VOICE CONVERSION BASED ON SPECTRAL ENVELOPE TRANSFORMATION

Robert Vích and Martin Vondra

Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic
vich@ure.cas.cz

Abstract: In this paper a new voice conversion algorithm is presented, which transforms the utterance of a source speaker into the utterance of a target speaker or into the utterance of a new unknown speaker. The voice conversion algorithm is based on spectral speech analysis, frequency transformation, spectral envelope warping, spectrum interpolation and parametrical high quality IIR or FIR cepstral speech synthesis. The cepstral speech model is realized using short time discrete Fourier transform of overlapping pitch-asynchronous speech frames and on speech deconvolution in the cepstral domain. Cepstral speech synthesis is implemented pitch-synchronous and in contrary to the LPC speech model, the cepstral speech model is of the pole/zero type and contains also information about the vocal tract excitation. Several approaches to frequency transformation of the speech spectrum are compared, e.g. linear frequency scaling, piecewise linear frequency warping and nonlinear lowpass to lowpass frequency transformation. The type of spectral warping depends on the wanted accuracy of the formant mapping of the source into the target spectrum. Prosodic transformations, i.e. fundamental frequency, time and intensity scale modifications are also shortly mentioned.

1 Introduction

Voice conversion is mainly applicable in Text-to-Speech (TTS) systems to generate different types of output voice without creating new speech databases from several speakers and without time consuming inventory labeling. Further applications are also meaningful, e.g. for change of speaker identity in e-mail reading adapted to senders voice or in voice-mail, in low-bandwidth speech coding not preserving the speaker's identity, in improving the intelligibility of abnormal uttered speech, in designing hearing aids appropriate for specific hearing problems, in language learning and last but not least in speech recognition.

State of the art of voice conversion was the topic of the special issue of Speech Communication in 1995 [1]. An extensive bibliography on voice transformation may be found in the dissertation of Kain in [2]. Voice conversion is also several years in the centre of interest in the Institute of Radio Engineering and Electronics of the Czech Academy of Sciences in Prague. Different approaches have been studied, e.g. using spline interpolation and harmonic speech modeling [3], by decimation and interpolation of the speech signal and cepstral speech synthesis [4] and last but not least using PSOLA and resampling [5]. In [6] a nonlinear frequency scale mapping for voice conversion combined with spline interpolation and implemented in the harmonic speech model was presented.

The voice conversion approach proposed in this paper is based on spectral speech analysis and cepstral speech synthesis. The cepstral vocoder with voice conversion is shown in Fig. 1. There are the speech analysis and speech synthesis blocks. Speech analysis is performed pitch-asynchronous. Speech spectrum envelope is estimated using cepstral deconvolution. The fundamental frequency is determined by looking for the location of the main peak of the autocorrelation and of the real cepstrum. For speech reconstruction the composite cepstral model is applied excited for voiced speech by an impulse generator and for unvoiced speech by a white noise generator [7, 8].

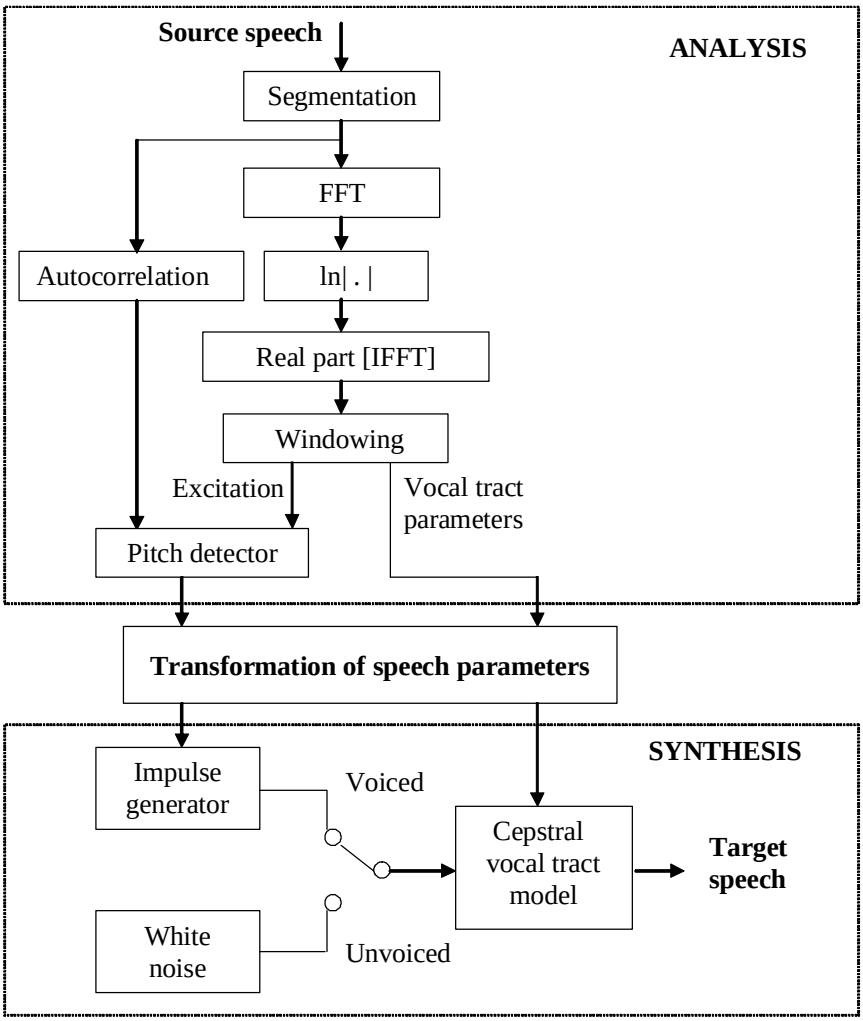

Figure 1 - The cepstral vocoder with speech parameter transformation.

In this paper several approaches to frequency warping of the speech spectrum envelope are compared, e.g. linear frequency scaling, piecewise linear and nonlinear lowpass to lowpass frequency transformation. The type of spectral warping depends on the wanted accuracy of the formant mapping of the source into the target speaker.

The fundamental frequency of the utterance in the proposed voice conversion algorithm is scaled by matching the average and variance of the fundamental frequency of the source and the target speaker. Time and intensity scale modifications of the output speech can also be simply realized thanks to the parametric description of the cepstral speech model. In TTS application the time-scale modification may be implemented speech element adaptive.

The proposed voice conversion algorithm can be incorporated in a vocoder as an output block or in TTS applications it can be applied for the conversion of the speech element inventory prior to its parameterization. The different proposed voice conversion approaches will be illustrated by audio presentations.

2 Transformation of the speech spectrum envelope

The short-time speech spectrum envelope is the basic segmental characteristic not only in cepstral speech synthesis. It can be characterized using resonances/formants and antiresonances/antiformants, and their resonant frequencies, bandwidths and relative amplitudes. The envelope of the short-time speech spectrum depends on the dimension and on the adjustment of the human vocal tract. Great differences in the vocal tracts are first of all between male, female and childish, but differences exist also within these classes. It can be

stated that the formant frequencies for female voices are approximately 16-20 % higher than that for male voices. In Fig. 2 the short-time LPC speech spectrums corresponding to the vowel “e” for a male and female speaker are depicted. The different shape and the differences in formants can be seen at first glance. The speech spectrum envelope together with the speaker’s fundamental frequency, are the main parameters characteristic for different voices.

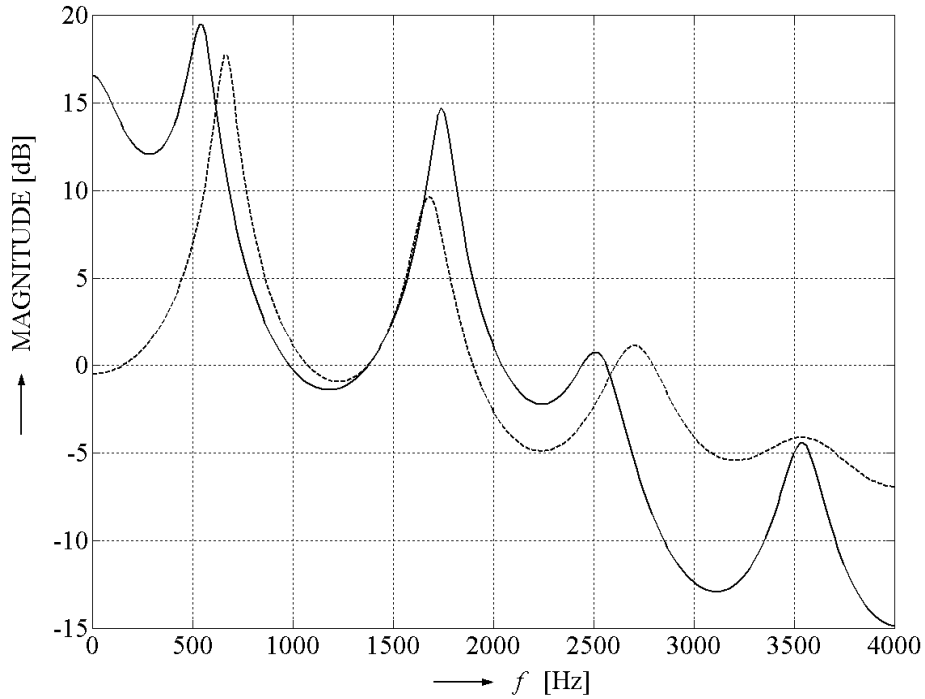

Figure 2 – Short-time spectrum envelope of the stationary part of the vowel „e“ for male (solid line) and female (dashed line) voice.

In this contribution a new approach to voice conversion by frequency warping of the speech spectrum envelope is described. Frequency warping has the greatest influence on the voice change. The change of the spectrum frequency tilt given by the relative amplitudes of formants will not be solved.

The warping of the speech spectrum envelope is given by the requirement

$$|S_T(f)| = |S_S(F)|. \tag{1}$$

$|S_T(f)|$ and $|S_S(F)|$ are the short time spectrum envelopes of the *target* speaker and the *source* speaker, respectively. The variables f and F are the corresponding frequency variables. In speech spectrum analysis using DFT the source spectrum $|S_S(F)|$ is estimated equidistantly for

$$F_k = k\frac{F_S}{N_F}, k = 0,1,2...(N_F - 1). \tag{2}$$

F_S is the sampling frequency and N_F is the dimension of the applied DFT.

The transformation of the source speaker spectrum $|S_S(F)|$ into the target speaker spectrum $|S_T(f)|$ is accomplished by mapping of the variable F into the variable f ; i.e.

$$f = Q(F). \tag{3}$$

This function may be given numerically by a table or analytically and generally it is nonlinear. That means that even if F is equidistantly sampled, f is not equidistantly

sampled. Therefore the transformed $|S_T(f)|$ must be interpolated for further application at points

$$f_k = k\frac{F_S}{N_F}, k = 0,1,2...(N_F - 1). \quad (4)$$

2.1 Piecewise linear frequency warping

Let us suppose that the mapping function $f = Q(F)$ is given, for example, by some chosen formant frequencies of the source and target speaker. We shall call these points as significant frequencies. The number of significant frequencies is $(M+1)$ and is equal for the source and the target speaker. Further we shall assume that the boundary significant frequencies $F = 0$ and $F = F_M = F_S/2$ are mapped into $f = 0$ and $f = f_M = F_S/2$, respectively.

Let the sequence of significant frequencies for the source speaker be $\{F_l\} = \{0, F_1, F_2, ...F_l, ...F_M\}$ and for the target speaker $\{f_l\} = \{0, f_1, f_2, ...f_l, ...f_M\}$, $l = 0,...M$.

Then, between the significant frequencies F_l and F_{l+1} of the source speaker the transformed frequency f can be linearly interpolated using the relation

$$f = \frac{f_{l+1} - f_l}{F_{l+1} - F_l}(F - F_l) + f_l. \quad (5)$$

The index l changes in the interval $l = 0,...M-1$. The complete frequency mapping $f = Q(F)$ is given by composition of the individual interpolated frequency intervals f.

This piecewise frequency warping together with the corresponding spectrum transformation is shown in Fig. 3 for $M = 5$.

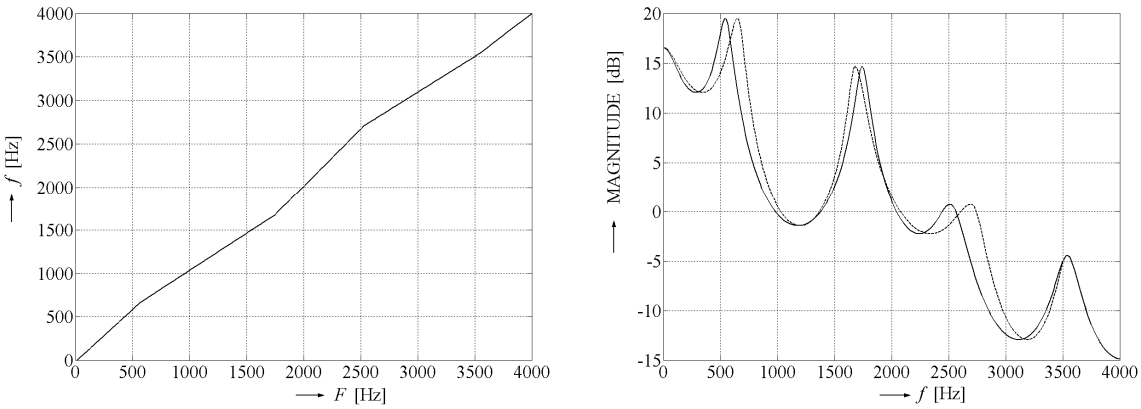

Figure 3 - Piecewise frequency warping for *M*=5 together with the corresponding spectrum transformation. Source spectrum is shown by solid line, target spectrum by dashed line.

2.2 Linear frequency scaling

Linear frequency scaling is a special case of piecewise linear frequency warping for $M = 1$. From (5) follows

$$f = \frac{f_1}{F_1}F = KF, \quad (6)$$

where $K = \frac{f_1}{F_1}$ is the scaling parameter for linear frequency warping.

In Fig. 4 linear frequency scaling for $K > 0$ together with the corresponding spectrum transformation is shown. For $K > 1$ the last part of the source spectrum is not transformed, it

is omitted. For $K<1$ the last part of the source spectrum must be extrapolated in a convenient way, for instant set equal to a small constant value.

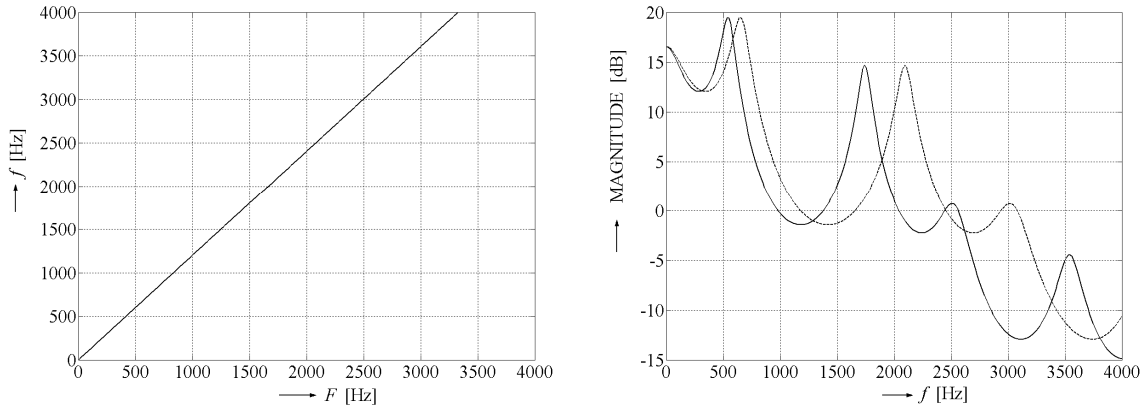

Figure 4 – Linear frequency scaling together with the corresponding spectrum transformation. Source spectrum is shown by solid line, target spectrum by dashed line.

2.3 Nonlinear frequency warping

In general frequency transformations of digital filters can be obtained by using transformations similar to the bilinear transformation [9]. In our case of voice conversion we need not to transform the transfer function of the digital vocal tract model, we may apply only the frequency warping corresponding to the allpass transfer function

$$Z = \frac{z-a}{1-az}. \tag{7}$$

This transfer function transforms a lowpass filter with the cutoff frequency $F = F_c < F_S/2$ - the source filter - into a new lowpass filter – the target filter - with the cutoff frequency $f = f_c < F_S/2$. Setting e.g. the cutoff frequency F_c equal to the 1st formant frequency of the source speaker, i.e. $F_c = F_1$ and the cutoff frequency f_c equal to the 1st formant frequency of the target speaker, i.e. $f_c = f_1$, then the transformation parameter a is given by

$$a = \frac{\sin(p(F_1 - f_1)/F_S)}{\sin(p(F_1 + f_1)/F_S)}. \tag{8}$$

The frequency warping function $f = Q(F)$ follows after some manipulation

$$f = F + \frac{F_S}{p}\arctan\frac{a\sin(2pF/F_S)}{1-a\cos(2pF/F_S)}. \tag{9}$$

This lowpass-to-lowpass nonlinear frequency warping together with the corresponding spectrum transformation is depicted in Fig. 5.

3 Transformation of prosodic parameters

In addition to the speech spectrum envelope transformation described in Chapter 2 the fundamental frequency F_0, the speech rate and speech intensity of the source speaker are also modified to mimic that of the target speaker. These parameters of the speech model can be modified simply thanks to the parametric description of the cepstral speech model.

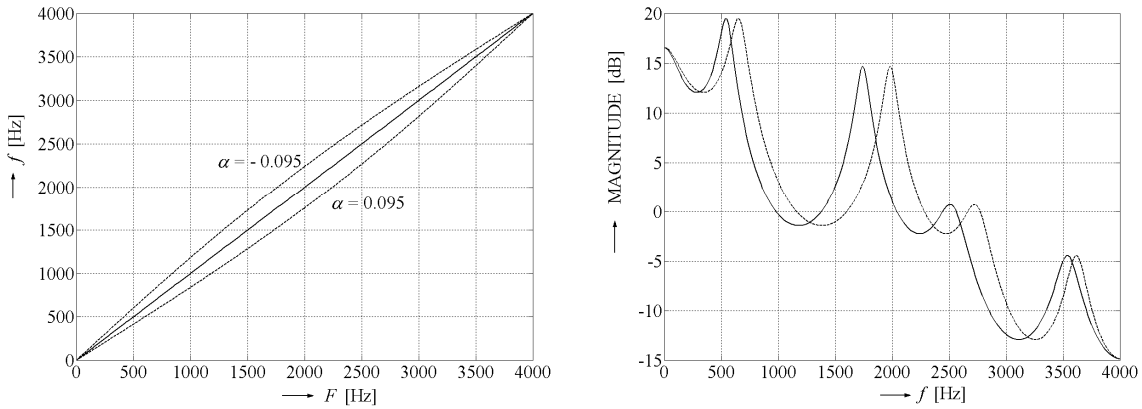

Figure 5 – Nonlinear frequency warping using lowpass-to-lowpass frequency transformation together with the corresponding spectrum transformation. Source spectrum is shown by solid line, target spectrum by dashed line.

3.1 F_0 modification

The fundamental frequency F_0 modification applied in the described speech conversion system matches the means and standard deviations of the source and target speakers [10]. This can be realized using

$$F_{0T}(t) = aF_{0S}(t) + b\,. \tag{10}$$

$F_{0T}(t)$ denotes the target instantaneous fundamental frequency and $F_{0S}(t)$ is the source instantaneous fundamental frequency,

$$a = \sqrt{s_{TF}^2 / s_{SF}^2} = s_{TF} / s_{SF}\,,\; b = m_{TF} - am_{SF}\,. \tag{11}$$

The values s_{TF}^2, s_{SF}^2 and m_{TF}, m_{SF} correspond to the F_0 variances and mean values for the target and the source speakers, respectively. This expression may be rearranged into the following form

$$F_{0T}(t) = a(F_{0S}(t) - m_{SF}) + m_{TF}\,. \tag{12}$$

3.2 Speech rate modification

Speech rate modification may be realized by changing the duration of the utterance. The basic assumption of this approach is that it does not affect the speech spectrum envelope and the fundamental frequency F_0. In cepstral speech synthesis uniform speaking rate modification can be simply performed by multiplication of the analysis frame length N by a factor b. The speech rate decreases for $b > 1$, for $b < 1$ the speech rate increases. The parameter b can be estimated as the ratio of the duration of the target speaker's utterance T_T to that of the source speaker's T_S, i.e. $b = T_T / T_S$ or as the ratio of the mean values of the target and source speaker durations of phonemes. In this way the speech rate of the source speaker can be adapted to that of the target speaker.

This speech rate modification is in general not convenient first of all for large $|b|$ because it does not respect the duration differences between phonemes and pauses and also does not take into account the duration characteristics of speakers given by context, accent and dialect. In TTS application of speech conversion the duration transformation can be implemented speech inventory dependent by rules [11] or phoneme adaptive using neural networks.

3.3 Intensity modification

The instantaneous speech intensity of the source speaker given by the gain factor $I_S(t)$ of its cepstral speech model may also be adapted in the same manner as the fundamental frequency to the speech intensity of the target speaker. It holds

$$I_T(t) = cI_S(t) + d\,, \tag{13}$$

where $I_T(t)$ are the target and $I_S(t)$ the source instantaneous gain factors,

$$c = \sqrt{s_{TI}^2 / s_{SI}^2} = s_{TI} / s_{SI}\,,\quad d = m_{TI} - cm_{SI}\,. \tag{14}$$

The values s_{TI}^2, s_{SI}^2 and m_{TI}, m_{SI} correspond to $I_T(t)$ and $I_S(t)$ variances and mean values of the target and the source speakers, respectively.

For equal mean speech intensity of the target and source speaker $m_{TI} = m_{SI} = m_I$ it approximately holds

$$I_T(t) = cI_S(t) + m_I(1-c)\,. \tag{15}$$

4 Experiments

Experiments have been performed with voice conversion from male to female and to childish voices and vice versa. Records of identical utterances of several male and female voices have been analyzed and the F_0 and intensity contours have been estimated. From the stationary parts of vowels the formant frequencies have been estimated and used for transformation of the speech spectrum envelope. The fundamental frequency modification matching the F_0 variances and mean values for the target and the source speakers described in Section 3.1 has been applied. The experiments have been realized using the cepstral vocoder combined with voice conversion for 16 kHz sampling frequency.

In Fig. 6 the spectrograms of female voice and of the converted male voice are shown.

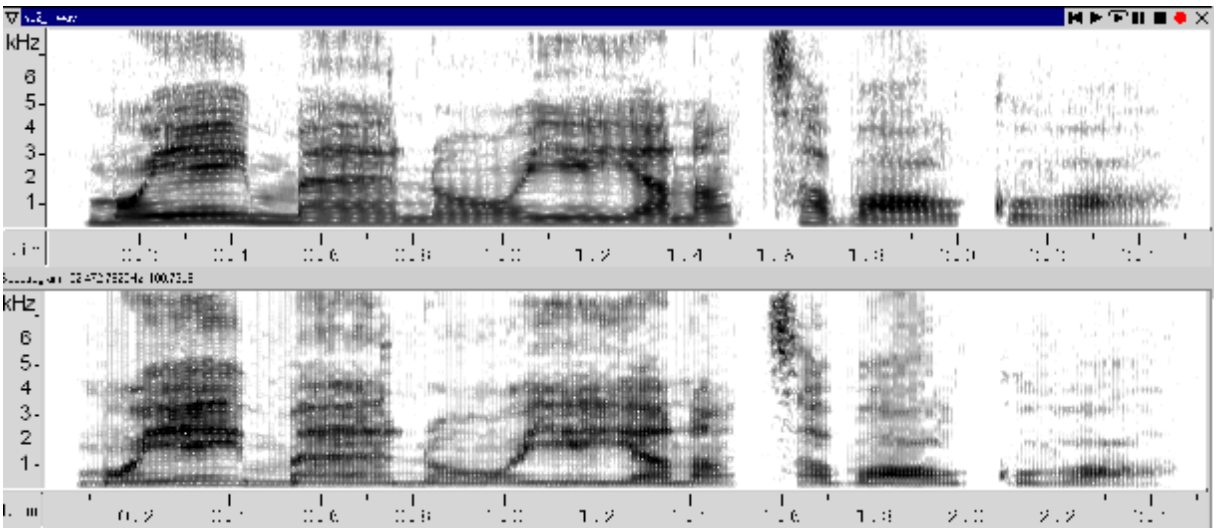

Figure 6 - Spectrogram of the source female speech (upper part) and the spectrogram of the output of the female-male voice conversion by nonlinear frequency warping and fundamental frequency matching (lower part).

5 Conclusion

In all solved examples only one transforming function for all phonemes in the utterance has been used without respect to the phonetic content. The subjective voice conversion quality has

been judged quite well. The algorithm is, in contrary to other approaches, numerically modest and does not require a training phase. The total similarity of the transformed speech to that of the target speaker is not achieved because the transformation parameters approach the optimal parameter set only in the mean.

The evaluation of speech conversion was performed by subjective listening tests. The differences between the described approaches of speech spectrum envelope transformation are very subtle. Nevertheless, better results have been achieved by piecewise linear and nonlinear spectrum envelope transformation than with linear frequency scaling. The identity of the transformed voice is perceived as different from that of the source speaker and close to that of the target speaker. Generally, better results have been obtained with female-male voice conversion than in the reverse direction.

Acknowledgements

This research has been supported by the Grant Agency of the Czech Republic (GA102/02/0124 "Voice Technologies for Support of Information Society") and by the Ministry of Education, Youth and Sports of the Czech Republic (OC 277.001 "Transformation of Segmental and Suprasegmental Speech Models").

References

[1] Moulines, E., Sagisaka, Y. (Eds.): Voice Conversion: State of the Art and Perspectives. Special issue of Speech Communication, Vol. 16, No. 2, February 1995.

[2] Kain, A. B.: High Resolution Voice Transformation. PhD Thesis, Oregon Graduate Institute of Science and Technology, October 2001.

[3] Přibilová, A.: Speech Spectrum Envelope Modification. In: R. Vích (Ed.): Proc. of the 13th Czech-German Workshop on Speech Processing, Prague, September 15-17, 2003, pp. 30-37.

[4] Vondra, M.: Voice Transformation in Parametric Speech Synthesis. In: R. Vích (Ed.): Proc. of the 13th Czech-German Workshop on Speech Processing, Prague, September 15-17, 2003, pp. 35-37.

[5] Nemšák, S.: Pitch Shifting and Voice Transformation Using PSOLA. In: R. Vích (Ed.): Proc. of the 13th Czech-German Workshop on Speech Processing, Prague, September 15-17, 2003, pp. 38-41.

[6] Přibilová, A., Vích, R.: Non-Linear Frequency Scale Mapping for Voice Conversion. In: Proc. of the 14th International Czech-Slovak Scientific Conference Radioelektronika 2004, Bratislava, Slovak Republic, April 27–28, 2004, pp. 100–103.

[7] Vích, R.: Cepstrales Sprachmodell, Kettenbrüche und Anregungsanpassung in der Sprachsynthese. Wissenschaftliche Zeitschrift der Technischen Universität Dresden, Vol. 49, No. 4/5, 2000, pp. 116 -121.

[8] Vondra, M., Smékal, Z.: Composite Cepstral Models for TTS Synthesis. In: R. Vích (Ed.): Speech Processing, Proc. of the 11th Czech-German Workshop on Speech Processing, Prague, September 17-19, 2001, pp. 76-78.

[9] Oppenheim, A. V., Schafer, R. W: Discrete-Time Signal Processing. Prentice-Hall, 1989.

[10] Arslan, L. M.: Speaker Transformation Algorithm using Segmental Codebooks (STASC). Speech Communication, Vol. 28, 1999, pp. 211-226.

[11] Horák, P.: Rule Based Sounds Duration Model for the Czech TTS System. In: 15. Konferenz Elektronische Sprachsignalverarbeitung ESSV 2004, Cottbus, 20.-22. September 2004.