Two Vocoder Techniques for Neutral to Emotional Timbre Conversion

Fabio Tesser[1], *Enrico Zovato*[2], *Mauro Nicolao*[1], *Piero Cosi*[1]

[1]Institute of Cognitive Sciences and Technologies, Italian National Research Council,
Padova, Italy
[2]Loquendo S.p.A., Torino, Italy

fabio.tesser@gmail.com, enrico.zovato@loquendo.com,
mauro.nicolao@pd.istc.cnr.it, piero.cosi@pd.istc.cnr.it

Abstract

In this paper, we describe the application of two vocoder techniques for an experiment of spectral envelope transformation. We processed speech data in a neutral standard reading style in order to reproduce the spectral shapes of two emotional speaking styles: *happy* and *sad*. This was achieved by means of conversion functions which operate in the frequency domain and are trained with aligned source-target pairs of spectral features. The first vocoder is based on the source-filter model of speech production and exploits the Mel Log Spectral Approximation filter, while the second is the Phase vocoder. Objective distance measures were calculated in order to evaluate the effectiveness of the conversion framework in predicting the target spectral envelopes. Subjective listening tests also provided interesting elements for the evaluation.

Index Terms: emotional speech, spectral transformation, GMM, mel-cepstral analysis, phase vocoder, MLSA filter

1. Introduction

The study of emotions in human communication has seen a growing interest in the recent years. The achievements within this research field have also been exploited in some applications for improving human-machine interaction. Examples of these applications are Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) synthesis, in which the capability to recognize and generate emotional behaviors can actually improve the naturalness of vocal interfaces in man-machine interaction. Many studies were focused on the characteristics of emotional and expressive speech production in terms of correlates occurring between acoustic patterns and emotion categories [1, 2]. Two main groups of vocal parameters were considered, related to prosody and voice quality respectively. Speech rate, intensity and fundamental frequency (F0, F0 range) are among the most studied prosodic features. Frequency formants, spectral energy distribution and spectral noise are well known features belonging to the second category.

With regards to speech synthesis, the algorithms for processing emotional speech were mainly focused on the control of prosody. In our previous investigations [3], we proposed a framework for emotional speech synthesis that provided a module for prosody modeling along with a rule-based module for voice quality modification. These processing modules were applied to a diphone synthesiser.

Although the rule-based approach provided good results, it was not flexible enough to adapt to the sudden variations occurring in speech. For this reason, we decided to adopt a statistical approach, as it is more suitable for processing both spectral and prosodic features.

A number of signal processing techniques have recently been proposed to solve a somewhat similar task known to the speech community as voice conversion, namely the transformation of a source speaker voice into the voice of a target speaker, while preserving the semantic content of the utterances. It was found that these solutions can also be very effective for emotional speech conversion tasks [4, 5].

Following this trend, our work was aimed at exploiting spectral conversion techniques to improve the emotional perception of processed speech. Of course this technique could be applied to any speech signal. For example, it could be used to convert the speech data of a concatenative TTS system, into something more expressive. This approach could therefore be used to overcome the limited control over expressive features of many state of the art speech synthesizers, which nonetheless provide high quality and intelligibility.

In our experiment we converted the spectral envelopes of neutral speech data into the corresponding envelopes of two emotional targets: happy and sad. Spectral envelopes were calculated exploiting the mel-cepstral analysis [6, 7], because of its capability for extracting spectral features on the basis of perceptual scales.

The prediction of the target emotional spectral envelope, starting from the neutral source, was handled by means of a parametric conversion function which was automatically trained using a data-driven approach. Thus, a parallel corpus of neutral and emotional (happy and sad) speech was recorded by an Italian male speaker. The statistical transformation between the source and target dataset, was executed by training Gaussian Mixture Models (GMM) [8]. In this technique, a conversion function, defined by means of statistical clustering of neutral spectral envelopes, was modeled using the target emotional speech data, with the objective of reducing the mean square prediction error.

The predicted "emotional" spectral envelopes were then used in a vocoder schema in order to effectively modify the spectral timbre of neutral speech utterances. Many digital signal processing techniques could be used for high quality spectral envelope modification, among which we could mention methods such as STRAIGHT [9], Harmonic plus Noise Model [10], Phase vocoder [11] and MLSA filter [6]. Two of these vocoders (Phase Vocoder and MLSA filter) were considered in our experiment.

The paper is organized as follows: the next section describes the spectral analysis that has been exploited. Section 3 consists of the training process of the conversion function while the paragraph after describes the integration of the two vocoder

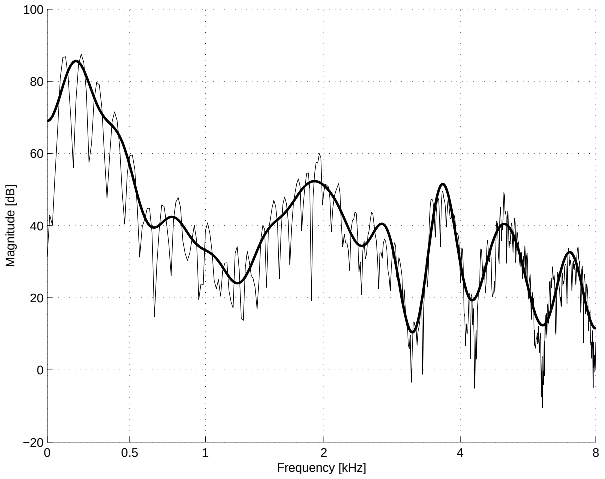

Figure 1: *Signal spectrum (DFT) and Mel-Cepstral spectral envelope ($M = 26$) of a particular frame of speech. x-axis is in warped frequency scale. Thin line represents the short-term spectrum (DFT), while bold line represents the mel-cepstral spectral envelope.*

techniques into our system. Sections 5 and 6 explain the data used in the experiment and how it has been processed. Section 7 reports some results based on objective spectral distance measures and subjective evaluations and finally Section 8 concludes the paper.

2. Spectral envelope extraction

Mel-Cepstral analysis [6, 7] represents the spectral envelope $H(e^{j\omega})$ using $M+1$ mel-cepstral coefficients $\tilde{c}(m)$ as in:

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

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

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

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

3. Mapping function estimation

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

For the purpose of aligning the corresponding frames between the source and target utterances a Dynamic Time Warping (DTW) algorithm [12] has been used. To increase the accuracy of this alignment the DTW algorithm uses the phonetic boundaries information that comes from a forced alignment procedure (see section 5).

The problem of estimating the transformation function can be described as: given the source neutral spectral envelope $\boldsymbol{x}_t$, 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 $\boldsymbol{y}_t$ has to be found for all data in the learning set ($t \in \mathbb{L}$). Following the solution proposed by Stylianou et al. [8], the probability distribution of the neutral acoustical space is modelled with a GMM:

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

and the transformation function has the following parametric form:

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

where Q is the total number of GMM components, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$[1] 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 neutral and emotional speech data, using the Mel-Cepstral analysis described in Section 2, and two sets of paired data are obtained using the DTW algorithm. The GMM model parameters representing the neutral acoustical space are then estimated using the HTK-based Expectation Maximization algorithm [13], and finally $\boldsymbol{\nu}_i$ and $\boldsymbol{\Gamma}_i$ are computed solving an overdetermined system of linear equations by means of the Least Squares Method on the paired data [8].

One drawback of this frame by frame transformation is the lack of dynamic coherence. Beyond the mel-cepstral coefficients, their first and second order derivatives ($\Delta + \Delta^2$) have also been used in the training procedure so as to add dynamic information. Results of experiments with and without dynamic features are compared in Section 7.1.

4. Vocoder techniques

4.1. Phase vocoder based conversion

The Phase vocoder [14], mainly used for operations of pitch shifting and time stretching, can also be used to modify the timbre of an audio signal. Its implementation is based on the spectral representation of time windowed signal intervals and on the overlap-add method. Because of this, it is possible to manipulate some attributes of the original audio signal, both in the frequency and the time domain.

In order to modify the speech timbre, a different shape can be imposed on the moduli of each short-time Fourier windowed frames, according to the desired spectral envelopes.

The framework with the FFT-based phase vocoder [11] that we implemented for this purpose, is shown in Figure 3.

Firstly, the source mel-cepstral vector $\boldsymbol{x}_t$ is computed through the use of mel-cepstral analysis of neutral speech. The predicted target spectral envelope $\boldsymbol{y}'_t$ is then computed using the

[1] Because of the propriety of cepstral analysis that tends to minimise correlation between coefficients, diagonal covariance matrices $\boldsymbol{\Sigma}_i$ are used here as approximation of full covariance matrices.

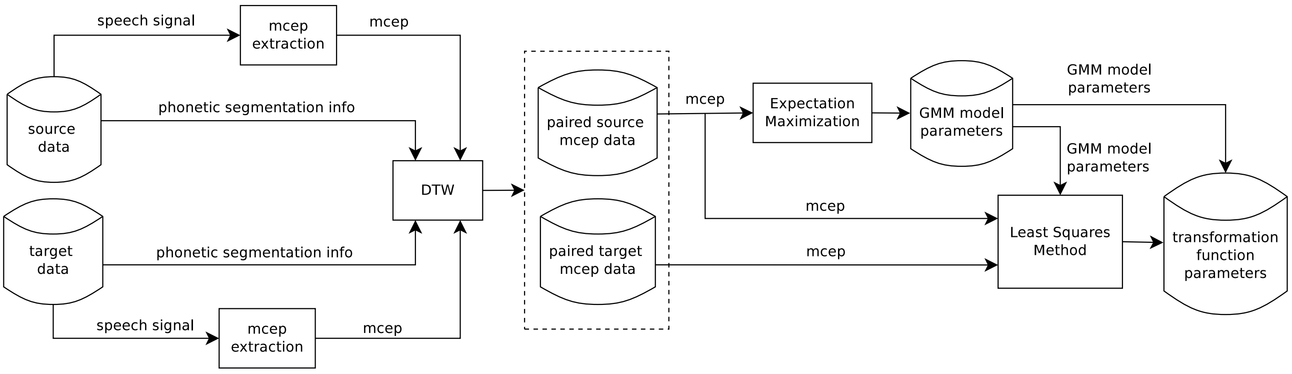

Figure 2: *Functional diagram of the learning procedure.*

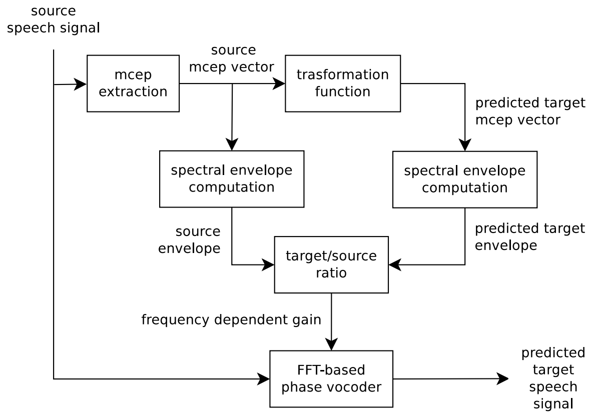

Figure 3: *Functional diagram of spectral envelope transformation system using the phase vocoder technique.*

transformation function $\mathcal{F}(\boldsymbol{x}_t)$ obtained in the training phase. These mel-cepstral vectors are then used to calculate their spectral envelopes in the FFT domain ($Y_t'(f)$ and $X_t(f)$), and their ratio:

$$R(f) = \frac{Y_t'(f)}{X_t(f)} \tag{4}$$

is used as frequency dependent gain[2] applied to the FFT modulus of the source speech frames in the phase vocoder scheme. As a consequence, the shape of the predicted target spectral envelopes are imposed to the FFT frames, which are then joined together by the overlap-add method.

4.2. MLSA filter based conversion

Considering that mel-cepstral vectors represent speech spectral envelopes, they can be used in a source-filter model of speech production. This simplified framework assumes that 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.

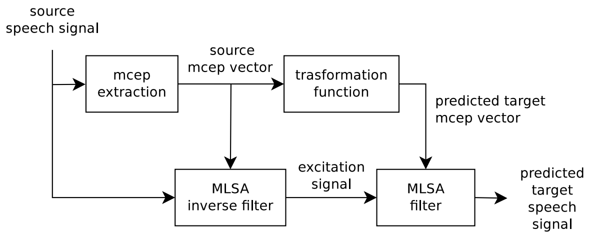

Figure 4: *Functional diagram of spectral envelope transformation system using the MLSA filter technique.*

In this scheme the Mel Log Spectral Approximation digital filter [6, 7] is used to synthesise speech with a particular spectral envelope: this technique derives the coefficients of a zero-pole filter directly from the mel-cepstral coefficients $\tilde{c}(m)$.

To convert the original spectral envelope, two MLSA filters, used in a spectral whitening-reshape scheme, are controlled by mel-cepstral vectors as shown in the functional diagram of Figure 4.

The source mel-cepstral spectral envelope $\boldsymbol{x}_t$ and the predicted target vector $\boldsymbol{y}_t'$ are obtained in the same way as the phase vocoder scheme.

The spectral envelope vector $\boldsymbol{x}_t$ is used to control the inverse MLSA filter in order to whiten the spectrum of the source speech signal, this is employed as excitation signal for another MLSA filter controlled by the predicted mel-cepstral vector $\boldsymbol{y}_t'$, obtaining a speech signal with the predicted spectral envelope.

5. Emotional speech database

In this experiment, speech data were recorded by one Italian male speaker. In order to train the voice transformation model, two sets of data were necessary: the source data and the target one. Source data were extracted from the neutral voice of the speaker while target data corresponded to the emotional voice of the same speaker.

As with what 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: happiness and sadness. The corpus consisted of of 200 sentences, (generally 10-15 words each), extracted from a big newspaper corpora. These sentences provided adequate contextual coverage of the Italian phonetic inventory.

[2] $R(f)$ represents the frequency response of the filter needed to transform the neutral speech timbre into the emotional one. Moreover $R(f)$ can be computed directly from the mel-cepstral coefficients difference $(\boldsymbol{y}_t' - \boldsymbol{x}_t)$ saving some computation load.

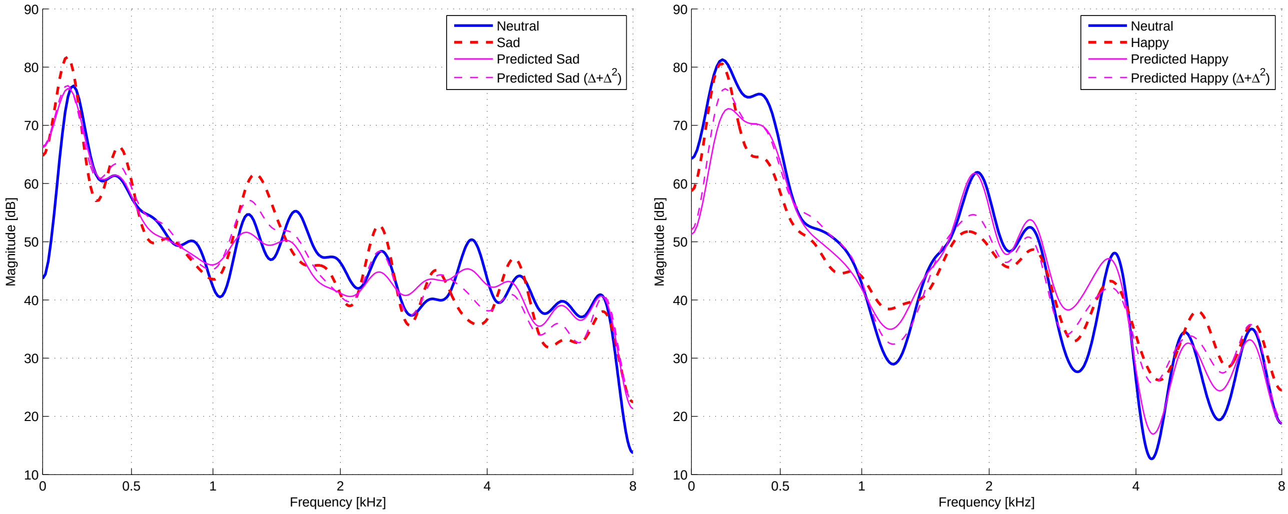

Figure 5: *Two examples, taken from two pairs of frames in the test set, of the mel-cepstral spectral envelopes involved in the transformation. Blue bold lines represent the neutral envelope. Red bold dashed lines represent the real affective envelope. Red thin lines represent the predicted affective envelopes estimated using 16 Gaussian mixture components and respectively the first and second order derivatives coefficients (dashed lines) or not (solid lines). Left: neutral-to-sad case. Right: neutral-to-happy case. x-axis is in warped frequency scale.*

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 down sampling at 16 kHz 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 [15] to detect their boundaries in the corresponding waveforms. This tool is based on Hidden Markov Models (HMMs). For our task 3-5 states (loop-forward) models were used. At the beginning we used a set of speaker independent bootstrap models of the Italian phonemes. These have then been refined through the adaptation with supervised data of our speaker.

The whole corpus was split into a training set, used to estimate the parameters of the voice conversion system, and a test set (unseen data) to measure its performance. We used 180 sentences for training the parameters of the mapping function and the remaining 20 sentences for evaluation.

6. Speech-timbre conversion experiments

In the present study, the mel-cepstral analysis of order 26 was performed using the SPTK toolkit [16]. We used fixed length windows both for mel-cepstral extraction and spectral modification. Frame-analysis size and hop size were set respectively to 40 ms and 10 ms.

The predicted first mel-cepstral coefficient $\tilde{c}(0)$, that represents the energy of the speech frame, was not taken into consideration in both vocoder frameworks. In this way, the same intensity of the original source signal was maintained in the converted signal. In fact, in this work we were only interested in the spectral shape of the target envelope, not in the variation of intensity occurring in the speech datasets.

The transformation function parameters were estimated using the training part of the database, making experiments with different numbers of GMM components and with or without dynamic coefficients ($\Delta + \Delta^2$). The resulting models were then evaluated on the test set (see Section 7.1).

Figure 5 shows an example of the transformation of spectral envelopes involved in the neutral-to-sad (left) and neutral-to-happy (right) conversion. In these figures the initial distance between neutral and emotional envelopes is clearly observable. Moreover, we can note that the predicted envelopes computed with dynamic features, provide better approximations of the target “emotional” envelopes than those computed without dynamic features.

The final part of this experiment consisted of modifying the vocal timbre of some neutral test utterances using the spectral envelopes predicted through the voice conversion procedure (neutral-to-sad and neutral-to-happy) and the two vocoding techniques.

7. Evaluation

7.1. Objective evaluation of the transformation function

A good perceptual measure of the distance between two spectral envelopes $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ is the mel-cepstral distance (mcd):

$$mcd_{[dB]}(\boldsymbol{x}_t,\boldsymbol{y}_t)=\sqrt{\int_{-\pi}^{\pi}\left(20\log_{10}\left|\frac{H_{\boldsymbol{x}_t}(e^{j\tilde{\omega}})}{H_{\boldsymbol{y}_t}(e^{j\tilde{\omega}})}\right|\right)^2\frac{d\tilde{\omega}}{2\pi}} \tag{5}$$

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

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

$$mcd_{[dB]}(\boldsymbol{x}_t,\boldsymbol{y}_t)=\frac{20}{\ln(10)}\sqrt{\sum_{m=1}^{M}\left[\tilde{c}_{\boldsymbol{x}_t}(m)-\tilde{c}_{\boldsymbol{y}_t}(m)\right]^2} \tag{6}$$

where, as above explained, the first mel-cepstral coefficient $(m=0)$ is not taken into consideration in this experiment.

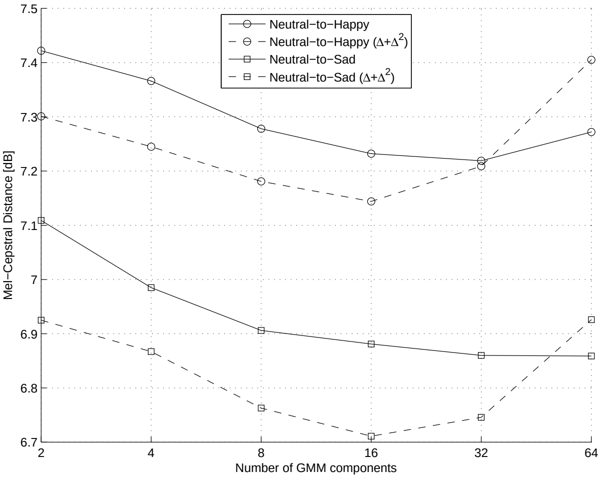

Figure 6: *Prediction error (pe) as a function of the number of GMM components computed in the test set.*

In order to evaluate the spectral envelopes conversion framework, the original spectral distance (osd) between neutral and emotional data, and the prediction error (pe) obtained with the samples included in the test set ($t \in \mathbb{T}$) were used.

The first value is the average mel-cepstral distance between the affective spectral envelopes and the neutral ones:

$$osd = \left\langle mcd_{[dB]}(\boldsymbol{y}_t, \boldsymbol{x}_t) \right\rangle_{t \in \mathbb{T}} \tag{7}$$

whereas the prediction error is the average mel-cepstral distance between the predicted spectral envelopes and the target ones:

$$pe = \left\langle mcd_{[dB]}(\boldsymbol{y_t}\boldsymbol{\prime}, \boldsymbol{y}_t) \right\rangle_{t \in \mathbb{T}} \tag{8}$$

The original spectral distance values are shown in Table 1, while Figure 6 shows the prediction error resulting from models built with different numbers of Gaussian components, with or without dynamic features ($\Delta + \Delta^2$).

Table 1: *Original spectral distance (osd) computed in the test set.*

Original Spectral Distance [dB]	
Neutral-to-Sad	Neutral-to-Happy
10.87	9.34

It is interesting to note that although the original spectral distance (Table 1) is higher for the neutral-to-sad case (10.87 dB) than for the neutral-to-happy one (9.34 dB), the neutral-to-sad conversion provides lower prediction errors than the neutral-to-happy one (see Figure 6).

This probably derives from the different cues of the two emotions. Sadness has a lower speech rate and more static characteristics compared to happiness. Low speech rate also implies a larger amount of speech data when using the same sentences. This could improve the accuracy of the statistical model. Furthermore, sadness probably produces more static spectral envelope vector distributions with less acoustic variability with respect to happiness, and it is then easily modelled through a transformation that involves means and variances.

With reference to Figure 6, we can also notice the overfitting phenomenon: increasing the number of GMM components, a minimum in the prediction error is reached, but after this point performances get worse. This is due to the fact that the models built with a large number of Gaussian components are more complex, and they have too many degrees of freedom (the transformation function parameters), in relation to the amount of data available in the learning set. As a consequence, the model built from the training data provides poor predictive performances when applied to the test set, loosing its generalization characteristics.

Finally, these results show that the inclusion of dynamic features, reduces the prediction errors (for $Q < 64$).

To sum up, the best result in the test set is obtained including the dynamic features and using 16 Gaussian components, and this model was used to produce the sentences for the subjective listening test.

7.2. Subjective evaluation

Beyond objective distance measures, we decided to collect subjective ratings in order to verify whether the perception of the synthesised stimuli was coherent with the intended spectral transformations. We set-up an evaluation schema aimed at assessing the naturalness and the emotional characteristics of the timbre of the synthesised phrases. In particular some subjects had to judge the style of each sample by selecting an option among five labels: sad, slightly sad, neutral, slightly happy and happy. As for naturalness we used a 5 points MOS scale with semantic labels: 5=Excellent, 4=Good, 3=Fair, 2=Poor, 1=Very poor.

The evaluation data set consisted of 8 samples extracted from 5 different groups. The first group included neutral utterances selected from the original recordings of the speaker. The other groups included the phrases in two emotional speaking styles and synthesised with the two vocoders considered: happy MLSA, happy Phase Vocoder, sad MLSA and sad Phase Vocoder respectively.

The test was executed by 30 subjects through an interactive web interface. Some instructions were provided in the home page. In particular, we suggested using the headphones or, alternatively, to do the listening session in a quiet environment. Samples were proposed in random order and subjects could play each stimulus as many times as needed and re-listen to the items previously evaluated.

Results, reported in Figure 7 (top), show that the degree of naturalness for the synthesised phrases is not so distant from the reference cases (neutral), which on average have been judged as “Good”. There are no significant differences between the MLSA filter and the Phase vocoder techniques for the Happy style, while the second one is slightly better for sad transformations. This means that the two vocoder techniques do not introduce perceivable artifacts and are actually suitable for this kind of speech conversion.

Regarding the evaluation of the emotional style, as shown in Figure 7 (bottom), the processing neutral-to-sad seems to be more effective than the neutral-to-happy for which the average values of the ratings are similar. We could hypothesize that the combination of spectral and prosodic features, (which were not considered in this conversion experiment), could lead to better results in terms of emotion perception. The MLSA technique was judged better than the phase vocoder when used to render the sad emotional style. This is probably due to the capability of this filter to model more accurately the spectral energy distribution, and consequently low frequency patterns which are likely to be an important feature in the sad speaking style. In the case of neutral-to-happy conversions the two techniques re-

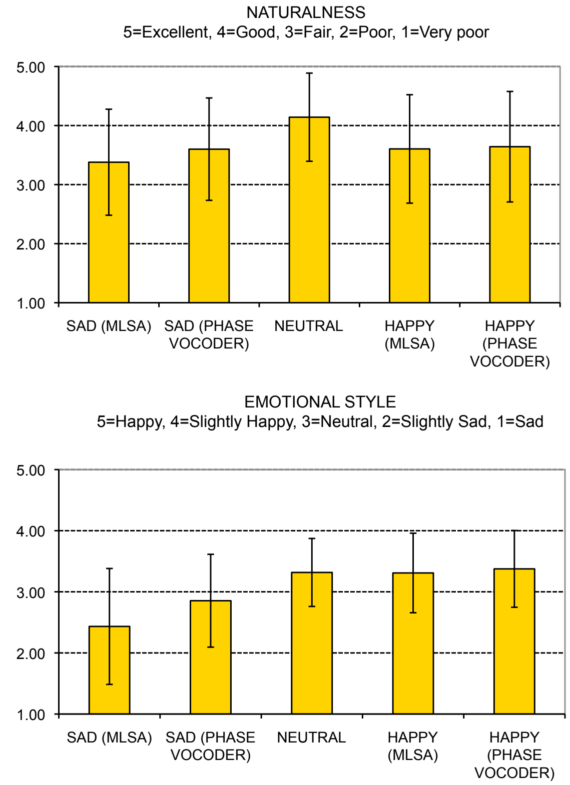

Figure 7: *Subjective evaluation results. Top: Naturalness assessment. Bottom: Emotional perception assessment.*

ceived similar evaluations.

8. Conclusions

In this paper, a framework for timbre conversion from neutral to emotional was presented. This transformation was substantially a signal post processing applied to speech waveforms recorded by a male speaker. In order to describe the modifications between the two spectral acoustic spaces, a statistical conversion function was trained. After obtaining the transformed spectrum, these changes were applied to speech signals with two different vocoding techniques. From both objective measures and subjective assessments we obtained encouraging results.

The objective evaluations showed that the transformation system performed successfully and on average it yielded to a reduction of the spectral distance between transformed neutral speech and target emotional speech. This means that it provided a conversion in the right direction. From these measurements, therefore, we noticed the effectiveness in bridging the gap between neutral and emotional speech, especially when the target emotional style was *sad* rather than *happy*.

The subjective assessment was generally coherent with the objective measures and the neutral-to-sad conversion was often positively recognised. In this particular conversion, MLSA filter was perceived slightly better than the Phase vocoder. In the neutral-to-happy case, the difference between the two vocoders is not so significant to suggest a preference for one of them.

Speech naturalness was tested along with the emotion perception. Most of the listeners agreed that, even though the neutral speech was recognised as being more natural, the converted speech using different vocoders and rendering different emotions, was not perceived so dissimilar. This confirms that the conversion performed well and no unwanted artifacts were added.

9. Acknowledgements

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). Author F.T. would like to thank the Speech, Music and Hearing Department of KTH, Stockholm, Sweden, for the opportunity of visiting their lab. Author M.N. has recently joined to the Speech and Hearing Research Group (SpandH) of the University of Sheffield. Authors also would like to thank the participants of the subjective test.

10. 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] F. Tesser, P. Cosi, C. Drioli, and G. Tisato, "Emotional Festival-Mbrola TTS Synthesis," in *INTERSPEECH*, 2005, pp. 505–508.

[4] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[5] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *EUROSPEECH*, Geneva, Switzerland, 2003, pp. 2401–2404.

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

[7] 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*, vol. 92, no. 1, 1992, pp. 137–140.

[8] 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.

[9] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveign, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187 – 207, 1999.

[10] Y. Stylianou, J. Laroche, and E. Moulines, "High-Quality. Speech Modification based on a Harmonic + Noise Model," in *EUROSPEECH*, no. September, 1995, pp. 451–454.

[11] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Transactions on acoustics, speech and signal processing*, vol. 24, no. 3, p. 243248, 1976.

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

[13] 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.

[14] J. L. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[15] 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.

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