VOICE CONVERSION: ADAPTATION OF RELATIVE LOCAL SPEECH RATE BY MPEG-4 HVXC

Lutz Leutelt, Ulrich Heute
Institute for Circuits and Systems Theory
Christian-Albrechts University
Kaiserstraße 2, D-24143 Kiel, Germany
Phone: ++49 431 880-6132, Fax: ++49 431 880-6128
e-mail:{ll,uh}@tf.uni-kiel.de

ABSTRACT

This paper proposes a modified speech coder, namely the MPEG-4 harmonic vector excitation coder (HVXC), for adaptation of the relative speech rate. Applied to the conversion of voice characteristics it can be used to partially restore the speaking style of a specific speaker which can get lost during the process of a frame-by-frame voice conversion. Furthermore, a method for estimation of the relative speech rate between two different speakers will be given.The results show that the speech rate can be adapted by the proposed system with almost no quality decrease of the synthesized speech.

1 INTRODUCTION

The presented system for adaptation of the speech rate is used in the context of *voice conversion*, the transformation of the voice individuality of one speaker to that of another[2]. By this technique, it is possible to derive various voices (“target” speakers) from one single reference speaker (“source” speaker).

A possible application of such a voice conversion system to a multi-speaker text-to-speech (TTS) system [1] is shown in Figure 1. Instead of using one separate speech database for each speaker as in conventional TTS systems (Figure 1-a), various voices can be derived from one standard speaker by voice conversion. The advantage of voice conversion in this application is that costly studio recording and hand labeling of the speech corpus is only necessary for the source speaker who guarantees the correct operation of the TTS system, whereas the other speaker can be generated by conversion algorithms for which less training data is required. In this way, it is also possible to generate voices that are not available for studio recordings and the memory requirements of the TTS system can be reduced. Using an extended speech decoder for voice conversion as depicted in Figure 1-b allows further reduction of the memory requirements by compression of the source speaker’s database. In this paper, a harmonic speech coder, the Harmonic Vector Excitation Coder (HVXC) of the MPEG-4 standard[3], is used for extraction and synthesis of relevant speaker-characteristic parameters[5].

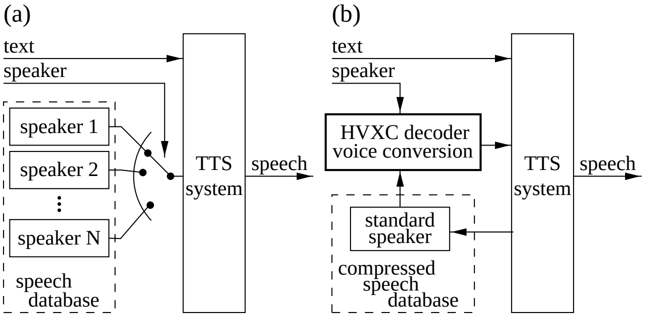

Figure 1: (a) Conventional multi-speaker TTS system and (b) HVXC based system with voice conversion.

The main focus of this work is on an algorithm for restoring the speech rate contour of the target speaker. Restoring can become necessary since voice conversion systems usually generate the target speaker with the speech rate contour of the source speaker when speech parameter sets of the source are extracted on a frame-by-frame basis and mapped directly onto the target-speaker parameter sets. In this way, the (to some extend) speaker-characteristic speech-rate contour, reflecting part of the speaking style of the target speaker, is lost. A method for estimation of the relative speech rate is subject of section 2 and the speech rate adaptation module is presented in section 3.

2 RELATIVE LOCAL SPEECH RATE

The speech rate can be differentiated in global, local and relative speech rate[7]. The *global* speech rate is a measure for the average number of speech segments, e.g. phonemes or syllables, per unit of time for an overall utterance, whereas the *local* speech rate takes into account the variations of speech rate within one utterances due to prosody (stress, emotions, etc.) and speaking style. In contrast to these absolute measures, the *relative rate* (either global or local) is measured with reference to another utterance. In this paper, the variations of the relative speech rate between the target and the source speaker are considered. Figure 2a,b demonstrates the differing global and local speech rate by spectrograms with added phoneme boundaries for two speakers reading the same sentence.

2.1 Estimation

As for the estimation of the relative speech rate, a method is adapted that was proposed for analysis of speech-rate variations of single speakers speaking in different manners [6].

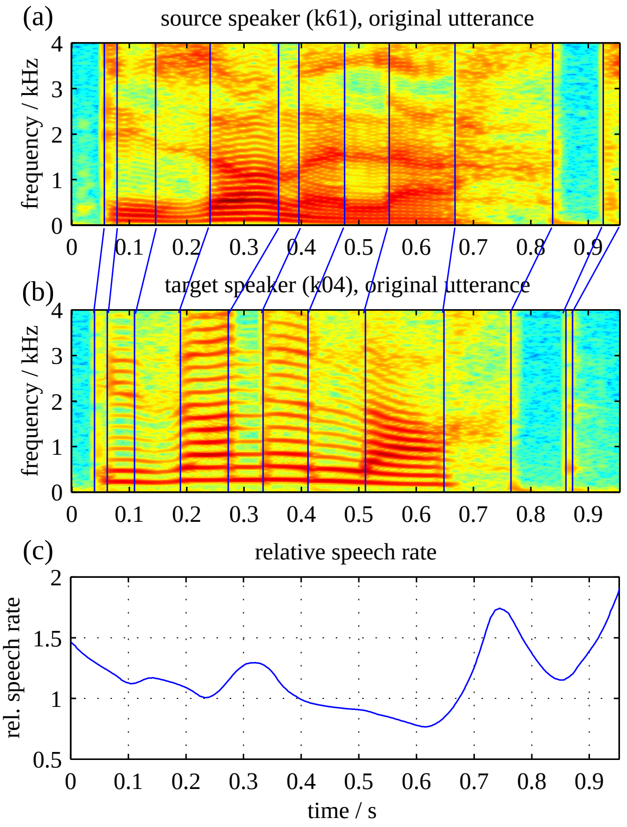

Figure 2: Two corresponding speech spectrograms (sentence BE002: “Die Sonne lacht.”, engl.:“The sun laughs (shines).”) illustrating the different global and local speech rate of source (male, k61) and target speaker (female, k04), and (c) the estimated relative local speech $\tilde{r}(t)$.

The method consists of two steps. At first, reference points $(t_{S,i}; t_{T,i})$ of the time-warping function $w(t)$—that maps the source speaker’s time-axis onto the target speaker’s—are determined by dynamic-time warping (DTW). The employed spectral feature used for alignment of the speaker utterances are Mel-Frequency Cepstral Coefficients (MFCC) and for constraint of the search path a $[1/3, 3]$–scheme is used [6].

The relative local speech rate $r(t)$ can be defined from the warping function $w(t)$ by

$$r(t) = 1 \Big/ \frac{\mathrm{d}w(t)}{\mathrm{d}t}.$$

Since the reference points $(t_{S,i}; t_{T,i})$ describing $w(t)$ are neither strictly monotonic nor equally spaced, interpolation by linear regression lines is employed to determine an estimate of $r(t)$ at any arbitrary time instant. The estimate $\tilde{r}(t)$ is found by minimizing the weighted mean-square error

$$\sum_{i=1}^{N} g\left(t_{S,i}-t\right)\left[\,t_{T,i}-\left(\tilde{r}(t)\,t_{S,i}+\tilde{c}(t)\right)\right]^2 \xrightarrow{\tilde{r}(t),\tilde{c}(t)} \min.$$

yielding [6]

$$\tilde{r}(t) = \frac{\sum\limits_i g_i \cdot \sum\limits_i g_i\, t_{S,i}^2 - \left[\,\sum\limits_i g_i t_{S,i}\right]^2}{\sum\limits_i g_i \cdot \sum\limits_i g_i\, t_{S,i} t_{T,i} - \sum\limits_i g_i\, t_{S,i} \cdot \sum\limits_i g_i\, t_{T,i}},$$

where

$$g_i = g(t_{S,i}-t) = \begin{cases} 1-\left|\frac{2(t_{S,i}-t)}{T_D}\right|, & t-\frac{T_D}{2} \le t_{S,i} \le t+\frac{T_D}{2} \\ 0\,, & \text{otherwise} \end{cases}$$

denotes a triangular window centered around the current value of t. A choice of $T_D = 200$ms was found to be an adequate compromise between time resolution and smoothness of the speech-rate contour.

2.2 Measurements with the Speech Corpus

Five speakers (two female: k04, k06 and three male: k05, k61, k65) reading 100 sentences each were chosen from a labelled speech corpus (“Kiel Corpus of Read Speech”[4]) for evaluation of the rate adaptation module. For the pair of source and target speaker in Figure 2, the estimated relative speech-rate contour $\tilde{r}(t)$ is shown in subplot 2-c. Apparently, the contour $\tilde{r}(t)$ coincides chiefly with the phoneme duration patterns in the spectrograms. The speaker dependency of the $\tilde{r}(t)$-contour is illustrated on a different sentence from the corpus for all four target speakers (with reference to k61) in Figure 3. Except for the fact that all target speakers begin the utterance at a significantly higher speech rate, no further distinct similarities can be directly found. That these variations in relative speech rate are not only random but to some extent speaker characteristic can be seen from Table 1 where the mean global speech rate and the mean (local) speech rate of selected phonemes, averaged over all 100 sentences, are given. Although the global speech rates vary only by a few percent, some of the speakers exhibit a significantly differing speech rate for some phonemes *on average*. However, the *exact prediction* of the target speaker dependent part of the $\tilde{r}(t)$—influenced by a multitude of factors—is beyond the scope of this paper. Instead, the focus is on the subsequent adaptation of the speech rate during speech synthesis without affecting the speech quality.

3 RATE ADAPTATION MODULE

3.1 Analysis-Synthesis-System

The adaptation module for the speech rate is embedded into the MPEG-4 HVXC decoder (see Figure 4), which is part of the overall analysis-synthesis system for extraction and resynthesis of speaker-characteristic parameters. Two different schemes are employed for the framewise

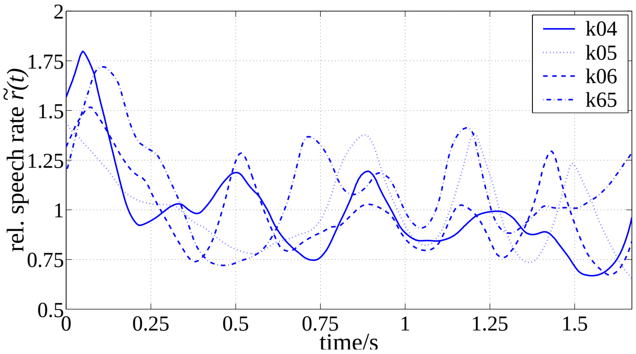

Figure 3: Estimated speech-rate contours $\tilde{r}(t)$ for sentence BE013 with reference to target speaker k61.

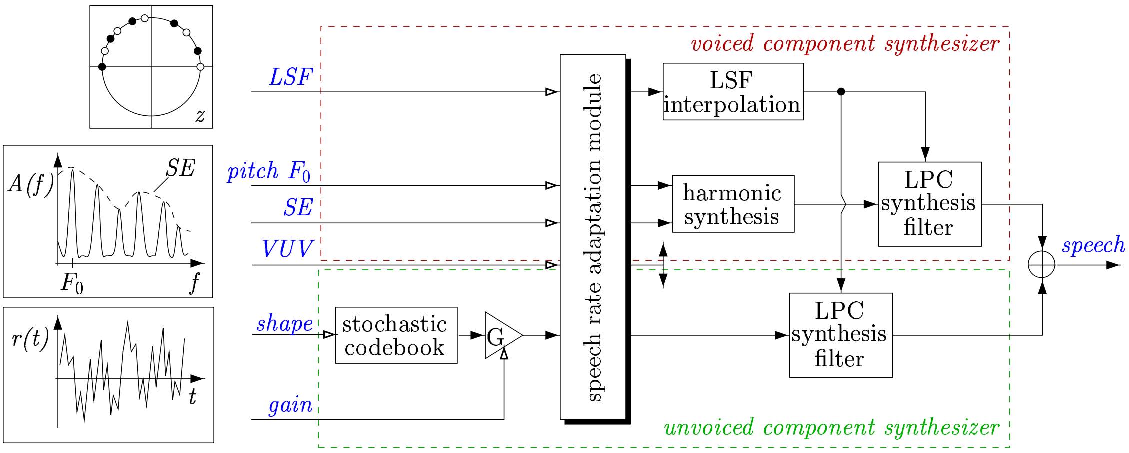

Figure 4: Simplified scheme of the MPEG-4 HVXC decoder with speech rate adaptation module.

(length: 20 ms) speech synthesis (voiced/unvoiced component synthesizer)[3]: speech frames classified as voiced are synthesized by a harmonic scheme in which the residual signal is represented by a harmonic spectral pulse-train (at multiples of the pitch frequency F0) shaped with the spectral envelope SE. In addition, bandlimited colored Gausssian noise is added, depending on the degree of voicedness VUV. In the case of an unvoiced speech frame, the speech signal is represented by a CELP-like (Code Excited Linear Prediction) scheme in which the residual signal is generated in time-domain by a stochastic codebook vector and a gain factor. The synthesis is accomplished by an LPC-filter derived from interpolated Line Spectrum Frequencies (LSF). Continuity of the transitions between adjacent parameter sets is preserved by linear interpolation of the above speech parameters.

3.2 Adaptation Algorithm

The favourable interpolation properties of the HVXC speech parameters are utilized in the speech rate adaptation module. It was derived from the speed module of the HVXC [3] originally intended to provide a “fast-forward”-functionality. The principal procedure of the rate adaptation in conjunction with voice conversion is shown in Figure 5. The speech of the source speaker is analysed framewise around equally spaced time instances $t_S = t_{S,\lambda}$ and for each analysis frame the speech parameters described in the above section, here summarized by $P_S(t_{S,\lambda})$, are extracted. In the voice-conversion step the source parameters $P_S(t_{S,\lambda})$ are mapped onto the corresponding target parameters $P_T(t_{S,\lambda})$. Synthesizing the target speech with these parameter sets would yield a *target* speaker utterance using the speech-rate contour of the *source* speaker since speech-rate variations have not been considered yet. Therefore in the next step, the relative speech rate (with reference to the source speaker) is adapted subsequently. The synthesis of the rate adapted speech is realized on a different time grid at equally spaced time instances $t_T = t_{T,\mu}$, where the synthesis frame length equals the analysis frame length, i.e. $t_{T,\mu} - t_{T,\mu-1} = t_{S,\lambda} - t_{S,\lambda-1} = 20$ms. The corresponding time instances $t'_{S,\mu}$ on the source speaker axis can be found by the (estimated) inverse warping function $t'_{S,\mu} = \tilde{w}^{-1}(t_{T,\mu})$ which derives from the estimated relative speech rate $\tilde{r}(t)$ by

$$\tilde{w}^{-1}(t) = \left[\int_0^t \frac{1}{\tilde{r}(\tau)} \mathrm{d}\tau\right]^{-1}.$$

In general, the time instances $t'_{S,\mu}$ do not coincide with the time instances $t_{S,\lambda}$ for which the target parameters $P_T(t_{S,\lambda})$ are known. Therefore, the parameter sets $P_T(t'_{S,\mu})$ are determined by linear interpolation. Assuming $t_{S,\lambda-1} \leq t'_{S,\mu} \leq t_{S,\lambda}$, then the interpolated parameter set is calculated by

$$P_T(t'_{S,\mu}) = \frac{t'_{S,\mu} - t_{S,\lambda-1}}{t_{S,\lambda} - t_{S,\lambda-1}} P_T(t_{S,\lambda-1}) + \frac{t_{S,\lambda} - t'_{S,\mu}}{t_{S,\lambda} - t_{S,\lambda-1}} P_T(t_{S,\lambda}).$$

As for the unvoiced residual signal, instead of interpolating the stochastic signals according to the above scheme, Gaussian noise of the same energy as the residual signal centered around $t'_{S,\mu}$ is inserted. In case of transients from voiced to unvoiced frames and vice versa, interpolation is not possible and the parameter set $P_T(t_{S,\lambda})$ being closest to $t'_{S,\mu}$ is simply repeated.

	speaker				
	k04	k05	k06	k65	k61
rel. global rate	1.05	1.04	0.94	0.93	1.00
/a/	0.96	0.98	1.03	0.86	1.00
/e/	0.83	0.90	0.98	0.81	1.00
/O/	0.94	0.87	0.94	0.77	1.00
/u/	0.75	0.81	0.91	0.71	1.00
/U/	1.06	0.88	1.01	0.95	1.00
/@/	1.06	0.84	0.93	1.03	1.00
/s/	0.83	0.85	1.04	1.14	1.00
/z/	0.82	0.95	1.03	1.12	1.00
/S/	0.91	0.87	1.09	1.00	1.00

Table 1: Mean relative global speech rate and mean relative speech rate of selected phonemes (normalized by the relative global speech rate) with reference to speaker k61 and averaged over 100 sentences.

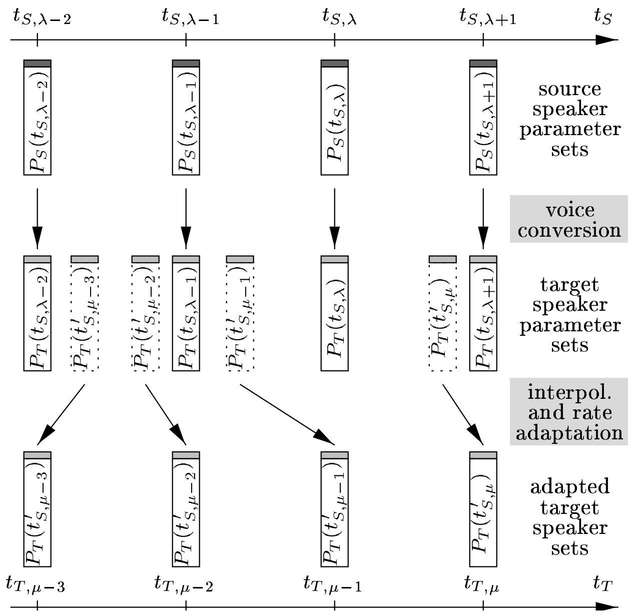

Figure 5: Schematic illustration of the speech rate adaptation module.

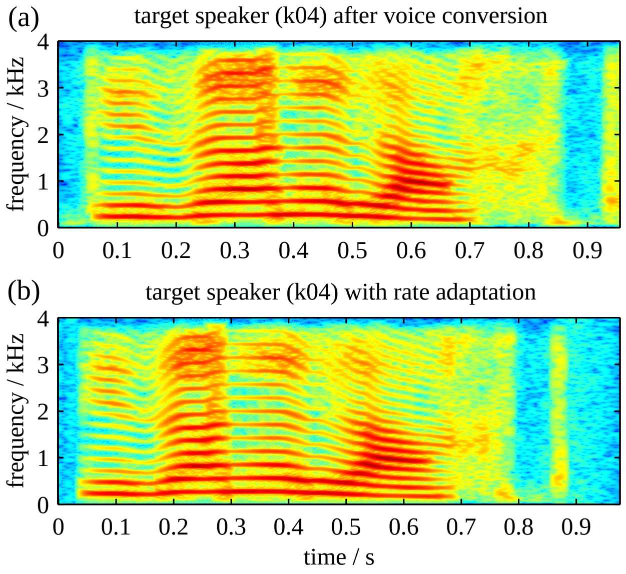

Figure 6: Spectrograms of the target speaker k04 generated from k61 (a) without and (b) with speech-rate adaptation.

3.3 Results

The proposed speech rate adaptation module was evaluated with test sentences from the Kiel Corpus. "Ideal" voice conversion was applied to let the speech quality remain unaffected by the voice conversion step: Parameter sets $P_S(t_{S,\lambda})$ were extracted from the source-speaker utterances and mapped onto corresponding sets of the target speaker $P_T(t_{S,\lambda})$ which were derived directly from the original target utterance. Figure 6-a shows the result of this ideal voice conversion without rate adaptation for the example pair of source and target speakers in Figure 2. Though the spectrogram in Figure 6-a exhibits the spectral characteristics of the target speaker k04, the time patterns are still those of the source speaker k61.

In Figure 6-b the speech-rate adaptation was applied. It can be clearly seen that both, the global (duration of the utterance) and the local speech rate match again the original utterance in Figure 2-b, while spectral and pitch characteristics are preserved. Informal listening tests showed that the speaking style of the speaker with rate adaptation is judged to be closer to the original utterance than without. A quality decrease is not percievable except in some cases when stretching of unvoiced phonemes by a large factor leads to slight roughness or unnaturalness. Further steps will be taken in future work to improve the rate adaptation for unvoiced speech segments.

4 CONCLUSIONS

A system for adaptation of the relative speech rate, characteristic for the speaking style of a person to some extend, was proposed in this work. The relative speech rate between source and target speaker, considered in a voice-conversion process, was determined by a method using dynamic-time warping and weighted interpolation with regression lines. The speech rate adaptation module, embedded into the standard speech coder MPEG-4 HVXC, was shown to be capable of adjusting the correct speech rate of the target speaker by taking advantage of the interpolation properties of the harmonic synthesis in the HVXC with almost no quality decrease. This system for determination and adaptation of the relative speech rate can be seen as the starting point for a derivation of general rules for prediction of the speaker-dependent part in the speech-rate contour in future work.

REFERENCES

[1] A. Kain and M. W. Macon, "Spectral Voice Conversion for Text-To-Speech Synthesis", *Proceedings of ICASSP 1998*, vol.1, pp. 285–288, 1998.

[2] H. Kuwabara and Y. Sagisaka, "Acoustic characteristics of speaker individuality: Control and conversion", *Speech Communication*, vol.16, No. 2, pp. 165–173, 1995.

[3] MPEG-4, "Information Technology – Very Low Bitrate Audio-Visual Coding, Part 3: Audio, Subpart 2: Parametric Coding", *ISO/IEC 14496-3*, May 2000.

[4] K.J. Kohler, "Labelled data bank of spoken standard German: The Kiel Corpus of Read/Spontaneous Speech",Proc. ICSLP'96, Philadelphia, Pennsylvania, USA, 1996

[5] L. Leutelt, "An MPEG-4 HVXC Based Interspeaker Pitch Prediction for Voice Conversion", Proc. ECMCS'01, pp. 313-316, Budapest, Hungary, 2001

[6] S. Ohno and H. Fujisaki, "A Method for Quantitative Analysis of the Local Speech Rate", *Proceedings of Eurospeech 1995*, pp. 421–424, 1995.

[7] H.R. Pfitzinger, "Two Approaches to Speech Rate Estimation", Proc. SST, pp. 421–426, Adelaide,1996