Cluster Analysis of Differential Spectral Envelopes on Emotional Speech

Giampiero Salvi[1] Fabio Tesser[2] Enrico Zovato[3] Piero Cosi[2]

[1]KTH, School of Computer Science and Communication, Dept. of Speech, Music and Hearing, Stockholm, Sweden
[2]Institute of Cognitive Sciences and Technologies, Italian National Research Council, Padova, Italy
[3]Loquendo S.p.A., Torino, Italy

1) Contribution

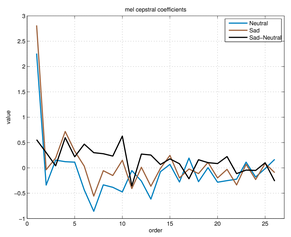
This paper reports on the **analysis** of the **spectral variation** from **neutral** to **emotional speech**.

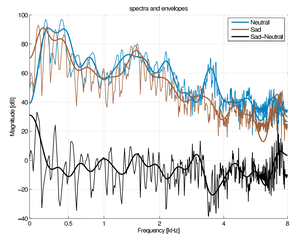
- The analysis is based on **differential spectral envelopes** computed from **mel cepstrum** (no prosody)
- . . . performed by **clustering** the **statistical distributions** of the **differential envelopes**
- Motivation 1: study **speech production**
- Motivation 2: collect useful knowledge for **voice transformation**

2) Data: Parallel corpora

- One **Italian male speaker**
- **Acted speech** (known limitations)
- **Neutral**, **happy** and **sad** emotions
- **200 utterances**/emotion (same content)
- 44.1 kHz sampling rate, down-sampled at **16 kHz**
- **Forced alignment** to detect the phonetic boundaries

3) Method: Differential Mel-Cepstral Analysis





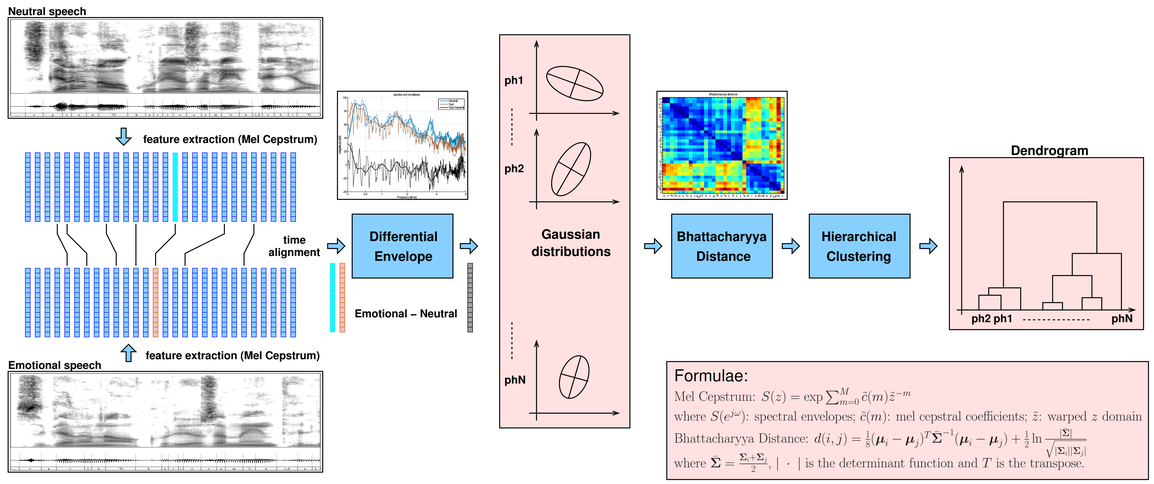
Mel-cepstral analysis:

- Optimal mel-cepstral coefficients estimated from short-time spectrum minimising the spectral envelope representation error directly in the **perceptual** relevant mel-cepstral domain

Differential analysis (DMC):

- Corresponding frames in two different expressive speaking styles are matched by means of **DTW**
- **Feature vectors**: **differences in neutral-emotional pairs** of corresponding mel-cepstral coefficients

5) Method: Schema



Formulae:

Mel Cepstrum: $S(z) = \exp \sum_{m=0}^{M} \tilde{c}(m) \tilde{z}^{-m}$

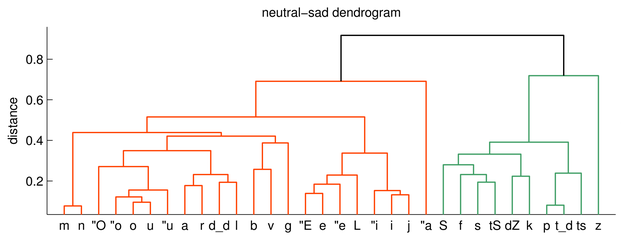
where $S(e^{j\omega})$: spectral envelopes; $\tilde{c}(m)$: mel cepstral coefficients; $\tilde{z}$: warped z domain

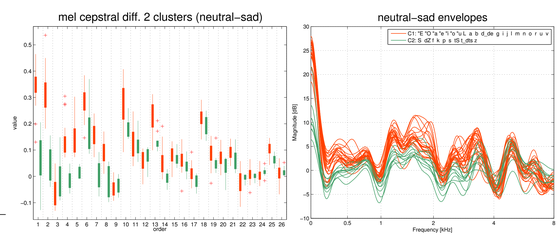
Bhattacharyya Distance: $d(i,j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \bar{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2} \ln \frac{|\bar{\boldsymbol{\Sigma}}|}{\sqrt{|\boldsymbol{\Sigma}_i||\boldsymbol{\Sigma}_j|}}$

where $\bar{\boldsymbol{\Sigma}} = \frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}$, $|\cdot|$ is the determinant function and T is the transpose.

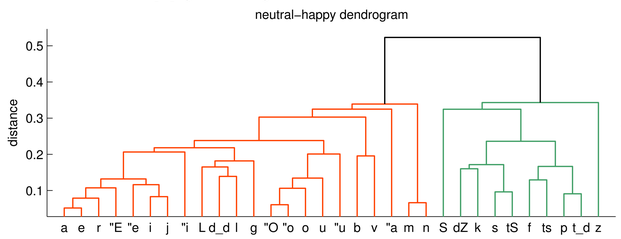
6) Results

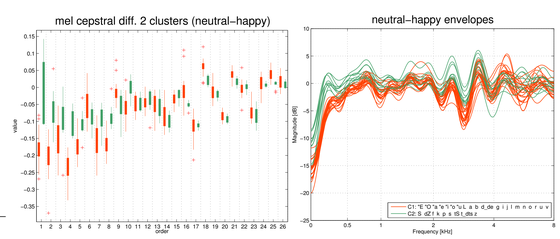
Neutral-Sad analysis





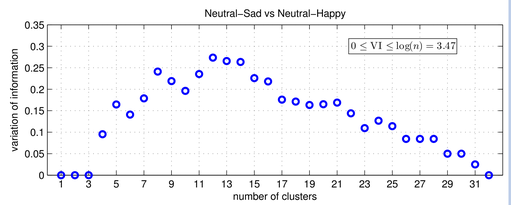
Neutral-Happy analysis





Cluster Validation

- The **Cophenetic correlation coefficient** is **0.78** for the **neutral-sad** and **0.76** for the **neutral-happy** dendrogram (good modelling of the distances)
- The **Variation of Information** (plot to the right) shows that the two dendrograms are similar



4) Method: Clustering and Cl. Validation

Clustering:

- Based on **statistics** of the data for each phoneme (**means** and **covariances**)
- Dissimilarity criterion: **Bhattacharyya distance**
- Method: **Agglomerative hierarchical clustering**
- Linkage: **Average**

Cluster validation:

- **Cophenetic correlation coefficient** (**COPH**): how well a dendrogram models the distance matrix (the closer to **1.0 the better**)
- **Variation of Information** (**VI**): compare different partitions (0.0 if identical partitions, max is $\log(n)$)

7) Discussion

- **COPH** shows that the **dendrograms model the distance matrices well**
- **VI** shows good **degree of similarity** between neutral-sad and neutral-happy **dendrograms**
- The **partition of order 2** separates **voiced** and **unvoiced** both in neutral-sad and neutral-happy
- **Largest timbre deviations** at low-frequencies (< 200 Hz) (influenced by **pitch variation**? $\rightarrow$ need for pitch normalised analysis?)
- **Voiced/unvoiced** separation only below **4 kHz**

8) Conclusions

- **Timbre deviation** from neutral speech is **emotion** and **phoneme dependent**
- Within the **same emotion**, **voicing** plays an important role
- The deviations are **specular** for **neutral-sad** and **neutral-happy** comparisons
- The dendrograms suggest groups of **homogenous transformations** for **voice conversion**
- Emotional expression is **speaker dependent** and should be confirmed on a number of **other subjects**

http://www.speech.kth.se/~giampi Mail: giampi@kth.se, fabio.tesser@gmail.com, enrico.zovato@loquendo.com, piero.cosi@pd.istc.cnr.it