# Unveiling the Acoustic Properties that Describe the Valence Dimension

**Carlos Busso and Tauhidur Rahman**

**Multimodal Signal Processing (MSP) Laboratory**
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083, U.S.A.

INTERSPEECH 2012
Spoken Language Processing and Biomedicine

Interspeech 2012: September 9-13, 2012
Portland, Oregon

## Introduction

- Speech is a valuable source to recognize emotional behaviors
- Previous studies have shown that acoustic features
  - can discriminate between emotions with low or high arousal
  - cannot robustly separate emotions differing in valence domain
    - Happiness versus anger [Busso et al., 2009]
- Major limitation in many behavioral areas:
  - Depression
  - Post-traumatic stress disorder (PTSD)

[Cowie et al. 2000]

**Identify speech traits that characterize valence dimension**

## Motivation: Regression Analysis

- Vera am Mittag (VAM) database [Grimm et al., 2009]
  - Realistic audiovisual recordings of emotional behaviors
  - 12-hour recordings from 47 speakers (947 utterance)
  - Activation, valence and dominance (17 annotators)
- Linear kernel SVR with Correlation feature selection (CFS)
  - Exhaustive set of 4,368 sentence level features, IS 2011
  - 4 speaker independent partitions (cross validation)

| Attribute | With CFS [Correlation] | Without CFS [Correlation] |
|---|---|---|
| Valence | 0.2161 | **0.3245** |
| Activation | 0.5497 | 0.8035 |
| Dominance | 0.5650 | 0.7637 |

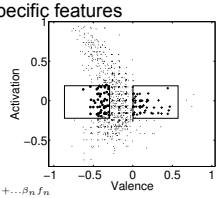
## Regression: Feature Group

- Analysis per acoustic feature group in the valence dimension
  - Energy, F0, voice quality, spectral, MFCCs and RASTA
- We train separate regression models for each feature group
  - Linear kernel SVR with SMO, CFS
  - Fourfold speaker independent cross-validation
- Results:
  - VQ features do not produce accurate prediction ($\rho = 0.08$)
  - MFCC, Spectral and F0 features give better estimates of valence

| Feature Group | [Correlation] |
|---|---|
| $\lambda_{Energy}$ | 0.1555 |
| $\lambda_{F0}$ | 0.2749 |
| $\lambda_{VQ}$ | 0.0817 |
| $\lambda_{Spectral}$ | 0.2721 |
| $\lambda_{MFCC}$ | 0.2843 |
| $\lambda_{RASTA}$ | 0.1606 |

## Positive Versus Negative Classification

- Controlled evaluation to identify specific features
  - 2 groups with similar activation, but with different valence
  - At least 50 samples per group




- Logistic regression framework

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

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

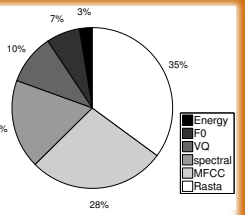
- Benefits of including features can be statistically measured
  - Log-likelihood ratio test between two nested models

$$H_0: \beta_0 = 0 \quad g_0(f_i) = \beta_0$$
$$H_1: \beta_1 \neq 0 \quad g_1(f_i) = \beta_0+\beta_1 f_i$$

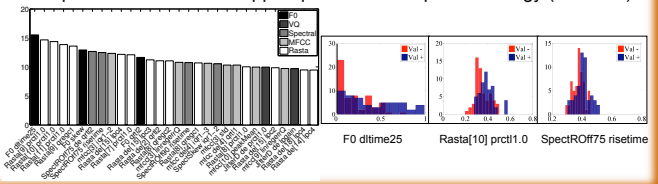
- We compare model with one feature with constant model

### "Share" between feature groups:

- Only 435 features relevant (p-value=0.05)
  - Spectral, RASTA and MFCC (80%)
  - Energy and F0 features (10%)




- F0 dltime25: duration when F0 is below its 25% range
- Rfilt[9,10,11] prctl1.0: 1% percentile of RASTA coefficients [900-1300Hz]
- SpectROff75 risetime: Upper quartile of the spectral energy (rise time)



F0 dltime25 Rasta[10] prctl1.0 SpectROff75 risetime

## Analysis & Conclusions

- F0 and spectral features are the most discriminative groups
- Characteristic trends in F0 distribution for positive sentences
  - There are longer segments with small F0 values
  - Positive skewness
- Characteristic trends in the spectrum for positive sentences
  - Higher 1% percentile of RASTA coefficient [900-1300Hz]
  - Increase in rise time duration for spectral roll-off [75%]

### Future Directions:

- Consider sentences with high/low activation values
  - Moving up/down the rectangles
- Study articulatory feature (USC-EMA)

### 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 R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. September 2000, pp. 19–24. Newcastle, Northern Ireland, UK: ISCA.
③ 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), Honolulu, HI, USA, April 2007, pp. 1085–1088.