STATISTICAL SPECTRAL ENVELOPE TRANSFORMATION APPLIED TO EMOTIONAL SPEECH

Fabio Tesser,
Institute of Cognitive Sciences and Technologies,
Italian National Research Council
Padova, Italy
fabio.tesser@gmail.com

Enrico Zovato,
Loquendo S.p.A.,
Torino, Italy
enrico.zovato@loquendo.com

Piero Cosi,
Institute of Cognitive Sciences and Technologies,
Italian National Research Council
Padova, Italy
piero.cosi@pd.istc.cnr.it

ABSTRACT

Transformation of sound by statistical techniques is a promising method for a new range of digital audio effects. In this paper a data driven voice transformation algorithm is used to alter the timbre of a neutral (non-emotional) voice in order to reproduce a particular emotional vocal timbre.

Perceptually based Mel-Cepstral analysis and Mel Log Spectral Approximation digital filter are used to represent the speech timbre and to synthesize speech with modified spectral envelope.

The transformation function adopts a GMM (Gaussian Mixture Model) based parametrization in order convert the spectral envelopes. Experiments with the first and second order derivatives of the mel-cepstral coefficients have been undertaken to prove the benefit of including dynamic information in the model.

The proposed algorithm has been evaluated by means of objective measures in the neutral-to-happy and neutral-to-sad tasks.

1. INTRODUCTION

Human voice is capable of producing different tone colors depending on many factors like personal attitudes, context and, not least, the emotional state of the speaker. Regarding emotional speech, many scholars have identified clear correlates between emotional categories and acoustic features such as intonation, loudness, rhythm [1], and voice quality.

In this project we were mainly interested in the latter aspect. We tested some solutions to modify neutral speech data, with the goal of adding expressive characteristics by means of spectral transformations.

Among the applications that could benefit from spectral transformation there are Text-To-Speech systems. For example, concatenative TTS have reached good levels of quality and intelligibility, but have very limited emotional features, if not at all. Experiments with diphones concatenative TTS systems have already been done in this direction [2], and it is likely that effective spectral transformations could improve the overall naturalness and flexibility of Unit Selection TTS systems too.

To this end, probabilistic techniques for voice conversion were adopted [3], and conversion functions were defined by means of spectral analysis and statistical clustering of the acoustic units. Two emotional speaking styles, with opposite valence, were considered: happy and sad. Speech data from a male Italian speaker was recorded accordingly. In fact a corpus of sentences was read by the speaker in a neutral non-emphatic style and afterward the same sentences were read trying to simulate the emotional styles considered.

At this early stage we modeled the spectral envelope of the expressive speech data, i.e. the transformation involves only the magnitude of the transfer function. In order to model the spectral envelope, a mel-cepstral analysis was chosen. Beyond mel-cepstral coefficients, their first and second order derivatives were also included in the model, in order to better account for dynamic variations. In the next section a formal description of the spectral analysis that has been exploited, is reported. Section 3 describes the data resources that have been used for this experiment. The following paragraph is about the training process, while section 5 gives more details about the implementation of the conversion system. Section 6 reports some results, mainly based on objective spectral distance measures.

2. MEL-CEPSTRAL ANALYSIS AND SYNTHESIS

The Mel-Cepstral analysis method [4, 5] is used in order to extract the spectral envelope from speech data.

Mel-Cepstral analysis represents the spectral envelope $H(e^{j\omega})$ using $M+1$ mel-cepstral coefficients $\tilde{c}(m)$ as:

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

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

To compute the mel-cepstral coefficients $\tilde{c}(m)$, the algorithm adopts an optimization method that minimizes the spectral envelope representation error directly in the perceptual-relevant mel-cepstrum domain. An example of the spectral envelope obtained from the Mel-Cepstral analysis of a frame of speech is shown in Figure 1.

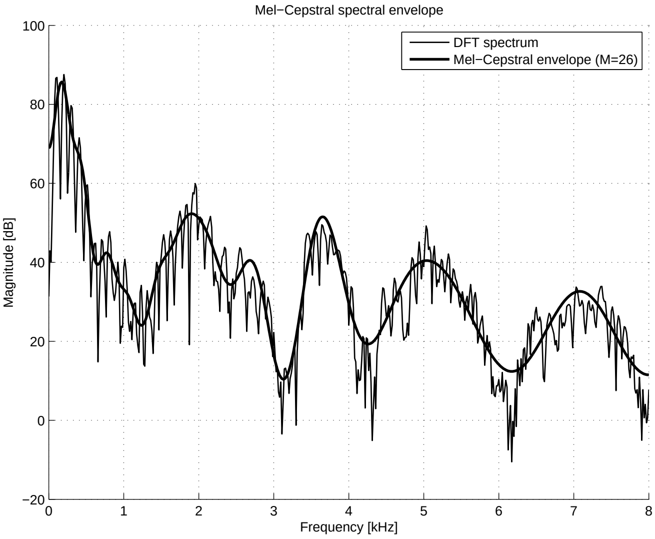

Figure 1: *Signal spectrum (DFT) and Mel-Cepstral spectral envelope ($M = 26$) of a particular frame of speech.*

In the work described here, the notation of spectral envelope vector $\boldsymbol{x}_t$ corresponds to the vector composed by the $M + 1$ mel-cepstral coefficients $\tilde{c}(m)$ computed at the speech frame t.

Mel-Cepstral analysis can be used in the context of the source-filter model of speech production. In this assumption the vocal folds are the source of a spectrally flat sound (the excitation signal), and the vocal tract acts as a filter to spectrally shape the various components of speech.

In this scheme the Mel Log Spectral Approximation digital filter [4, 5] is used in order to synthesize speech with a particular spectral envelope: this technique derives the coefficients of a zero-pole filter directly from the mel-cepstral coefficients $\tilde{c}(m)$.

In the present study, the Mel-Cepstral analysis and synthesis of order $M = 26$ is performed using the SPTK toolkit [6].

3. SPEECH DATA

In this experiment, speech data from one Italian male speaker have been recorded. In order to train the voice transformation model, two sets of data are necessary: the source data and the target one. In the emotional voice transformation case, the source data are extracted from the neutral voice of the speaker while the target data correspond to the emotional voice of the same speaker.

As concerns 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 project. The corpus is composed of 200 sentences, (generally 10-15 words each), extracted from a big newspaper corpora. These sentences provide adequate contextual coverage of the Italian phonetic inventory.

Recording sessions were held in a silent environment, with good digital acquisition equipment. Linear PCM files were produced at 44.1 kHz sampling rate. Post-production included some manual editing to remove voice artefacts, and automatic noise reduction based on spectral subtraction. Signals were then down sampled at 16kHz for analysis and synthesis purposes.

A rule based automatic grapheme-to-phoneme processor was used in order to obtain the phonetic transcriptions of the scripts. Given the phonetic sequences, we have then applied a forced alignment tool [7] to detect their boundaries in the corresponding waveforms.

4. ESTIMATION OF THE TRANSFORMATION FUNCTION

The transformation function $\mathcal{F}(\cdot)$ is a parametrization of the mapping function between coherent pairs of spectral envelope vectors belonging to the neutral and emotional sets respectively.

For the purpose of aligning the corresponding frames between the source and target utterances a Dynamic Time Warping (DTW) algorithm [8] is used. To increase the accuracy of this alignment the DTW algorithm uses the phonetic boundaries information that comes from the force alignment procedure.

The problem of estimating the transformation function can be described as: given the source neutral spectral envelope $\boldsymbol{x}_t$, find the transformation function $\mathcal{F}(\cdot)$ such that the transformed spectral envelope $\boldsymbol{y}'_t = \mathcal{F}(\boldsymbol{x}_t)$ has the best correspondence with the target emotional spectral envelope $\boldsymbol{y}_t$, for all data in the learning set $(t = 1, \ldots, N)$. Following the solution proposed by Stylianou et al.[3], the probability distribution of the neutral acoustical space is modelled with a Gaussian Mixture Model (GMM):

$$p(\boldsymbol{x}_t) = \sum_{i=1}^{Q} \alpha_i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2}$$

and the transformation function has the following parametric form:

$$\boldsymbol{y}'_t = \mathcal{F}(\boldsymbol{x}_t) = \sum_{i=1}^{Q} P(\mathcal{C}_i|\boldsymbol{x}_t) \left[\boldsymbol{\nu}_i + \boldsymbol{\Gamma}_i \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_i)\right] \tag{3}$$

where Q is the total number of GMM components, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance of the mixture component $\mathcal{C}_i$, $P(\mathcal{C}_i|\boldsymbol{x}_t)$ is the conditional probability that $\boldsymbol{x}_t$ belongs to the acoustic class $\mathcal{C}_i$, while $\boldsymbol{\nu}_i$ symbolize the target acoustical space and $\boldsymbol{\Gamma}_i$ stand for the relation between the source and target sets.

The purpose of the training procedure is then to find the transformation function parameters $(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \boldsymbol{\nu}_i, \boldsymbol{\Gamma}_i)$. Figure 2 shows the functional diagram of this operation.

The spectral envelopes are extracted, from both the neutral and emotional speech data, using the Mel-Cepstral analysis, and two sets of paired data are obtained using the DTW algorithm. Then the GMM model parameters representing the neutral acoustical space are estimated using the HTK-based Expectation Maximization algorithm [9], and finally $\boldsymbol{\nu}_i$ and $\boldsymbol{\Gamma}_i$ are computed solving an overdetermined system of linear equations by Least Squares Method on the paired data.

One drawback of this frame by frame transformation is the lack of dynamic coherence. In order to add dynamic information the mel-cepstral coefficient are used with their their first and second order derivatives ($\delta + \delta\delta$) in the training procedure. Results of experiments with and without dynamic features are compared in Section 6.

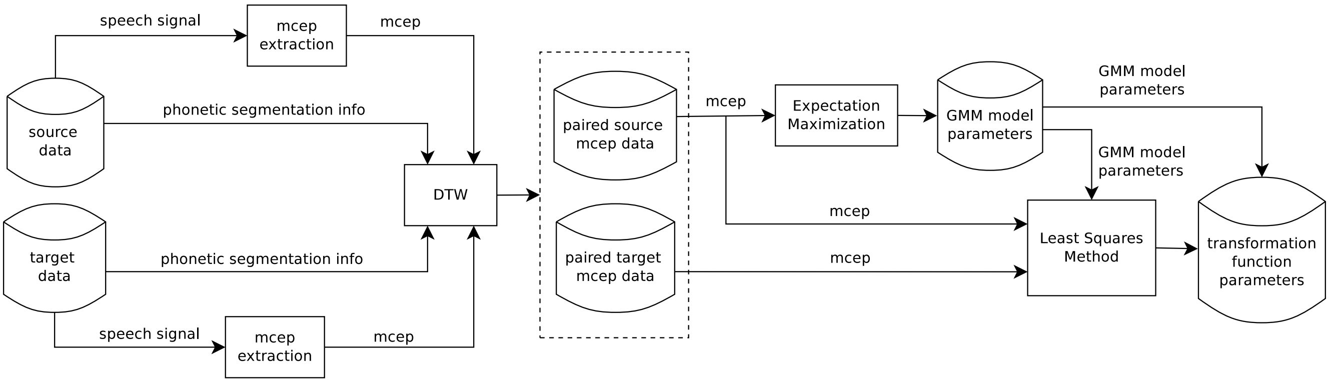

Figure 2: *Functional diagram of the learning procedure.*

5. SPECTRAL ENVELOPE TRANSFORMATION

In order to transform the spectral envelope, two MLSA filters, used in a spectral whitening-reshape scheme, are controlled by mel-cepstral coefficients as is shown in the functional diagram of Figure 3.

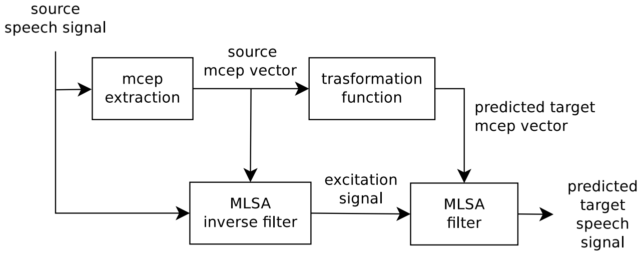

Figure 3: *Functional diagram of spectral envelope transformation system.*

First, the source mel-cepstral coefficients $\boldsymbol{x}_t$ are computed through the use of Mel-Cepstral analysis of neutral speech. This spectral envelope vector is used to control the inverse MLSA filter in order to whitening the spectrum of the source speech signal.

The predicted spectral envelope $\boldsymbol{y}'_t$ is then computed using the transformation function $\mathcal{F}(\boldsymbol{x}_t)$ obtained in the training phase.

In this re-synthesis scheme, the first cepstrum coefficient $\tilde{c}(0)$, that represents the energy of the speech frame, is not taken into consideration in order to maintain the same intensity of the original source signal.

Finally, the whitening signal is used as excitation signal for the MLSA filter controlled by the predicted mel-cepstral vector $\boldsymbol{y}'_t$ in order to synthesize the speech signal with the predicted emotional spectral envelope.

6. OBJECTIVE RESULTS

A good perceptual measure of the distance between two spectral envelopes $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ is the Mel-Cepstral Distance (MCD):

$$mcd_{[dB]}(\boldsymbol{x}_t, \boldsymbol{y}_t) = \sqrt{\int_{-\pi}^{\pi} \left(20 \log_{10} \left| \frac{H_{\boldsymbol{x}_t}(e^{j\tilde{\omega}})}{H_{\boldsymbol{y}_t}(e^{j\tilde{\omega}})} \right| \right)^2 \frac{d\tilde{\omega}}{2\pi}} \quad (4)$$

where $\tilde{\omega}$ represents the mel warped angular frequency.

This measure can be computed using the correspondents mel-cepstral coefficients as:

$$mcd_{[dB]}(\boldsymbol{x}_t, \boldsymbol{y}_t) = \frac{20}{\ln(10)} \sqrt{\sum_{m=1}^{M} [\tilde{c}_{\boldsymbol{x}_t}(m) - \tilde{c}_{\boldsymbol{y}_t}(m)]^2} \quad (5)$$

where the first mel-cepstral coefficient is omitted because is not taken into consideration in this experiment.

In order to evaluate the spectral transformation two coefficients are used: the prediction error and the normalized prediction error. The prediction error pe is the average Mel-Cepstral Distance between the predicted spectral envelope and the target one:

$$pe = \left\langle mcd_{[dB]}(\boldsymbol{y}_t\prime, \boldsymbol{y}_t) \right\rangle \quad (6)$$

The normalized prediction error pe_N takes into account the original distance between different source-target sets, measuring the improvement compared to the initial situation:

$$pe_N = \left\langle mcd_{[dB]}(\boldsymbol{y}_t\prime, \boldsymbol{y}_t) - mcd_{[dB]}(\boldsymbol{x}_t, \boldsymbol{y}_t) \right\rangle \quad (7)$$

Figures 4 and 5 show the results on both neutral-to-happy and neutral-to-sad voice transformation tasks using these two evaluation factors. Experiments with different numbers of GMM components and with or without dynamic components have been undertaken.

From these plots we observe that, the increase of the number of GMM components, and the inclusion of dynamic features, always reduce the prediction errors (pe and pe_N) and then improve the performance of the voice transformation system.

Another consideration is that neutral-to-sad conversion provides better results than neutral-to-happy. This comes from the different cues of the two emotions. Sadness has a lowest speech rate and more static characteristics in comparison with happiness. Lowest speech rate emotion collects a larger amount of speech data using the same sentences, increasing the reliability of the statistical model. Moreover sadness produces more static spectral envelopes vector distribution with less acoustic variability with respect to happiness, and it is then easily modelled through a transformation that involves mean and variance.

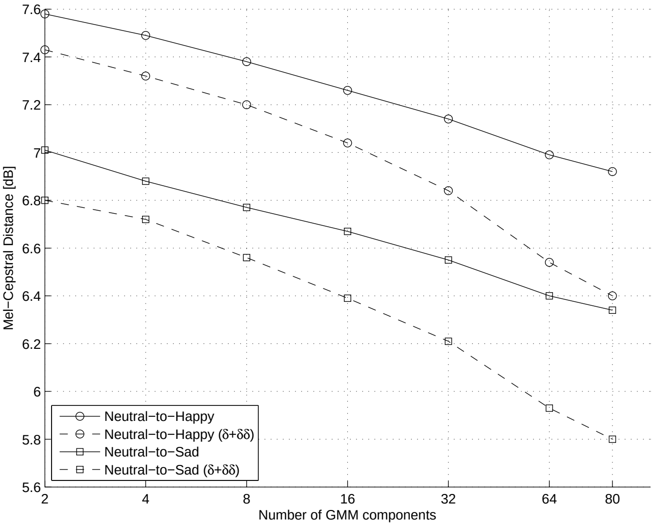

Figure 4: *Prediction error (pe) as a function of the number of GMM components. The original average Mel-Cepstral Distance between neutral and happy is* $9.62\ dB$ *and* $10.92\ dB$ *for the neutral-to-sad case.*

To prove this, it is sufficient to notice that the original average Mel-Cepstral Distance is higher for the neutral-to-sad transformation ($10.92\ dB$) with respect to the neutral-to-happy ($9.62\ dB$), but with only 2 GMM components the result is reversed: Figure 5 shows that in this case the normalized prediction error is reduced by $3.91\ dB$ in the neutral-to-sad case and only by $2.04\ dB$ in the neutral-to-happy case.

7. CONCLUSIONS

In this paper we have described an experiment aimed at transforming the spectral envelop of neutral speech data, according to targets composed of two sets of emotional data: happy and sad. Spectral distance measures have proved the effectiveness of the proposed method, and have demonstrated that the inclusion of dynamic features reduces the spectral distance with respect to target contours.

Informal listening tests have shown that the re-synthesised samples are adequately recognized as happy or sad, depending on the applied transformation. Of course more subjective listening tests will be necessary in order to better evaluate these results. Future plans also include experiments aimed at modelling the source part of the model, beyond its frequency response. Speaker independent spectral envelope transformation will be also attempted.

8. ACKNOWLEDGMENTS

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), http://aliz-e.org/. Author F.T. would like to thank the Speech, Music and Hearing Department of KTH, Stockholm, Sweden, for the opportunity of visiting their lab.

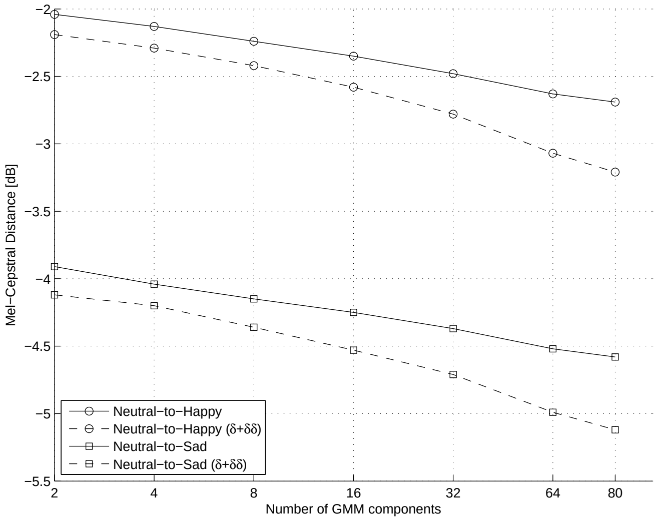

Figure 5: *Normalized prediction error* (pe_N) *as a function of the number of GMM components.*

9. 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] M. Nicolao, C. Drioli, and P. Cosi, "Voice GMM modelling for FESTIVAL/MBROLA emotive TTS synthesis," in *Proceedings of INTERSPEECH*, 2006, pp. 1794–1797.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[4] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1983, vol. 8, pp. 93–96.

[5] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1992, vol. 92, pp. 137–140.

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

[7] 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.

[8] L.R. Rabiner and B.H. Juang, *Fundamentals of speech recognition*, chapter 4, Prentice hall, 1993.

[9] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.-Y. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The Hidden Markov Model Toolkit (HTK) version 3.4," http://htk.eng.cam.ac.uk/, 2006.