN = interested, JE = jealous, LN = lonely, LV = love, RE = respectful, SA = sad, SR = surprised, and SS = suspicious.*

"confidence," in differentiating between positive and negative hesitations. Both durational parameters (VCR, rate, *f*0 peak rise time) and voice quality parameters ($alpha_{ST}$) corresponded to this dimension. Other variables to describe voice quality changes, such as modulations in amplitude, *f*0, and of the syllable rate were not explored, but may be worth investigating here.

The acoustic model predictions showed a close fit for Dimension 1, but a slight compression of the scale was observed for Dimensions 2 and 3 (linear regression slope < 1), with some emotions being better predicted than others, such as happy, angry, sad, anxious, annoyed, and surprised (these fell on the y = x slope). One method of reducing some of the scatter observed for Dimensions 2 and 3 is by examining a smaller set of the highly discriminated emotions. It is likely that not all of these emotions are equally well perceived in speech, such as jealous and lonely [4], and as a result, the inclusion of these emotions may have introduced some error into the model. The model may also be improved by using a two-dimensional model. Since the increase in R^2 from the two- to three-dimensional models was small (R^2 = 0.86 and 0.90, respectively), the addition of a third dimension may not have added a significant amount of information to two-dimensional model and may have resulted in a weak acoustic representation of this dimension (and the low R^2 observed for Dimension 2).

5. Conclusions

In the present study, a dimensional model of the perceived emotions (the perceptual model) using multidimensional scaling was identified and applied in the development of a computational acoustic model of emotions. MDS analysis of perceptual discrimination judgments of emotions showed that three dimensions were sufficient to explain differences in 19 emotions; however these did not exactly map onto the psychological dimensions, demonstrating the need to develop models of emotional prosody based on the perceived dimensions. Then, a dynamic feature set was introduced (i.e., one that does not require normalization to a speaker's "neutral" expression) for developing the acoustic model based on the perceptual dimensions. Stepwise regressions of the acoustic features onto the perceptual dimensions showed that the model predictions for Dimensions 1 and 3 closely matched the perceptual model, especially for some emotions such as happy, angry, sad, anxious, annoyed, and surprised. Further analysis using a two-dimensional model with a reduced set of emotions and a refined feature set is ongoing.

6. References

[1] Tato, R, Santos, R., Kompe, R., Pardo, and J. M., "Emotional space improves emotion recognition", in Proc. of the ICSLP, 2029-2032, 2002.

[2] Laukka, P., Juslin, P. N., and Bresin, R., "A dimensional approach to vocal expression of emotion", Cognition and Emotion, 19:633-653, 2005.

[3] Pereira, C., "Dimensions of emotional meaning in speech", in Proc. of the Speech-Emotion-2000 Tutorial and Research Workshop on Speech and Emotion,, 25-28, 2000.

[4] Patel, S., "An acoustic model of the emotions perceivable from the suprasegmental cues in speech", Ph.D. dissertation, Univ. of Florida, United States, August 2009. [Online]. Available: http://proquest.umi.com/pqdlink?did=1920819581&Fmt=7&clientId=45690&RQT=309&VName=PQD.

[5] Davitz, J. R., "The Communication of Emotional Meaning", 101–112, McGraw-Hill, 1964.

[6] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., and Gielen, S., "Acoustic correlates of emotion dimensions in view of speech synthesis", in Proc. of Eurospeech-2001, Aalborg, 1, 87–90, 2001.

[7] Huttar, G. L., "Relations between prosodic variables and emotions in normal American English utterances", J Speech Lang Hear Res., 11:481–487, 1968.

[8] 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), 189-244, Academic Press, 1985.

[9] Shrivastav, R., Sapienza, C., and Nandur V., "Application of psychometric theory to the measurement of voice quality using rating scales", J Speech Lang Hear Res., 48:323–335, 2005.

[10] Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A., "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts", Cognition and Emotion, 19:1113-1139, 2005.

[11] Juslin, P. N. and Laukka, P., "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion", Emotion, 1:381-412, 2001.

[12] Grandjean, D., Sander, D., Pourtois, G. , Schwartz, S., Seghier, M.L., Scherer, K.R., and Vuilleumier, P., "The voices of wrath: brain responses to angry prosody in meaningless speech", Nat. Neurosci., 8:145-146, 2005.

[13] Macmillan, N. A. and Creelman, C. D., "Detection Theory: A User's Guide (2nd Ed.)", 223, Psychology Press, 2005.

[14] Baird, J. C., and Noma, E., "Fundamentals of Scaling and Psychophysics," 180-190, John Wiley & Sons, Inc., 1978.