Voice GMM modelling for FESTIVAL/MBROLA emotive TTS synthesis

Mauro Nicolao, Carlo Drioli, Piero Cosi

Istituto di Scienze e Tecnologie della Cognizione - Sede di Padova "Fonetica e Dialettologia"
Consiglio Nazionale delle Ricerche, Via G. Anghinoni, 10 - 35121 Padova, Italy
nicolao@pd.istc.cnr.it, drioli@pd.istc.cnr.it, cosi@pd.istc.cnr.it

10.21437/Interspeech.2006-496

Abstract

Voice quality is recognized to play an important role for the rendering of emotions in verbal communication. In this paper we explore the effectiveness of a processing framework for voice transformations finalized to the analysis and synthesis of emotive speech. We use a GMM-based model to compute the differences between an MBROLA voice and an anger voice, and we address the modification of the MBROLA voice spectra by using a set of spectral conversion functions trained on the data.

We propose to organize the speech data for the training in such way that the target emotive speech data and the diphone database used for the text-to-speech synthesis, both come from the same speaker. A copy-synthesis procedure is used to produce synthesis speech utterances where pitch patterns, phoneme duration, and principal speaker characteristics are the same as in the target emotive utterances. This results in a better isolation of the voice quality differences due to the emotive arousal.

Three different models to represent voice quality differences are applied and compared. The models are all based on a GMM representation of the acoustic space. The performance of these models is discussed and the experimental results and assessment are presented.

Index Terms: Emotive Speech Synthesis, Voice Conversion, GMM , Italian Festival, MBROLA.

1. Introduction

The transmission of emotions in speech communication is a topic that has often received considerable attention. Automatic speech recognition (ASR) and text-to-speech (TTS) synthesis are examples of popular fields in which the processing of emotions can have a substantial impact and can improve the effectiveness and naturalness of the man-machine interaction. Many of the researches in the field have emphasized the importance of prosodic features (e.g., speech rate, intensity contour, F0, F0 range) and the importance of the voice quality in the rendering of different emotions in verbal communication [1].

In TTS technologies, voice processing algorithms for emotional speech synthesis have been mainly focusing on the control of phoneme duration and pitch, which are the principal parameters conveying the prosodic information. On the side of voice quality transformations for speech synthesis, some recent studies have addressed the exploitation of source models within the framework of articulatory synthesis to control the characteristics of voice phonation [1].

A number of signal processing techniques have also been recently proposed to solve a somehow similar task, known to the speech processing community as *voice conversion*, namely the transformation of a source speaker's voice to the voice of a target speaker, while preserving the semantic content of the utterances. Most of these techniques attempt at designing a functional mapping for the conversion of spectral envelopes on a statistical basis [2, 3, 4]. This study addresses the modeling of voice quality in emotive speech by the voice conversion techniques proposed in [3] and [5].

The paper is organized as follows. In Section 2 the voice material is described and the procedure for the construction of the neutral speech by TTS synthesis is illustrated. In Section 3 we define the signal processing framework for the modeling of voice quality characteristics of speech and for the neutral-to-anger conversion. The framework is evaluated on a set of examples from the database and the characteristics and limitations of the method are discussed.

2. Voice material

The voice material used in this work was recorded at the ISTC-CNR Institute in Padova, in an silent room, at a sampling frequency of 44.1 KHz by a male adult speaker. One recording session was specifically designed to collect the voice material needed to build a voice for the diphone MBROLA synthesizer [6]. This voice material was intended to be used for an emotionally "neutral" voice synthesis, and was not given any emotive characterization. The same speaker also performed a second recording session, in which a strong emotive characterization was imposed. The emotion selected was *Anger* (with a strong emphasis), being highly recognizable and being characterized by well perceivable differences in the phonation quality with respect to the neutral phonation. The audio was constituted by an Italian novel. The material was segmented in 47 speech audio files of 5-20 seconds to manipulate it easily. We will refer to this data as to the *target* speech.

Finally, another speech data set was generated through the MBROLA synthesizer, the voice being the one built from our speaker's recordings, by accurately reproducing the pitch countours and phonemes duration from the utterances in the target speech database (copy synthesis process). We will refer to this data as to the *source* speech. In this way, we have two sets of similar utterances, the emotive ones (target) and the neutral ones (source), which are prosodically identical and differ principally for the voice characteristics.

2.1. The copy synthesis process

The copy synthesis process is made of the following steps:

Phone labeling and phone duration through speech recognition performed by an Italian speech recognizer developed at ISTC-CNR, Padua [7]

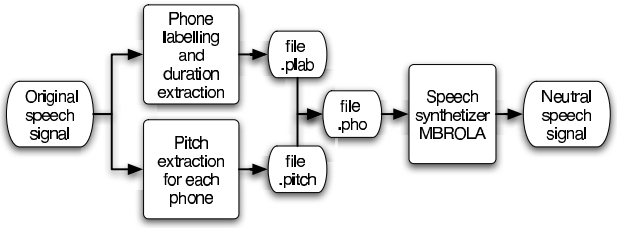

Figure 1: Voice material creation by *copy synthesis*

Pitch extraction by an analysis audio software (PRAAT[1])

Signal synthesis by the speech synthesizer MBROLA ([8])

The obtained audio neutral speech signal has the following peculiar features: time alignment with the target audio files, same duration of the phones, same pitch computed in each analysis frame, same voice.

3. Voice conversion framework

It is evident that the copy-synthesis process alone is not sufficient to reproduce all the subtle characteristics of the target emotional speech. The speech synthesis signal, even if aligned with the target speech as for phoneme timing, pitch contours, energy contours, and speaker identity, is lacking some relevant spectral characteristics both from an objective point of view (spectral distance) and from a subjective point of view (perceptually, the listeners are not satisfied by the rendering of the emotional characterization).

3.1. Speech signal and phonetic classes representation

The conversion process relies on a perceptual representation of the spectral envelopes. For each frame of the speech data, the mel-cepstral coefficients (mfccs) are computed[2], and a smoothed and warped versions of spectral envelope is obtained through an inverse discrete cosine transform. The number M of triangular mel-spaced analysis filters, and the number N_{cep} of coefficients used to represent the envelope, can be used to determine the level of accuracy of the spectral envelope and the resolution with respect to the spectral lobes. The mel-cepstral representation is used here to capture the perceptually meaningful differences between spectra by comparing the smoothed and warped versions of spectral envelopes. The spectral conversion functions will be designed to model these spectral differences.

Finally, instead of addressing the design of an unique conversion function for all phonemes as in, e.g., [3, 5], we decided here to design a set of conversion functions, each of which is aimed at representing the spectral transformations for a specific phonetic class. The phonetic classes were selected as being the ones of the Italian MBROLA voice database, i.e. 34 phonemes including consonants, stressed/non-stressed vowels, open/closed vowels, etc. Since the conversion function models mentioned above are all built upon an acoustic model of the phonetic material in the source speech data, we addressed this task by using the information available in the copy-synthesis phonetic (".pho") input files which drives the MBROLA synthesis in the generation of the source speech data (see Fig. 1). These files contain the label and duration information for each phoneme, thus providing a reliable segmentation of the source speech data to train the acoustic model.

3.2. Design of the conversion model

The synthesis framework, including the spectral transformation system, is shown in Figure 2. The role of the transformation system is to compute the required spectral filtering function to turn the source speech frame from the MBROLA synthesis into a frame with the spectral characteristics of anger.

During the synthesis process, the information on the phonetic class to select the correct conversion function can be extracted from the .pho file driving the MBROLA synthesis, or can in general be extracted by a speech recognition system as the one used in our case to segment the target speech data. Since we adopt a FESTIVAL/MBROLA Text-to-Speech framework, we rely on the .pho file produced by FESTIVAL to obtain the the phonetic class information.

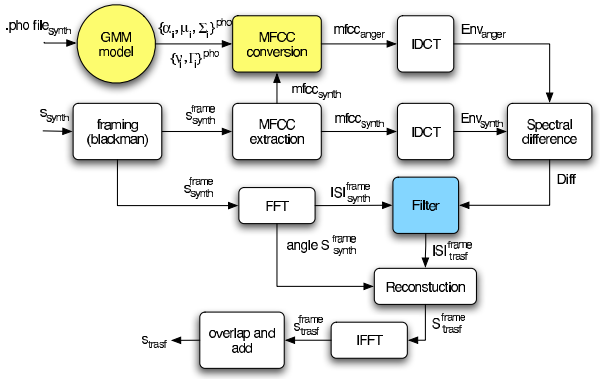

Figure 2: Spectral transformation scheme (Full conversion)

The transformation process in the scheme of Figure 2 is made of the following steps:

1. the neutral speech signal from the MBROLA synthesis, $\mathbf{s}_{\text{synth}}$, is sliced in overlapping frames and multiplied by a blackman window. ($\mathbf{s}^{\text{frame}}_{\text{synth}}$)
2. the MFCCs of the neutral signal are computed ($\mathbf{mfcc}_{\text{synth}}$), and the spectral envelope $\mathbf{Env}_{\text{synth}}$ is built
3. from these MFCCs, the MFCCs for "anger" ($\mathbf{mfcc}_{\text{anger}}$) are computed through the conversion function selected according to the phonetic class information, and the corresponding spectral envelope $\mathbf{Env}_{\text{anger}}$
4. the difference **Diff** between the two envelopes is computed
5. the signal $\mathbf{s}^{\text{frame}}_{\text{synth}}$ is FFT-transformed, and the modulus $|\mathbf{S}^{\text{frame}}_{\text{synth}}|$ and phase $\angle\mathbf{S}^{\text{frame}}_{\text{synth}}$ are computed
6. the spectral difference vector **Diff** is added to the modulus, while the phase is kept unchanged
7. the resulting frame in the time domain is computed through inverse FFT, and the output signal is composed by an overlap-and-add procedure

3.3. Gaussian model

In our experiments, various types of conversion functions were used to evaluate their performance. However, all functions are

[1] By P. Boersma. Available at http://www.fon.hum.uva.nl/praat/

[2] Mel-cepstral analysis was performed with the HTK Toolkit.

built upon a common component, i.e. a gaussian mixture model (GMM) trained on the source speech data. The M-dimensional GMM models the acoustic space of the neutral speech synthesis signal produced with the MBROLA diphone synthesis. The GMM was trained using the HTK toolkit: first, the parameters of the gaussians were initialized through the Viterbi algorithm to provide a first estimate, then a Baum-Welch (Forward-Backward) algorithm was used to refine the parametric identification. Moreover, as seen in Section 3.1, the acoustic space was splitted in phonetic classes in order to associate to each class a different conversion function. A different GMM was thus trained for each one of the 34 phonetic classes derived from the Italian voice database structure.

3.4. Conversion function

We discuss in this section the different types of conversion functions that were considered in this study and used in our synthesis experiments. Let us call $\{\mathbf{x}_n\}$, $n = 1, \ldots, N_T$, the set of MFCC vectors from the neutral signals (source data) that have to be converted, and $\{\mathbf{y}_n\}$ the corresponding target MFCC vectors. Since we introduced in Sec. 3.3 a division of the acoustic space in phonetic classes, each one modeled with a different GMM, similarly the corresponding set of conversion functions will be characterized by a different set of parameters for each phonetic class.

3.4.1. Full conversion

We assume that the full conversion function is ([3]):

$$\mathcal{F}(\mathbf{x}_n) = \sum_{i=1}^{M} P(\mathcal{C}_i \mid \mathbf{x}_n)[\nu_i + \mathbf{\Gamma}_i \mathbf{\Sigma}_i^{-1}(\mathbf{x}_n - \mu_i)] \quad (1)$$

where $P(\mathcal{C}_i \mid \mathbf{x}_n)$ is the probability that the vector $\mathbf{x}_n$ belongs to the i-th mixture, and μ_i and $\mathbf{\Sigma}_i$ are the means and covariances of the GMM model, computed from the source data. The parameters of the conversion function are the vector ν_i and the matrix $\mathbf{\Gamma}_i$, $i = 1, ..., M$, with M being the number of mixture components. The estimate of the parameters was performed with the "Diagonal Conversion" approach illustrated in [3], using as input training data $\mathbf{x}_n$ the first 46 speech files produced by copy synthesis, and using the corresponding 46 files of the target emotive data as training output $\mathbf{y}_n$.

3.4.2. Vector quantization conversion

This method is a simplified version of the full conversion function since it ignores the components accounting for the correlation between the source and target acoustic spaces. It is based on the observation that the matrix $\mathbf{\Gamma}_i \mathbf{\Sigma}_i^{-1}$ is often characterized by low-valued entries, and the contribution of the second part of (1) can be neglected. The resulting conversion function is

$$\mathcal{F}(\mathbf{x}_n) = \sum_{i=1}^{M} P(\mathcal{C}_i \mid \mathbf{x}_n)\, \nu_i \quad (2)$$

The computation of the parameters ν will be simpler in this case (see VQ-Type Convesion in [3]): the k-th component of the conversion vector ν_i is computed as

$$\nu^{(k)} = (\mathbf{P^T P})^{-1} \mathbf{P}^T \mathbf{y}^{(k)} \quad (3)$$

where P is the matrix in which the (n, i) element is $[p(i, k) = P(\mathcal{C}_i|\mathbf{x}_n)]$, $n = 1, \ldots, N$ and $i = 1, \ldots, M$.

3.4.3. Smoothed GMM and MAP adaptation conversion

A variant of the models proposed in [3] has been recently proposed in [5]. The authors highlight that the conversion method introduced by Stylianou et alii often produces an excessive smoothing of the spectral envelopes, and the resulting waveform is consequently affected by undesired distortions and artifacts. To overcome this undesired effect, Chen et alii propose to use the following function instead of (2):

$$\mathcal{F}(\mathbf{x}_n) = \mathbf{x}_n + \sum_{i=1}^{M} P(\mathcal{C}_i \mid \mathbf{x}_n)\, (\nu_i - \mu_i) \quad (4)$$

where μ_i is the usual mean vector of the GMM designed on the MFCCs from the MBROLA synthesis, and $P(\mathcal{C}_i|\mathbf{x}_n)$ is the probability that the training set input vectors belong to the i-th class of the GMM.

The parameter ν_i of the function corresponds to the vector of the means of a GMM designed on the MFCC space of the target data. The computation of this GMM model can be complicated by the lack of sufficient data, or by the difficulty in the alignment of the vector ν_i with the vector μ_i of the neutral GMM. To solve this problem, a MAP adaptation function was proposed, that allows to estimate ν_i from μ_i:

$$\nu_i = \frac{r}{r + \sum_{n=1}^{Q} p_i(\mathbf{x}_n)} \mu_i + \frac{\sum_{n=1}^{Q} p_i(\mathbf{x}_n)\mathbf{y}_n}{r + \sum_{n=1}^{Q} p_i(\mathbf{x}_n)} \quad (5)$$

where r is a tuning constant, $p_i(\mathbf{x}_n) = P(\mathcal{C}_i \mid \mathbf{x}_n)$, and Q is the number of vectors of the target space, aligned with the source. In our case is $Q = N$.

4. Experimental results

In this section we focus on the results obtained using the three conversion methods. In all cases the final choice of the model structural parameters, after testing various settings, was as follows: the MFCCs were computed using 100 triangular filters, equally spaced along the Mel scale, and the number of coefficient that were used to represent the smoothed spectrum envelopes was 26 (including c_0); the training speech samples were sliced in 32 msec frames with a hop size of 4 msec, and a 1024 points FFT was computed on the windowed frames; the number of gaussian functions in each GMM was 64; the training of the parameters of the conversion functions was performed on the set of phonemes from the first 46 speech files, while the remaining speech file was used as a test set. After training, the conversion functions were used to process both speech files from the training set and from the test set and showed appreciable results in both cases (see Fig. 3).

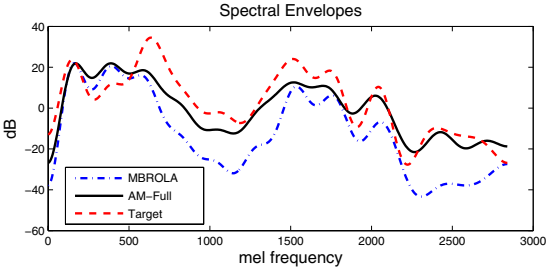

Figure 3: Envelope transformation

4.1. Objective assessment

To objectively assess the effectiveness of the models in the neutral-to-emotive conversion, the Itakura-Saito Distance was selected as the measure of the distance between the target spectral envelopes and the processed source spectral envelopes obtained from the MBROLA synthesis through the various conversion methods used. The different methods adopted are labeled as M (MBROLA), AM-Full (complete conversion), AM-VQ (vector quantization conversion), AM-MAP (MAP adaptation conversion plus Smoothed GMM). In Fig. 4 we show the distances (avaraged over the analysis frames and normalized with respect to the distances of the MBROLA synthesis (M) from the targets) between the target files and the transformed ones, computed for a single phoneme, for a sentence from the training set, and for a sentence from the test set.

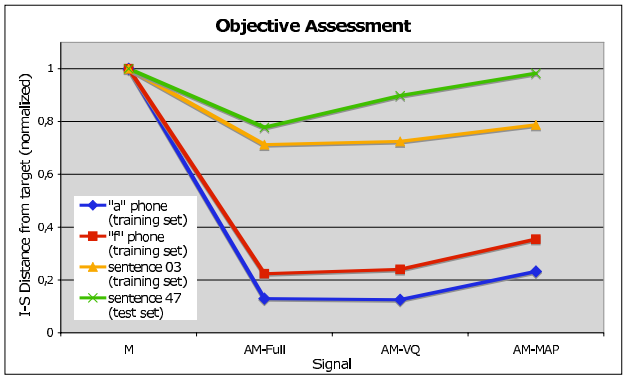

Figure 4: Spectral distance measures between the target signal and the MBROLA synthesis with voice conversion.

4.2. Perceptual assessment

The results of the synthesis were assessed through informal listening tests. The subjects in general agreed in recognizing a drastic improvement in the rendering of the emotive speech when comparing the MBROLA copy synthesis (M) and the MBROLA copy synthesis with the best (according to Itakura-Saito Distance) voice conversion spectral processing (AM), with the standard MBROLA synthesis (SM, i.e. rule-based prosody and neutral voice quality). Both samples from the training set and from the test set were proposed. The subjects were asked to answer the following question: "Is this an angry voice?" and to rate the intensity of anger from 0 to 5. The mean scores are reported in Figure 5.

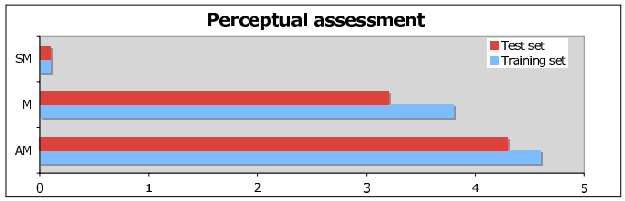

Figure 5: Score of the question of the perceptual assessment.

Finally, it is worth mentioning that the listeners also agreed on judging the synthesis with conversion as slightly degraded if compared to the standard MBROLA or copy-synthesis MBROLA sample, even if they underlined the difficulty to judge the quality of the *transformed* audio files because of the already "synthetic" quality of the *neutral* MBROLA voice.

5. Conclusions

A processing framework for voice transformations finalized to the analysis and synthesis of emotive speech has been proposed. We used a GMM-based approach to model the differences between an MBROLA voice and an anger voice, both from the same speaker, and we studied the problem of modifying the MBROLA voice spectra through a set of spectral conversion functions.

The peculiar organization of the voice training material, in that the target emotive speech data and the diphone database used for the text-to-speech synthesis both comes from the same speaker, permitted to effectively isolate the differences in the voice quality due to the switching from the neutral voice to the anger voice.

Three different conversion models were applied and compared. In all cases the improvement in the rendering of the emotive speech was clear. However, still some improvements are required, mainly to avoid the slight degradation audible in the converted speech synthesis, and is the subject of ongoing research. Moreover, a further model specialization could be obtained by the use of HMMs (Hidden Markov Models) for each phonemes instead of the "simple" GMMs.

6. Acknowledgements

Part of this work has been sponsored by PF-STAR (Preparing Future multiSensorial inTerAction Research, European Project IST-2001-37599, http://pfstar.itc.it).

7. References

[1] C. Gobl and A. N. Chasaide, "The role of the voice quality in communicating emotions, mood and attitude," *Speech Communication*, vol. 40, no. 2-3, pp. 189–212, April 2003.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantizzation," *International Conference on Acoustics, Speech, and Signal Processing*, pp. 655–658, 1988.

[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, March 1998.

[4] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 285–288, 1998.

[5] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, "Voice conversion with smoothed gmm and map adaption," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003, pp. 2413–2416.

[6] P. Cosi, F. Tesser, R. Gretter, and C. Avesani (with Introduction by Mike Macon), "Festival speaks italian!," in *Proceedings of EUROSPEECH 2001*, Aalborg, Denmark, Sept 2001, pp. 509–512.

[7] P. Cosi and J.P. Hosom, "High performance "general purpose" phonetic recognition for italian," in *Proceedings of International Conference on Spoken Language Processing*, Beijing, Cina, October 2000, vol. 2, pp. 527–530.

[8] T. Dutoit and H. Leich, "MBR-PSOLA : Text-To-Speech synthesis based on an MBE re-synthesis of the segments database," *Speech Commun.*, vol. 13, no. 3-4, pp. 167–184, November 1993.