Energy and F0 contour modeling with Functional Data Analysis for Emotional Speech Detection

Juan Pablo Arias and Nestor Becerra Yoma

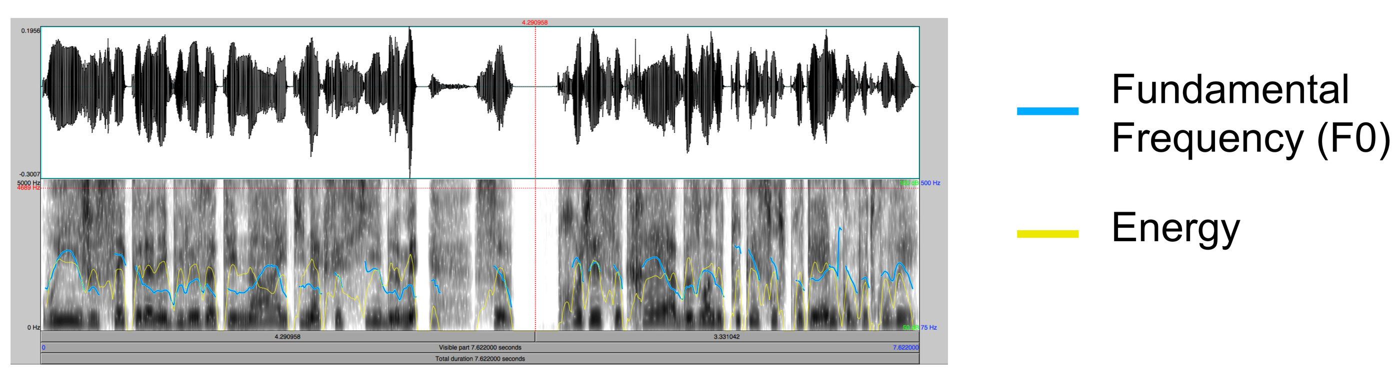
Speech Processing and Transmission Laboratory
Department of Electrical Enginering
University of Chile
Av Tupper 2007, Santiago, Chile

Carlos Busso

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

Introduction

- State-of-the-art: extract global statistics from acoustic features
- Do we capture all the emotional cues with global statistics?
  - Rising and falling F0 movements within accents [Paeschke & Sendlmeier, 2000]
  - Concavity and convexity of the F0 contour [Yang & Campbell, 2001]
  - Pitch accents and boundary tones [Liscombe et al., 2003]



- Goal: modeling the **shape** of the energy and F0 contours
  - Emotional prominence using shape-based neutral models

Modeling Approach with Functional Data Analysis (FDA)

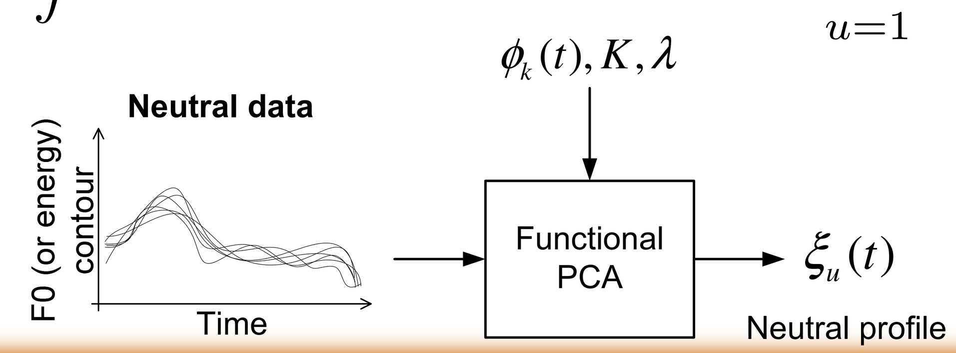
- FDA represents the structure of signals as functions
  - *x(t)* : signal *y(t)*: sampled value $\phi_k(t)$: basis functions

$$y_j = x(t_j) + \epsilon_j \qquad x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

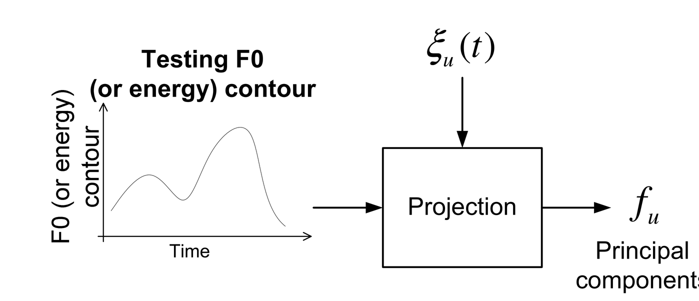
$$\hat{c_k} = argmin_{c_k} \sum_{j=1}^{n} [y_j - x(t_j)]^2 + \lambda \int [D^m x(s)]^2 ds$$

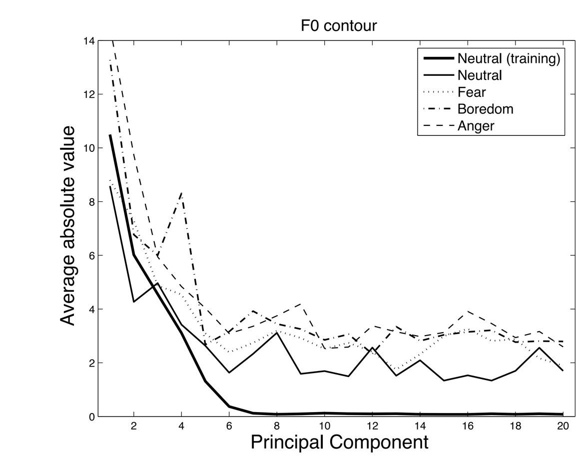
- Functional Principal Component Analysis (fPCA)
  - $x_v(t)$: set of functions $\xi_u(t)$: orthonormal basis

$$f_{u,v} = \int \xi_u(t) x_v(t) dt \qquad \hat{x}_v(t) = \sum_{u=1}^{U} f_{u,v} \xi_u(t)$$

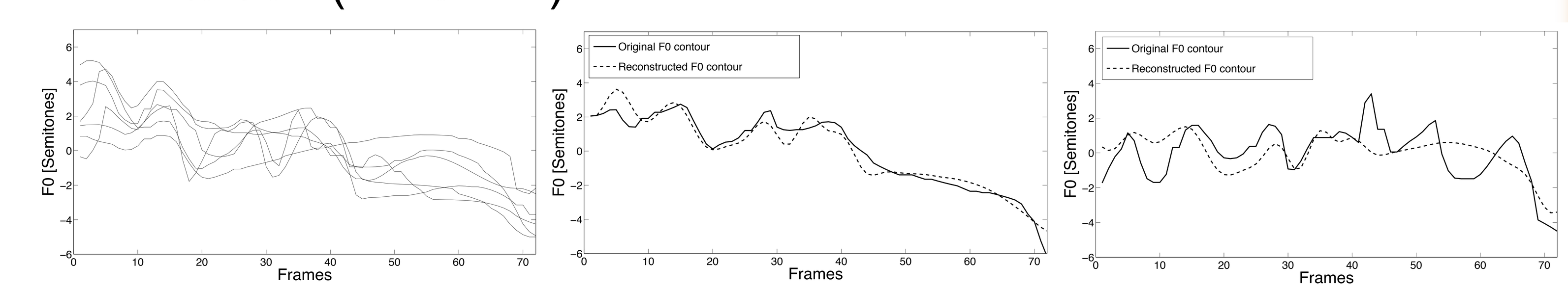


- We use fPCA to train neutral reference models
  - Projections $\{f_1, \ldots, f_U\}$ are used as features





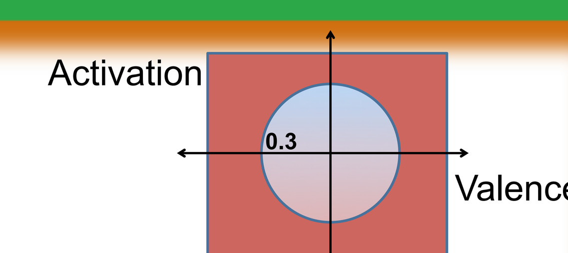
- Intuition (EMO-DB):



F0 contours of six neutral sentences | Reconstruction of F0 contour for **neutral** sentence (*U*=5) | Reconstruction of F0 contour for **happy** sentence (*U*=5)

- Implementation: ϕ_k ➔6th order B-spline with *K*=40 and *U*=20

Discriminant Analysis



- Emotional detection (neutral vs. emotional)
- Quadratic discriminant classifier (QDC)
  - SVM achieves similar performance
- We evaluate lexicon-independent models
  - Neutral speech with different lexical content
- Benchmark classifiers (QDC)
  - Trained with statistics from F0 and energy
  - Subset from IS challenge 2010 [Schuller et. Al, 2010]
  - Forward feature selection
    - 20 for (F0) or (E)
    - 40 for (F0+E)

EMO-DB corpus [Burkhardt et al. 2005]:

- Sentences' durations are linearly warped
- Speaker-independent cross-validation
  - Development, training, testing sets
- Emotional classes grouped into 1 class
  - Trained with under sampling (100 times)

EMO-DB	Accuracy	Average Precision	Average Recall	F-score
FDA (F0)	71.3 (3.6)	75.6	64.1	0.691
FDA (E)	75.9 (1.6)	80.0	69.2	0.742
FDA (E+F0)	80.4 (1.8)	88.3	70.3	0.782
Ben. (F0)	69.0 (9.7)	88.9	45.8	0.555
Ben. (E)	65.9 (7.3)	67.3	67.5	0.666
Ben. (E+F0)	62.8 (9.1)	95.9	27.2	0.390

SEMAINE corpus [McKeown et al., 2010]

- FDA neutral models trained with WSJ1
- Time based segmentation (1 sec)
- Neutral and emotional classes based on averaged activation-valence scores
- Two-fold cross-validation (5 train, 5 test)

SEMAINE	Accuracy	Average Precision	Average Recall	F-score
FDA (F0)	63.6	63.6	63.6	0.636
FDA (E)	57.6	57.1	59.0	0.570
FDA (E+F0)	64.2	64.3	64.2	0.642
Ben. (F0)	58.4	57.8	57.7	0.577
Ben. (E)	56.3	54.9	54.8	0.548
Ben. (E+F0)	57.4	56.5	56.3	0.563

Results & Conclusions

- EMO-DB:
  - fPCA projections increase performance up to 17.6%
  - The fPCA classifiers are more consistent (lower std)
- SEMAINE
  - Classifiers with fPCA projections are 6.9% better than benchmark
  - Performance is not affected by shorter segments (results on paper)
- Global statistics do not capture all emotional cues

Future Directions:

- Evaluation of the approach with prosodic & spectral features
- Detect localized emotional information in dialogs

References

① Juan Pablo Arias, Carlos Busso, and Nestor Becerra Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," Computer Speech and Language, vol. In Press, 2013.

Acknowledgements: Work funded by the Government of Chile (Fondecyt 1100195, Mecesup FSM0601), and NSF (IIS-1217104, IIS-1329659).