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

giampi@kth.se, fabio.tesser@gmail.com, enrico.zovato@loquendo.com, piero.cosi@pd.istc.cnr.it

Abstract

This paper reports on the analysis of the spectral variation of emotional speech. Spectral envelopes of time aligned speech frames are compared between emotionally neutral and active utterances. Statistics are computed over the resulting differential spectral envelopes for each phoneme. Finally, these statistics are classified using agglomerative hierarchical clustering and a measure of dissimilarity between statistical distributions and the resulting clusters are analysed. The results show that there are systematic changes in spectral envelopes when going from neutral to sad or happy speech, and those changes depend on the valence of the emotional content (negative, positive) as well as on the phonetic properties of the sounds such as voicing and place of articulation.

Index Terms: emotional speech, hierarchical clustering, spectral envelopes

1. Introduction

The communication of different emotions through voice takes place in the speech signal by changes of many acoustic parameters. Prosody correlates such as pitch contours, pausing structure, speech rate and intensity differences are among them [1, 2]. Other significant speech correlates of emotions rely on spectral analysis of speech, such as formants parameters [1], spectral energy distribution [2], and spectral noise [1].

Parameters related to timbre, described e.g. by spectral envelopes, belong to the second class. These parameters are relevant for voice conversion systems [3], and HMM-based speech synthesiser [4], where timbre can be modified and modelled in a statistical way on the basis of speech corpora. However, the models employed in these systems are often complex and hard to interpret.

The purpose of this paper is to analyse the variation of timbre when passing from neutral speech to emotional speech with opposite valence: happy and sad. The aim is not only to describe this variation for human speakers, but also to suggest strategies to improve voice conversion and speech synthesis systems.

We consider a differential approach based on the difference of spectral envelopes between affective and neutral speech segments, computed as mel-cepstral coefficients. This approach was already employed in the task of predicting emotional prosody [5] and has the advantage of eliminating, or reducing, the constant factors related to the speaker and channel characteristics.

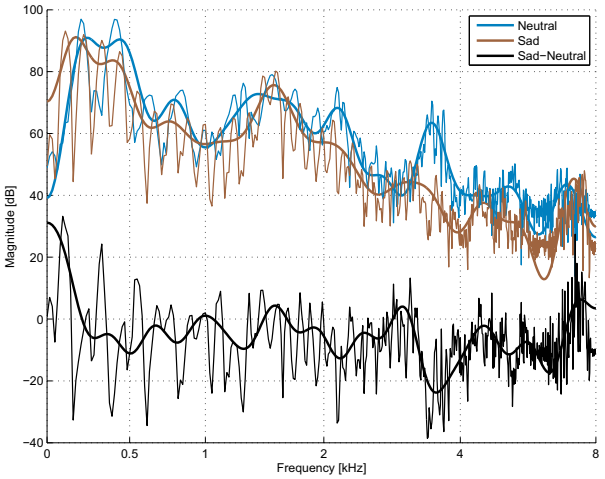

Figure 1: *Spectral envelopes and DFT of a pair of corresponding frames (sad and neutral) and their difference (sad-neutral). x-axis is in warped frequency scale. Thin lines represent the short-term spectrum (DFT), while bold lines represent the mel-cepstral spectral envelopes.*

Another property of differential mel-cepstral analysis comes from homomorphic system theory [6]. The Mel-cepstral transformation, as any homomorphic transformation, converts a convolution into a sum. The spectral envelope computed from differential mel-cepstrum, therefore, represents the frequency response of the filter needed to transform the neutral speech timbre into the emotional one (see, e.g., Figure 1). This is a further link between our analysis and studies on voice conversion, which deal with voice timbre transformation.

In order to perform the analysis on a moderately large data set, we make use of cluster analysis. In [7, 8] statistics were computed for each phoneme in a large corpus with regional accent variation. Then hierarchical clustering was performed on the statistical parameters rather than on the original data points, reducing significantly the computational demands. A similar method is used here based on the statistics of the differential mel-cepstrum for emotional-neutral speech. The advantage of hierarchical clustering, compared to partitional or density-based clustering, is its inherent property of displaying relational features at different levels of details. This fits the analysis scope of this paper where we want to explore the data.

The paper is organised as follows: Section 2 describes the feature extraction procedure and the clustering methods. The experimental settings and the results are described in Section 3. Finally, Section 4 concludes the paper.

2. Method

2.1. Differential Mel-Cepstral Analysis

Spectral envelopes are estimated using mel-cepstral analysis [9, 10]; in this technique optimal mel-cepstral coefficients $\tilde{c}(m)$ are estimated from the short-time spectrum of speech signals, minimising the spectral envelope representation error directly in the perceptual-relevant mel-cepstral domain.

The spectral envelope $S(e^{j\omega})$ is computed using $M+1$ mel-cepstral coefficients $\tilde{c}(m)$ as:

$$S(z) = \exp \sum_{m=0}^{M} \tilde{c}(m)\tilde{z}^{-m} \quad (1)$$

where $\tilde{z}$ is the warped z domain used to approximate the mel frequency scale [9].

Our method is based on differential analysis of spectral envelopes between corresponding frames in two different expressive speaking styles. This is achieved by aligning the material in our parallel corpus by means of the Dynamic Time Warping (DTW) algorithm [11]. In our implementation, both spectral similarities (based on the mel-cepstral coefficients) and phonetic boundaries are considered in the alignment. Differences in neutral-emotional pairs of corresponding mel-cepstral coefficients constitute our feature vectors. An example of this is shown in Figure 1.

2.2. Clustering

Similarly to [8], the cluster analysis is performed in two steps. Firstly, the statistics of the differential mel-cepstral coefficients are collected for each phoneme. These include means μ_i and covariance matrices Σ_i.

In the second step, iterative hierarchical clustering [12] is performed based on the Bhattacharyya distance [13]. The latter is a measure of dissimilarity between distribution i and j, defined as:

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

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

As a measure of how well the resulting dendrogram described the original distance matrix $d(i,j)$ we use the Cophenetic correlation coefficient [14] defined as:

$$\text{coph} = \frac{\sum_{i<j}(d(i,j)-d)(t(i,j)-t)}{\sqrt{\left[\sum_{i<j}(d(i,j)-d)^2\right]\left[\sum_{i<j}(t(i,j)-t)^2\right]}} \quad (3)$$

where d is the average of the distances $d(i,j)$, $t(i,j)$ is the distance in the dendrogram at which object i and j first meet, and t is the average of the $t(i,j)$s. The closer coph is to 1, the better the dendrogram represents the original distances $d(i,j)$.

We also use a measure of dissimilarity between alternative partitions of the data based on the variation of information [15] defined as:

$$VI(X;Y) = H(X) + H(Y) - 2I(X,Y) \quad (4)$$

where X and Y are two partitions of the data, $H(\cdot)$ is the entropy and $I(\cdot,\cdot)$ the mutual information. The probabilities used to compute entropy and mutual information are estimated from the frequencies of each symbol in each of the partitions X and Y. The measure is, therefore, independent on the arbitrary assignments of cluster labels, and can be used even if X and Y contain different numbers of clusters. VI has the properties of a metric, and is, therefore, 0 if and only if the partition X and Y are equivalent.

3. Experiments

3.1. Speech Data

We recorded speech data spoken by an Italian male speaker. In order to collect the parallel corpora, the speaker was instructed to read the same text set with neutral speaking style and two different emotions.

Concerning the neutral style, the speaker was asked to use a standard reading style, without any interpretation, focus or emphasis. In the case of emotional data, he was free to read the same scripts by simulating the two emotions considered in the experiment. The textual corpus is composed of 200 sentences, (generally 10-15 words each), extracted from a large newspaper corpus.

Recording sessions were held in a silent environment, with good digital acquisition equipment, producing linear PCM files at 44.1 kHz sampling rate. The signals were then down sampled at 16 kHz for the analysis.

A rule based automatic grapheme-to-phoneme processor was used in order to obtain the phonetic transcriptions of the scripts. We have then applied forced alignment [16] to detect the phonetic boundaries in the corresponding waveforms.

3.2. Analysis

Mel-cepstral coefficients of order 26 are computed with the SPTK toolkit [17], using 40 ms analysis windows and an hop size of 10 ms. The differential coefficients are calculated from pairs of frames in the neutral-sad and neutral-happy conditions, aligned by our DTW algorithm, and using the forced aligned phonetic boundaries as extra information.

In order to obtain enough samples for each phoneme, geminates are merged with their corresponding consonants. Then statistics are computed as described in Section 2.2 leaving out the first mel-cepstral coefficient $\tilde{c}(0)$. This is done in order to perform an intensity normalisation between sentences and to concentrate the analysis only on the spectral shape of the envelope, without taking into account prosodic intensity variations.

Because the number of samples for some phonemes is scarce, only diagonal covariance matrices are considered, in order to guarantee robust estimation of the parameters. This approximation is justified by the properties of cepstral analysis that tends to minimise correlation. We also performed some tests with full covariance matrices, confirming this observations.

3.3. Results

Figure 2 (left) shows the dendrograms for the neutral-sad and neutral-happy conditions. The Cophenetic correlation coefficient is 0.78 for the neutral-sad dendrogram, and 0.76 neutral-happy dendrogram, indicating that the clustering is a fairly good representation of the pairwise distances. At the bottom of the dendrograms, each phoneme, shown in X-SAMPA labels [18],

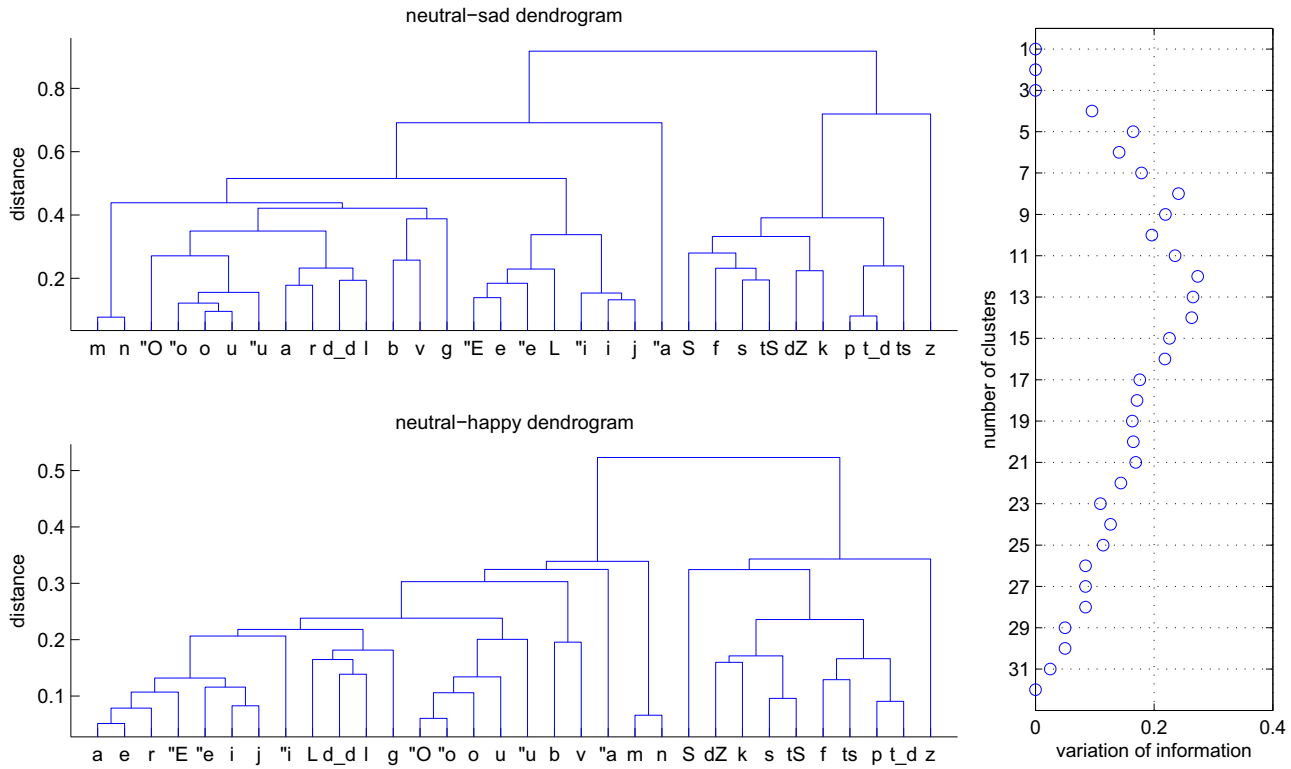

Figure 2: *Left: dendrogram for neutral-sad (top) and neutral-happy (bottom). Right: Variation of information for pairs of partitions of order 1 to 32 extracted from the neutral-sad and neutral-happy dendrograms respectively.*

constitutes a single cluster. Moving upward, clusters are iteratively merged until all phonemes belong to one single group. Considering the trees from the top, in both cases, the partition of order 2 separates mainly voiced and unvoiced phonemes (an informal inspection showed that /dZ/ and /z/ are sometimes unvoiced in the corpus). The next split, in both cases, the phoneme /z/ forms it's own cluster.

Further splits are different in the two trees, although some similarities can be seen. For example, there is a clear cluster including /m/ and /n/ in both trees. Another cluster that appears clearly in both trees includes all the stressed and unstressed back vowels (/"O "o "u o u/).

In order to quantify the degree of similarity between the two dendrograms, Figure 2 (right) shows the variation of information (VI) for each level. As expected, VI is zero for the first, second, third and last level because the partitions of order 1, 2, 3 and 32 are identical. The values of VI in between are smaller than 0.25. To give an idea of the scale, VI is bounded by $\log n$ where n is the number of elements to cluster, and the bound is reached when we compare the partition of order 1 with the one of order n. In our case $n = 32$ and $\log n \approx 1.5$.

In Figure 3, spectral envelopes represent the means of the differential mel-cepstral coefficients for each phoneme, corresponding to the frequency response of the neutral-to-emotional conversion filter.

We can notice from Figure 3 (top), that all the envelopes in the neutral-sad case present a strong amplification in low-frequency band (< 200 Hz), while, on the contrary, an attenuation can be found on the same frequencies range in the neutral-happy case as is shown in Figure 3 (bottom). This phenomenon is in agreement with the results in studies on spectral energy distribution in emotional speech, e.g. [2].

For voiced phonemes this behaviour can be partially explained by the shift in pitch going from neutral to emotional speech. There is usually a decrease in pitch from neutral to sad and a rise from neutral to happy. The corresponding shift in first harmonic affects the estimation of the spectral envelope at lower frequencies (see Figure 1).

The envelopes in Figure 3 are displayed with two different colours indicating the clusters obtained by cutting the corresponding dendrograms at the second order partition. As we have already noticed, this partition separates voiced and unvoiced phonemes in both cases. In the neutral-sad case voiced phonemes have stronger amplification for frequencies less than 4 kHz with respect to unvoiced ones. Similarly in the neutral-happy case a greater attenuation for frequencies less than 4 kHz is found for voiced phonemes with respect to unvoiced ones. In both cases, for frequencies above 4 kHz, there is no longer a clear separation between envelopes belonging to the two clusters. Interestingly, this frequency value corresponds to the value suggested as reference value for the maximum voiced frequency [19].

This observation suggests the hypothesis that, above the maximum voiced frequency, the differential neutral-emotional envelopes of voiced or unvoiced frames have the same behaviour.

4. Conclusions

In this paper we have shown how spectral envelopes, computed using the mel-cepstrum, vary when we consider speech with different emotional content. The differential mel-cepstrum (DMC)

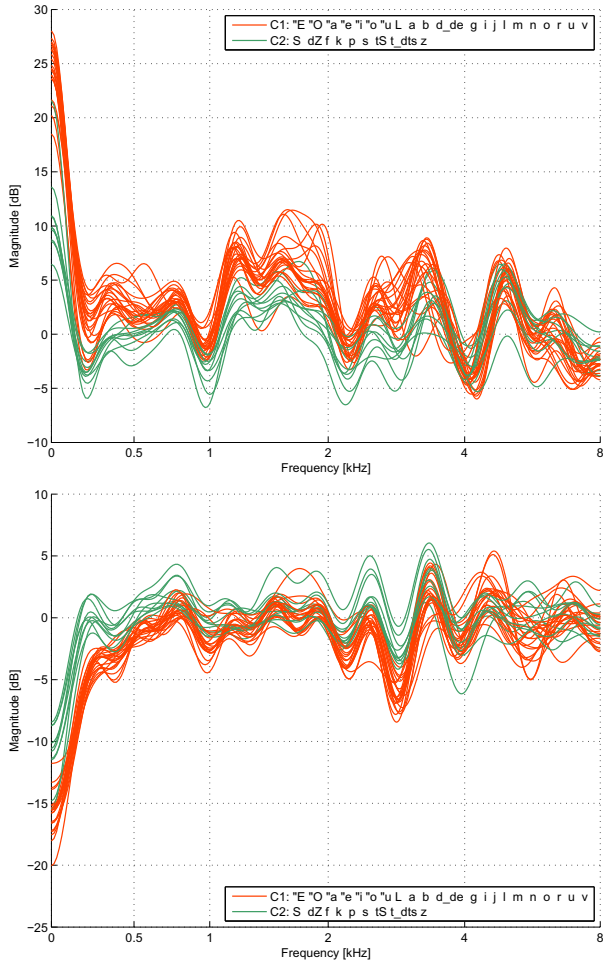

Figure 3: *Spectral envelopes representing the mean of the differential mel-cepstral coefficients for each phoneme. x-axis is in warped frequency scale. Two colours are used to show the clusters obtained from the partition of order 2 in the corresponding dendrogram. Top: neutral-sad case. Bottom: neutral-happy case.*

between neutral and affective speech proved to be a meaningful basis for the analysis. These features also have the advantage of describing a filter that can be applied to the neutral speech to obtain the affective speech, if we are interested in voice modification.

Clustering the distributions of the data, rather than the original data points, proved to be a powerful method to analyse and summarise large amounts of data. The resulting partitions of the DMC statistics for different phonemes can be interpreted in terms of phonetic features. In particular, voicing seems to play a fundamental role in the way spectral envelopes vary across emotional intentions. Back vowels also seem to share similar modifications in the spectral envelopes.

Looking more in details at the average spectral envelopes, we conclude that while up to 4 kHz voiced and unvoiced envelopes have distinct behaviours, above this frequency there is no longer a clear separation between them. In addition, the behaviour of the low frequencies band (< 200 Hz) is the main discriminating factor between neutral-sad and neutral-happy envelope modifications. We suspect this may not be directly associated with vocal timbre but with pitch variations. Although this fact should be further investigated, our results suggest normalising the spectral envelopes with pitch both for analysis and synthesis (voice conversion) tasks.

Finally, it is important to note that emotional expression is speaker dependent and, thus, although the analysis method is applicable regardless the speaker, the generality of the results shown here should be confirmed on a number of other subjects.

5. Acknowledgements

Author G.S. thanks the Swedish Research Council (Vetenskapsrådet grant 2009-4599). Author T.F. would like to thank the Speech, Music and Hearing Department of KTH, Stockholm, Sweden, for the opportunity of visiting their lab. This work was partially supported by the EU's 6th framework project COMPANIONS, "www.companions-project.org", IST 034434 and by the EU FP7 "ALIZ-E" project (grant number 248116).

6. References

[1] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[2] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614–36, March 1996.

[3] A. Kain and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *IEEE ICASSP*, vol. 2, 2001, pp. 813–816.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[5] F. Tesser, P. Cosi, C. Drioli, and G. Tisato, "Emotional FestivalMbrola TTS Synthesis," in *INTERSPEECH*, 2005, pp. 505–508.

[6] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Transactions on Audio and Electroacoustics*, vol. 16, no. 2, p. 221–226, 1968.

[7] G. Salvi, "Accent clustering in Swedish using the Bhattacharyya distance," in *15th ICPhS*, August 2003.

[8] ——, "Advances in regional accent clustering in Swedish," in *Eurospeech*, 2005, pp. 2841–2844.

[9] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *IEEE ICASSP*, vol. 8, 1983, pp. 93–96.

[10] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE ICASSP*, vol. 92, no. 1, 1992, pp. 137–140.

[11] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Prentice hall, 1993, ch. 4.

[12] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967.

[13] B. Mak and E. Barnard, "Phone clustering using the bhattacharyya distance," in *ICSLP96*, vol. 4, 1996, pp. 2005–2008.

[14] R. R. Sokal and F. J. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, vol. 11, pp. 33–40, 1962.

[15] M. Meilă, "Comparing clusterings—an information based distance," *J. Multivar. Anal.*, vol. 98, no. 5, pp. 873–895, 2007.

[16] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[17] SPTK Working Group, "Speech Signal Processing Toolkit (SPTK) version 3.3," http://www.sp-tk.sourceforge.net/, December 2009.

[18] J. Wells, "Computer-coding the IPA: a proposed extension of SAMPA," 1995.

[19] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *IEEE ICASSP*, 2008, pp. 4609–4612.