

Unveiling the Acoustic Properties that Describe the Valence Dimension

Carlos Busso and Tauhidur Rahman

Multimodal Signal Processing (MSP)
Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
busso@utdallas.edu, txr100020@utdallas.edu

Abstract

One of the main challenges in emotion recognition from speech is to discriminate emotions in the valence domain (positive versus negative). While acoustic features provide good characterization in the activation/arousal dimension (excited versus calm), they usually fail to discriminate between sentences with different valence attributes (e.g., happy versus anger). This paper focuses on this dimension, which is key in many behavioral problems (e.g., depression). First, a regression analysis is conducted to identify the most informative features. Separate *support vector regression* (SVR) models are trained with various feature groups. The results reveal that spectral and F0 features produce the most accurate predictions of valence. Then, sentences with similar activation, but with different valence are carefully studied. The discriminative power in valence domain of individual features is studied with logistic regression analysis. This controlled experiment reveals differences between positive and negative emotions in the F0 distribution (e.g., positive skewness). The study also uncovers characteristic trends in the spectral domain.

Index Terms: valence, emotion recognition, speech analysis, emotion representation

1. Introduction

Emotion recognition is an important problem in the context of *behavioral signal processing* (BSP) and *human machine interfaces* (HMIs). Among different modalities, speech is a valuable source of information to recognize expressive behaviors. In many scenarios and practical problems, it is the only source of information (e.g., call center). Previous studies have reported important progress in affective computing. While acoustic features have been successfully used to discriminate emotions characterized with low or high arousal, previous efforts have failed to robustly discriminate emotions that differ in the valence domain (e.g., happy versus anger) [1, 2]. The lack of discrimination in the valence domain is a major problem in many behavioral problems such as depression and *post-traumatic stress disorder* (PTSD). This paper aims to identify traits in speech that characterize the valence dimension.

Emotional primitives are a popular alternative representation of expressive behaviors. Instead of defining a limited, usually incomplete, set of discrete emotional labels, two or three continuous primitives are defined which cover the entire space. The most common attributes are activation/arousal (excited versus calm), valence (positive versus negative) and dominance (weak versus strong). Several studies have attempted to predict the emotion in terms of these attributes. For example, Grimm et al. proposed several regression models for primitives-based emotion recognition [3, 4]. They were able to achieve small

classification errors using *support vector regression* (SVR) [4]. Wöllmer et al. proposed *recurrent neural network* (RNN) and *conditional random field* (CRF) to track continuous dimensions [5]. In all these studies, the valence dimension was found to be the most challenging attribute. For example, while a linear kernel SVR could predict activation ($\rho_{act.} = 0.80$) and dominance ($\rho_{dom.} = 0.77$) with high accuracy, the correlation for valence was only $\rho_{val.} = 0.37$ [4]. This problem is not only observed with continuous attributes but also with discrete emotional labels [6]. For example, Yildirim et al. reported that angry and happy sentences, and neutral and sad sentences share similar acoustic patterns [2]. Therefore, they are highly confused in the acoustic domain. Notice that these pairs of emotions are similar in the activation domain, but they are different in the valence domain. These studies suggest that finding relevant features to discriminate in the valence domain is one of the main challenges in emotion recognition.

Previous studies have reported some acoustic patterns that are relevant for valence. Spectral tilt, type of phrase accent and boundary tone were found to be useful to discriminate valence [7]. Perez-Espinosa et al. measured the performance of several acoustic feature groups to predict continuous emotion primitives, including valence [8]. Goudbeek et al. reported that positive valence increases the mean value of the second formant [9]. In contrast to previous work, this paper aims to systematically analyze the discriminative power of an exhaustive set of acoustic features in the valence dimension.

This paper systematically studies acoustic properties describing valence. First, acoustic features are clustered into energy, F0, voice quality, spectral, MFCCs and RASTA features. A separate SVR model is trained for each of the feature group. The analysis shows that spectral and F0 features provide the most accurate predictions of valence. Then, a controlled experiment is proposed, in which sentences with similar activation, but different valence are carefully studied. With this novel approach, we analyze individual features by using logistic regression analysis. The study reveals differences between positive and negative emotions in the F0 distribution (e.g., positive skewness). It also uncovers characteristic trends in the spectral domain.

2. Motivation

To motivate the proposed study, we implement a regression analysis to predict the emotional content in terms of activation, valence and dominance. The study relies on the *Vera am Mittag* (VAM) database, which provides realistic audiovisual recordings of emotional behaviors [10]. The corpus was recorded from a German TV show, in which the guests discuss their personal problems. Although the corpus includes video recording,

Table 1: Low level descriptors from speech. The derivatives of these LLDs are estimated and included for analysis (the suffix *de* is included to denote derivatives).

| Group | Low level descriptors | Nomenclature |
|-------------------------|--|-------------------|
| Energy | Sum of RASTA style Auditory Spectrum | SumAudSpecRasta |
| | Sum of Auditory Spectrum | SumAudSpec |
| | RMS Energy | RMSenergy |
| | Zero Crossing Rate | ZCR |
| F0 | Fundamental frequency | F0 |
| | Probability of Voicing | ProbVoicing |
| Voice Quality | Jitter (Local) | JitterL |
| | Jitter (Delta) | JitterD |
| | Shimmer (Local) | ShimmerL |
| Spectral | Spectral Flux | SpectFlux |
| | Spectral Entropy | SpectEnt |
| | Spectral Variance | SpectVar |
| | Spectral Skewness | SpectSkew |
| | Spectral Kurtosis | SpectKurt |
| | Spectral Slope | SpectSlope |
| | Spectral Rolloff 0.25 | SpectROff25 |
| | Spectral Rolloff 0.50 | SpectROff50 |
| | Spectral Rolloff 0.75 | SpectROff75 |
| | Spectral Rolloff 0.90 | SpectROff90 |
| | Spectral Energy 25-650 Hz | Spectfband 25-650 |
| Spectral Energy 1k-4kHz | Spectfband 1k-4kHz | |
| MFCC | Mel-frequency cepstrum coefficients | mfcc |
| RASTA | Rasta-Style Filtered-Auditory Spectral bands[1-26] | Rasta[1-26] |

our study includes only the acoustic modality. The VAM corpus consists of 12 hours of recordings from 47 speakers. The dialogues were segmented into utterances, which were emotionally annotated by 17 raters in terms of the continuous attributes activation, valence and dominance. The perceptual evaluation was implemented using the icon-based, text-free method *self assessment manikins* (SAMs), in which the raters are required to select the pictorial representation that best describes their perceived emotions. For each sentence, the average value across raters is mapped into the range [-1, 1], which is used as ground truth. The study uses 947 utterances from 47 speakers (11 male and 36 female).

We estimate an exhaustive set of sentence level features including prosodic, spectral and voice quality features. The set corresponds to the features provided for the Interspeech 2011 Speaker State Challenge [11]. The feature set includes 59 *low-level descriptors* (LLDs) related to energy, spectral feature, and voiced related features (Table 1). For each of the LLDs, we estimate *high level descriptors* (HLDs) consisting of 33 base functionals and 6 F0 functionals (Table 2). Altogether, the study uses 4,368 sentence level features. The feature set is reduced using *correlation feature selection* (CFS). The features are compared and ranked-ordered according to their correlation. The underlying hypothesis is that a good feature subset should contain features highly correlated with the target class. At the same time, the features should be uncorrelated with each other to avoid collinearity, which is important in regression problems. Notice that the criterion does not use any particular learning algorithm for optimization. The features are sequentially included using forward feature selection.

For regression, we use a linear kernel *support vector regression* (SVR) framework with *sequential minimal optimization* (SMO). SVR is a regression approach based on *support vector machine* (SVM). While SVMs aims to determine the maximum margin separation hyperplane between two classes for classification, SVR aims to find the optimal regression hyperplane in which most of the training samples lie within a margin. The SVR is trained and tested with WEKA data mining toolkit. The corpus is split in four speaker independent partitions (i.e., speech from one speaker is included only in one set). Then, a

Table 2: High level descriptors derived from LLD.

| Functionals | suffix |
|---|--------------------------|
| Quartiles 1-3 | qrtl 1-3 |
| Inter-quartile ranges | iqr 1-2, iqr2-3, iqr1-3 |
| Percentile (1%,99%) | prctl1.0, prctl99.0 |
| Arithmetic Mean, Standard deviation | amean, std |
| Skewness, Kurtosis | skew, kurt |
| Mean of peak distances | meanPeakDist |
| Standard Deviation of peak distances | peakDistStd |
| Mean of peaks | peakMean |
| Arithmetic Mean of mean peaks | peakMMDist |
| Linear Regression Slope and Quadratic error | linregc1, linregerrQ |
| Quadratic Regression coefficients and Quadratic error | qregc1, qregc2, qregerrQ |
| Contour Centroid | centroid |
| Duration when Signal below 25% range | dltime25 |
| Duration when Signal above 90% range | ultime90 |
| Duration when Signal rising/falling | risetime, falltime |
| Gain of linear prediction (LP) | lpgain |
| LP Coefficients | lpc 0-4 |
| Percentage of non-zero frames | nnz |
| mean, max of segment length | meanSegLen, maxSegLen |
| min, std. dev. of segment length | minSegLen, StdsegLen |
| Input duration in seconds | duration |

Table 3: Prediction of valence, activation and dominance using linear kernel SVR.

| Attribute | Without CFS Feature Selection | |
|------------|-------------------------------|---------------------|
| | Correlation | Mean Absolute Error |
| Valence | 0.2161 | 0.3483 |
| Activation | 0.5497 | 0.3866 |
| Dominance | 0.5650 | 0.3756 |
| Attribute | With CFS Feature Selection | |
| | Correlation | Mean Absolute Error |
| Valence | 0.3245 | 0.1452 |
| Activation | 0.8035 | 0.1690 |
| Dominance | 0.7637 | 0.1465 |

fourfold cross validation test is implemented with three sets for training and one set for testing.

Table 3 shows the regression results in terms of correlation and mean absolute error. The results are provided with and without CFS. The table suggest that reducing the feature set with the proposed feature selection scheme improves the regression performance. More important, the results indicate that acoustic features provide discriminative information to predict activation ($\rho_{act.} = 0.80$) and dominance ($\rho_{dom.} = 0.76$). However, we observe significantly lower performance in predicting valence ($\rho_{val.} = 0.32$). These results agree with previous studies that indicate the lack of discrimination of acoustic features in the valence dimension [1, 4, 12]. Given these limitation on the acoustic domain, this paper aims to identify the most relevant acoustic features describing valence. The findings can guide the selection of new acoustic features with better discrimination in this important dimension.

3. Acoustic Features Describing Valence

This section describes the proposed analysis to identify the most relevant acoustic features in the valence dimension. For this purpose, we group low level descriptors into energy, F0 (fundamental frequency), voice quality, spectral, MFCCs and RASTA features (see Table 1). While MFCCs and RASTA features are clearly spectral features, we decide to analyze them in separate groups to study the effect of *discrete cosine transform* (DCT) and temporal filtering after spectral analysis. Since the scope of this paper is valence, we only focus on this dimension.

Table 4: Prediction of valence from different acoustic feature groups using Support Vector Regression

| Group | Correlation | Mean Absolute Error |
|----------------------|-------------|---------------------|
| λ_{Energy} | 0.1555 | 0.1531 |
| λ_{F0} | 0.2749 | 0.1466 |
| λ_{VQ} | 0.0817 | 0.1679 |
| $\lambda_{Spectral}$ | 0.2721 | 0.1453 |
| λ_{MFCC} | 0.2843 | 0.1474 |
| λ_{RASTA} | 0.1606 | 0.1498 |

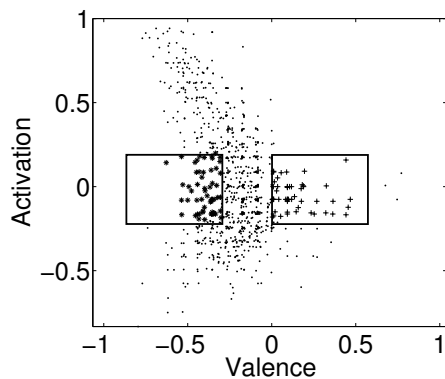


Figure 1: Cluster for binary problem

3.1. Regression per Feature Group

First, we train separate regression models for each feature group. This evaluation aims to identify the feature groups that are more informative in the valence dimension. The correlation coefficient between the predicted valence and the corresponding average value given by human evaluators is used as a metric of the discrimination associated with each feature group.

The analysis follows a similar approach as the one presented in section 2. For each feature group, we train a separate linear kernel SVR with SMO. We reduce the feature set using CFS. We use a fourfold speaker independent cross-validation scheme in which 75% of the data is used for training and 25% for testing. The reported results correspond to the average values across the folds. Although previous studies have argued that voice quality features can provide discriminant information in the valence domain [13], the selected features are not able to produce an accurate prediction ($\rho = 0.08$). Spectral and F0 features are the groups that give better estimates of valence.

3.2. Positive Versus Negative Classification

In the second part of the analysis, we present a novel controlled evaluation to identify specific features that describe valence. The evaluation consists in analyzing speech samples that are perceived with similar activation, but with different valence. Figure 1 plots the perceived activation-valence values of the samples in the VAM corpus. The figure also shows two rectangles that define the samples considered in this section. The two clusters have clearly different valence (negative versus positive). However, their activation values are close to zero. The rectangles includes at least 50 samples per group. Notice that dominance ratings are usually highly correlated with activation ratings. Therefore, it is expected that most of the selected samples will also have similar dominance values. By defining these two groups that are emotionally different only in the valence domain, we expect to directly observe acoustic features that effect the valence dimension.

The proposed analysis to identify the most informative feature for valence is based on logistic regression. Logistic regression is used to model binary or dichotomous variables. In this study, the binary categories correspond to positive or negative valence (Fig. 1). The conditional expectation of the variable given the observations $E(V|f_1, \dots, f_n)$ is given by equation 1. After applying the *logit transformation* (Eq. 2), the regression problem becomes linear in its parameters (β_0, \dots, β_n).

$$E(V|f_1, \dots, f_n) = \pi(\mathbf{f}) = \frac{e^{\beta_0 + \beta_1 f_1 + \dots + \beta_n f_n}}{1 + e^{\beta_0 + \beta_1 f_1 + \dots + \beta_n f_n}} \quad (1)$$

$$g(\mathbf{f}) = \ln \left[\frac{\pi(\mathbf{f})}{1 - \pi(\mathbf{f})} \right] = \beta_0 + \beta_1 f_1 + \dots + \beta_n f_n \quad (2)$$

A property of logistic regression is that the benefits of including new features in the model can be statistically measured by using the log-likelihood ratio test between two nested models (i.e., the variables of one model are included in the variables of the other model). The proposed approach to estimate the discriminative power of each input feature consists in comparing a constant model (Eq. 3) with a model trained with a single feature f_i (Eq. 4). Then, we estimate the statistic χ^2 -2 *log-likelihood ratio* of the models, which is approximately chi-square distributed. This statistic is used for hypothesis testing.

$$H_0 : \beta_0 = 0 \quad g_0(f_i) = \beta_0 \quad (3)$$

$$H_1 : \beta_1 \neq 0 \quad g_1(f_i) = \beta_0 + \beta_1 f_i \quad (4)$$

Among the 4,368 features, only 435 of them are found relevant to discriminate between the two groups (i.e., their regression models with one feature were significantly better, at p -value=0.05, than the constant model). Notice that the most discriminative features ranked with this approach may be correlated. Therefore, the selected set is expected to be different from the features selected with CFS. Figure 2 provides a pie chart with the distribution per group for these features (similar to the metric *share* defined in [14]). The figure shows that spectral features (including RASTA and MFCC) concentrate over 80% of the relevant features. Only 10% of the these features correspond to energy and F0 features.

Figure 3 gives the best 30 features according to their log-likelihood ratio. The features are colored per group. The best feature is *F0dtime25* which corresponds to the duration when F0 is below its 25% range. Further analysis on the data reveals that this feature is higher on sentences with positive valence. For these sentences, we observe that F0 median (*F0qrtl2*) is lower since there are longer duration with small F0 values. These changes in F0 distribution affect its skewness, which increases in positive sentences (*F0skew*).

According to the analysis, the best feature group for discriminating valence corresponds to RASTA style filtered auditory spectrum (14 features in Fig. 3). RASTA features capture the envelope of the true spectrum, reducing the effect produced by noise. From our analysis the ninth, tenth and eleventh coefficients are the best low level descriptors. These coefficients correspond to filter banks with center frequencies at 952Hz, 1111Hz and 1287Hz, respectively. Relevant information is found between 900 and 1300Hz. The functionals *Rfilt[9]prctl1.0*, *Rfilt[10]prctl1.0*, and *Rfilt[11]prctl1.0* are the best three RASTA features. The functional *prctl1.0* corresponds to the 1% percentile of the signal (Table 2). Further analysis on these features reveal higher values for positive sentences. The minimum energy for these coefficients tends to increase.

We observe that the statistics from *SpectROff75* and *SpectROff90* are selected among the best features. The LLD *SpectROff[X]* (*SpectROffX*) is defined as the frequency for

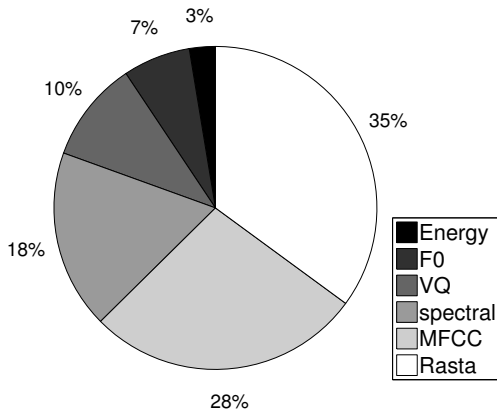


Figure 2: Distribution per group of relevant feature given by logistic regression analysis.

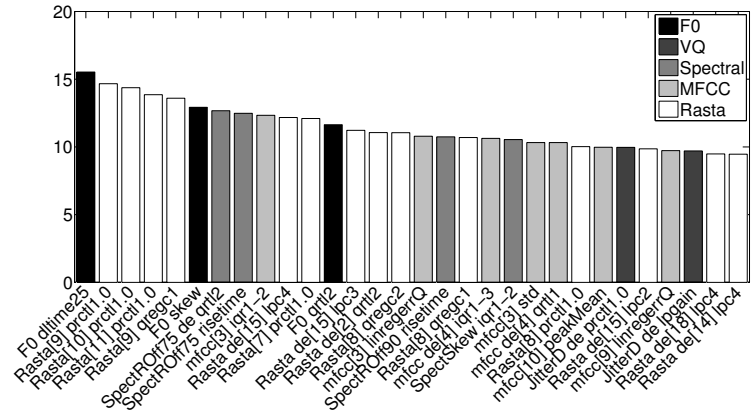


Figure 3: Best 30 features given by the logistic regression analysis. The y-axis is their log-likelihood ratio value.

which $X\%$ of the signal energy fall below that frequency. In particular, the median of the *SpectROff75* derivative (*SpectROff75deqrl2*), and the durations when *SpectROff75* and *SpectROff75* rise (*SpectROff75risetime*, *SpectROff90risetime*) are among the best features. We observe that these values increase for sentences with positive valence. These results suggest that in positive sentences the frequency tends to increase from frame to frame (longer durations with rising spectral roll off frequency).

4. Conclusions

This paper addressed the important problem of identifying relevant features in the valence domain. The analysis with separate regression models trained with various feature groups reveals that spectral and F0 features are the most discriminative acoustic groups. The paper also presented a novel controlled experiment to study valence. Sentences with similar activation, but with different valence were carefully studied. The analysis revealed differences in F0 distribution for positive sentences (e.g., positive skewness). The study also uncovered characteristic trends in the spectral domain.

The proposed analysis can be extended. The study only considered sentences with neutral activation values. We are planning to replicate the analysis with samples with similar high/low activation values, but with different valence values (moving up/down the rectangles in Fig. 1). The challenge in this analysis is that current emotional databases do not span the entire activation-valence space as shown in Figure 1. New emotionally balanced corpora will be needed. Since spectral properties are important to discriminate between positive and negative emotions, we are planning to study articulatory variables. This analysis will give us a better understanding of the key aspects that are used in expressive speech production to modulate changes in the valence domain. These valuable insights will guide us in the design of robust emotion recognition systems.

5. References

[1] C. Busso, M. Bulut, S. Lee, and S. Narayanan, "Fundamental frequency analysis for speech emotion processing," in *The Role of Prosody in Affective Speech*, S. Hancil, Ed. Berlin, Germany: Peter Lang Publishing Group, 2009, pp. 309–337.

[2] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 2193–2196.

[3] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[4] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, vol. 4, Honolulu, HI, USA, April 2007, pp. 1085–1088.

[5] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 597–600.

[6] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Network*, vol. 18, pp. 389–405, May 2005.

[7] J. Liscombe, J. Venditti, and J. Hirschberg, "Classifying subject ratings of emotional speech using acoustic features," in *8th European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, September 2003, pp. 725–728.

[8] H. Pérez-Espinosa, C. Reyes-García, and L. Villasenor-Pine, "Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model," *Biomedical Signal Processing and Control*, vol. 7, no. 1, pp. 79–87, January 2012.

[9] M. Goudbeek, J. Goldman, and K. R. Scherer, "Emotion dimensions and formant position," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1575–1578.

[10] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[11] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011.

[12] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[13] D. Ladd, K. Silverman, F. Tolkmitt, G. Bergmann, and K. Scherer, "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*, vol. 78, no. 2, pp. 435–444, August 1985.

[14] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, January 2011.