Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 48 (2006) 1691–1703
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description

Anna Přibilová [a,*], Jiří Přibil [b]

[a] Department of Radio Electronics, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovakia
[b] Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic, Chaberská 57, 182 51 Praha 8, Czech Republic

Received 23 February 2006; received in revised form 30 June 2006; accepted 3 August 2006

Abstract

Voice conversion, i.e. modification of a speech signal to sound as if spoken by a different speaker, finds its use in speech synthesis with a new voice without necessity of a new database. This paper introduces two new simple non-linear methods of frequency scale mapping for transformation of voice characteristics between male and female or childish. The frequency scale mapping methods were developed primarily for use in the Czech and Slovak text-to-speech (TTS) system designed for the blind and based on the Pocket PC device platform. It uses cepstral description of the diphone speech inventory of the male speaker using the source-filter speech model or the harmonic speech model. Three new diphone speech inventories corresponding to female, childish and young male voices are created from the original male speech inventory. Listening tests are used for evaluation of voice transformation and quality of synthetic speech.

Keywords: Speech synthesis; Source-filter speech model; Harmonic speech model; Text-to-speech system; Voice conversion

1. Introduction

Text-to-speech (TTS) synthesis is always nationally oriented. It holds much more for the Czech and Slovak languages with their complicated structure and diacritics. Our phonetic school has a long tradition (Janota, 1967; Janota et al., 1994; Janota and Vích, 1994). The first real-time Czech and Slovak TTS system had been developed in 1980s (Vích and Horák, 1996; Duběda et al., 2005).

Synthetic speech quality evaluation using listening tests must be performed by native Czech and Slovak people too.

From the beginning we had been orientated at construction of special aids for blind and partially sighted people based on the TTS systems. Later, we concentrated on the applications in the area of telecommunication services. Recently, the cooperative of blinds SPEKTRA addressed us to implement the Czech and Slovak TTS engine into the special aids – Braille notetaker BraillePen. The concept of the BraillePen software Easylink is based on the wireless BlueTooth connection between the small and handy Braille keyboard and the Pocket PC (PPC).

* Corresponding author. Tel.: +421 2 602 91 743; fax: +421 2 654 29 683.
E-mail address: pribilova@kre.elf.stuba.sk (A. Přibilová).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.08.001

According to the specific PPC properties (the size of the usable RAM memory and the used special main processor with lower computational performance) we must totally rebuild and simplify our last Czech and Slovak TTS system together with the required implementation of additional voices – new female and childish apart from the basic male voice necessary for the users of our TTS system (mainly the blind people). However, TTS systems typically generate speech by one speaker. The cost of employing more speakers is high because a separate database must be recorded, segmented and processed for each speaker. Voice conversion, i.e. modification of a speech signal to sound as if spoken by a different speaker may help create a speech inventory of a new voice without necessity of a new database. This is also our case and the reason why we have introduced new methods for easy creation of different voices in the frame of the TTS system.

Our paper describes the realized voice conversion methods for the Czech and Slovak TTS system. Section 2 presents an overview of voice conversion methods and an explanation of our approach to solution of the requirement to have several different synthetic voices without need for the particular voice identity using the general knowledge of Stevens (1997), Fant (1997). Then, we compare a linear method (Section 3) and two novel non-linear methods (Section 4) of frequency scale mapping for voice conversion with increasing or decreasing formants. The first non-linear method uses frequency scale mapping function as a circle arc. The second method uses a constant formant modification factor in a given frequency interval. Non-linear modification of male voice cepstral speech parameters is performed to create cepstral description files for new voices (Section 5). The original or modified cepstral description of the speech database is used in the Czech and Slovak TTS system. Synthetic speech with different voices is generated by means of the source-filter or the harmonic speech model with cepstral description (Section 6). No prosody transformation is discussed in our paper, as the prosody according to rules is generated by the TTS system for a given basic pitch frequency of a synthetic voice. The realized voice conversion methods are evaluated with the help of listening tests described in Section 7.

2. Overview of voice conversion methods

Voice conversion methods are based on the knowledge of speaker individuality. A study on speaker acoustic characteristics can be found in (Kuwabara and Sagisaka, 1995). According to them the voice individuality is affected by the voice source (the average pitch frequency, the pitch contour, the pitch frequency fluctuation, the glottal wave shape) and the vocal tract (the shape of spectral envelope and spectral tilt, the absolute values of formant frequencies, the formant trajectories, the long-term average speech spectrum, the formant bandwidth). The most important factors on individuality are the pitch frequency and the resonance characteristics of the vocal tract, though the order of the two factors differs in different research studies.

Slifka and Anderson (1995) evaluated the mean and the standard deviation of the angle and the radius of the pole locations corresponding to the formant resonances for 20 vowel classes of the source and the target speakers. The current frame angle and radius were normalized with the source speaker statistics and the normalized values were further modified to have the target speaker mean and standard deviation of the angle and the radius. However, the listening tests showed that apparent speaker identity was altered, yet not clearly and consistently to the identity of the target.

Mizuno and Abe (1995) proposed a piecewise linear conversion rules controlling formant frequencies, formant bandwidths, and spectral intensity to produce speech with the desired formant structure. They used vector quantized utterances with the codebook size of 256. Three conversions were performed from one man to another two men and a boy. Conversion to a boy was clearly identified by 10 listeners, while conversion to one of the men was indistinct.

Narendranath et al. (1995) proposed an artificial neural network with one input layer, two hidden layers, and an output layer for formant transformation. A feed forward neural network was trained using the back propagation algorithm and it captured a function transforming formants of the source speaker to that of the target speaker. The network, trained with formants from the steady voiced regions, transformed formant transitions well for male to female voice conversion.

Iwahashi and Sagisaka (1995) presented a speech spectrum transformation with multiple linear interpolating functions weighted by radial basis function networks (RBFN) using the time-aligned spectral parameters (e.g. cepstrum) of the same utterance by multiple speakers. Two male and two female

speakers' data were pre-stored and only several words spoken by the target speaker were necessary.

Gaussian mixture models (GMM) were used to transform speech spectral envelopes parametrized by real cepstral coefficients (Stylianou et al., 1998; Ye and Young, 2003) or line spectral frequencies (LSF) (Kain and Macon, 1998) trained by aligned source and target data. Modification of the GMM estimation to include unaligned data was introduced by Duxans and Bonafonte (2003) and further improved for exclusively unaligned data by Sündermann et al. (2004). Mouchtaris et al. (2004) have solved a similar task using the maximum likelihood constrained adaptation. Toda et al. (2001) used the GMM-based algorithm with dynamic frequency warping to avoid the over-smoothing of speech transformation and representation using adaptive interpolation of weighted spectrum.

Gutiérrez-Arriola et al. (2001) applied a conversion of gain contour, pitch contour, glottal source Liljencrants–Fant (LF) model, and vocal tract (formants and bandwidths) by means of linear transformation in the multi-speaker formant synthesizer. The synthesizer stored all the parameters for the basic speaker and linear transformation functions for each stable segment of 455 units (phones, diphones, triphones) of two other speakers. All combinations between one female and two male speakers showed that higher formants were better converted than lower ones.

Rentzos et al. (2003) used the formant transformation by a two-dimensional phoneme-dependent hidden Markov models (HMM), glottal pulse LF model transformation, and pitch transformation based on time-domain pitch-synchronous overlap-and-add (TD-PSOLA) method. Poles of the LPC model were used for formant estimation. Listening tests of voice conversion between three male and two female speakers showed that formants were the most important in speaker perception, however, bandwidth, intensity and spectral tilt transformation improved it.

Lee et al. (2002) performed the transformation of acoustic and phone-specific parts of LPC cepstrum vectors with soft-clustering approach. As the acoustic variation represented speaker's inherent characteristics common in all phone units, the averaged LPC cepstrum was assumed to represent the acoustic variation, while mean subtracted LPC cepstrum was used to represent the phonetic characteristics. Conversion rules for the time-varying part were constructed by the classified linear transformation matrix based on soft clustering techniques for LPC cepstrum expressed in Karhunen–Loève coefficients. Eighteen listeners evaluated male-to-male and male-to-female voice conversion of one female and two male speakers. Majority of listeners identified the transformed speech as the target speaker's.

Transformation of vocal tract represented by LSF, excitation, intonation, energy, and duration characteristics was performed using segmental codebooks by Arslan (1999). Sentence HMMs were trained using a segmental k-means algorithm followed by Baum–Welch algorithm and the best state sequence for each utterance was estimated using the Viterbi algorithm. Subjective tests by three listeners showed that male-to-female transformation was better than male-to-male transformation using three speakers. The method was further modified for subband voice conversion by Turk and Arslan (2002). Discrete wavelet transform (DWT) was employed for subband decomposition to estimate the speech spectrum better with high resolution. Four subbands were employed and only the first subband was converted. The subband based voice conversion output was preferred over the full-band based output by majority of 20 listeners.

Vondra and Vích (2005a,b) used sound-dependent spectral warping for voice conversion. They proposed frequency transformation non-stationary in time with the least difference between the speech spectral envelopes of the source and the target speakers using vector quantization without necessity of phonetic segmentation. Voice conversion implemented in the cepstral vocoder was evaluated by 10 listeners. Four voice transformations between two male and two female voices were mostly identified, however, speech quality was little worse.

Leutelt and Heute (2004) analyzed interspeaker and intraspeaker variations for voice conversion. They observed that while the different kinds of intraspeaker variations follow roughly the same pattern, the interspeaker variations are smaller in the region of the first two formants and higher near the fourth formant. Intraspeaker variations can exceed interspeaker variations and hence limit the quality of voice conversion if not taken into consideration, as the variations of phonemes occurring at the same position in the corpus uttered by different speakers are smaller than variations of phonemes enclosed by different neighbour phonemes of the same speaker.

In the references presented so far, the conversion to a particular voice was studied. The source voice

characteristics and the target voice characteristics were used. However, in many TTS applications, the requirement is only to produce different voices without a determinate individuality. Then, statistical values of the pitch and the formant frequencies corresponding to different genders can be used. Stevens (1997) specified that the frequency of vibration of the vocal folds during normal speech production is usually in the range 80–160 Hz for adult males, 170–340 Hz for adult females, and 250–500 Hz for younger children. It means that female pitch frequencies are about twice the male pitch frequencies, pitch frequencies of younger children are about 1.5-times higher than those of females and about 3-times higher than those of males. As regards the formant frequencies, Fant (1997) stated that females have them on average 20% higher than males, but the relation between male and female formant frequencies is non-uniform and deviates from a simple scale factor. For our aim to convert a male voice so that it sounds as an indeterminate female voice (or vice versa), it is sufficient to consider the mean ratio between male and female pitch frequencies and formant frequencies.

3. Linear frequency scale mapping

Shifting the formants by a constant factor may be done via linear expanding or compressing the speech spectrum envelope $E(f)$ along the frequency axis. It can be called a linear frequency scale mapping of the speech spectrum envelope. It corresponds to

$$E'(f) = E(\eta(f)), \tag{1}$$

where $E'(f)$ represents the modified log spectrum envelope. The linear frequency scale mapping function $\eta(f)$ is given by

$$\eta(f) = \frac{f}{\gamma}, \tag{2}$$

where γ is the constant formant modification factor corresponding to the ratio of the target speaker frequency f_{target} and the source speaker frequency f_{source}:

$$\gamma = \frac{f_{\text{target}}}{f_{\text{source}}}. \tag{3}$$

For 20% formant shift from male to female voice $\gamma = 1.2$, and from female to male $\gamma = 1/1.2$ (Fig. 1(a) and (b)).

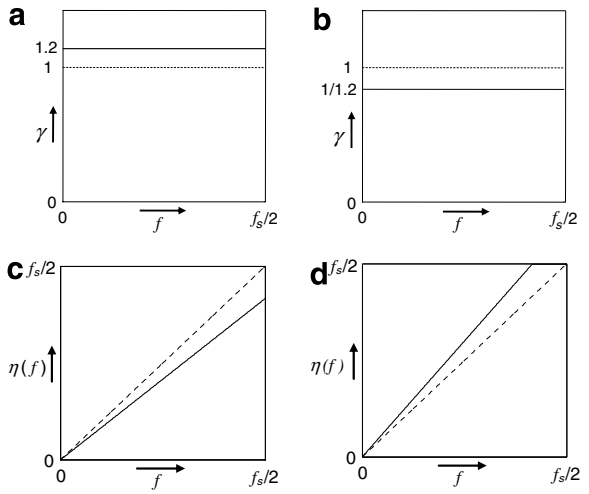

Fig. 1. Linear frequency scale mapping: constant formant modification factor $\gamma = 1.2$ (a), and $\gamma = 1/1.2$ (b), mapping function $\eta(f)$ for shifting formants to the right – expansion (c) and to the left – compression (d).

Fig. 1(c) represents a linear frequency scale mapping for $\gamma > 1$ (expansion of the speech spectrum envelope). The value of the modified spectrum envelope at the frequency f corresponds to the original envelope value at lower frequency, resulting in shifting formants to the right (compare Fig. 2(a) and (c)).

Fig. 1(d) corresponds to $\gamma < 1$ (compression of the speech spectrum envelope). Here the value of the modified spectrum envelope at the frequency f corresponds to the original envelope value at higher frequency, resulting in shifting formants to the left (compare Fig. 2(b) and (d)).

The disadvantage of the linear method is that for $\gamma > 1$ the frequencies higher than $f_s/2\gamma$ are missed (see Figs. 1(c) and 2(c)), and for $\gamma < 1$ the frequencies higher than $f_s/2\gamma$ must be padded by arbitrary values (see Figs. 1(d) and 2(d)). Furthermore, if N_F-point fast Fourier transform (FFT) is used for the original spectrum envelope computation, the number of points of the modified speech spectrum envelope is lower than N_F for $\gamma > 1$, and higher than N_F for $\gamma < 1$. To compute cepstral parameters using N_F-point FFT, the modified spectrum envelope would have to be resampled so that its length would be N_F.

4. Non-linear frequency scale mapping

To overcome the mentioned disadvantages of the linear method, new non-linear methods of mapping between the input frequency scale f and the output frequency scale $\eta(f)$ were proposed. The boundary

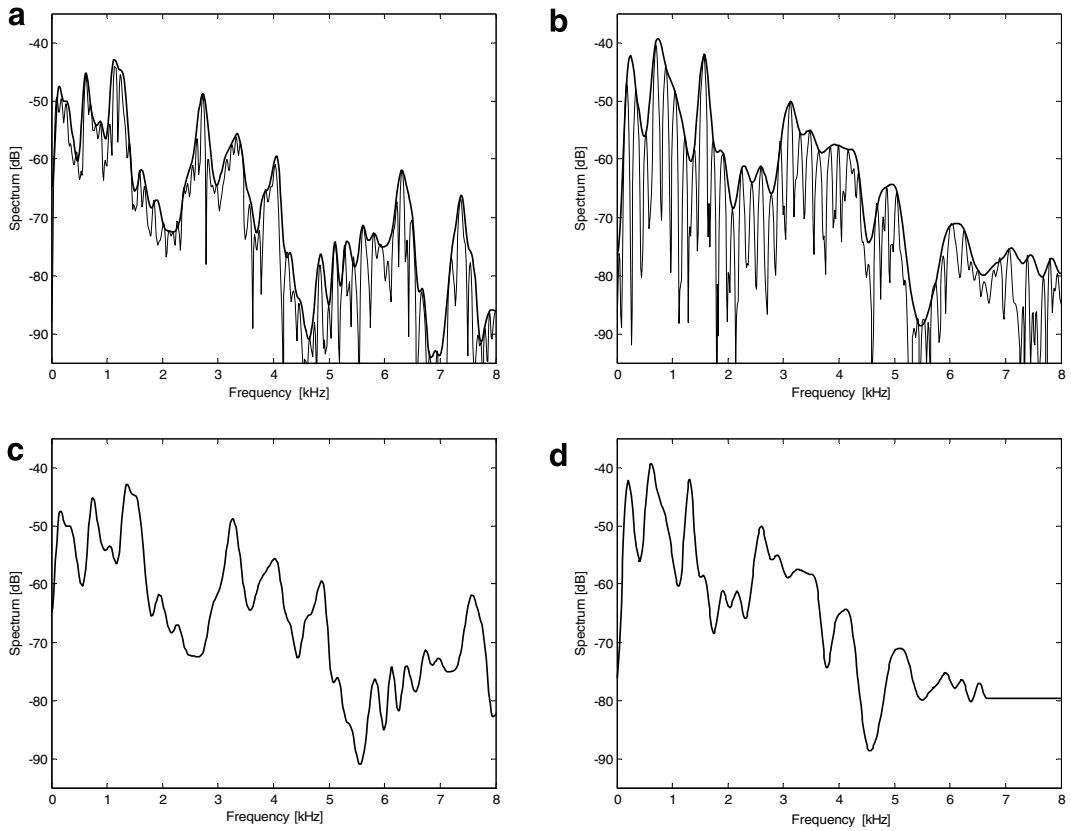

Fig. 2. Demonstration of linear frequency scale mapping: original male (a) and female (b) speech spectrum and its envelope, linear modification of male (c) and female (d) speech spectrum envelope (20% mean format shift).

points of the mapping function are $[0,0]$ and $[f_s/2,f_s/2]$ to preserve the number of points of the modified spectrum envelope while shifting formants.

4.1. First method

The first proposed non-linear method uses frequency scale mapping function as a circle arc going through the points $[0,0]$ and $[f_s/2,f_s/2]$. It is given by the equation

$$\eta(f) = y \pm \sqrt{r^2 - (f-x)^2}, \tag{4}$$

where

$$y = \frac{f_s}{8\Gamma} \cdot \frac{3\Gamma^2 - 1}{\Gamma - 1},$$
$$x = \frac{f_s}{2} - y, \tag{5}$$
$$r^2 = x^2 + y^2,$$

and Γ is the formant modification factor corresponding to the frequency $f_s/4$. The minus sign in the Eq. (4) corresponds to $\Gamma > 1$ (shifting formants to the right, see Figs. 3(a), 4(a) and (c)), the plus sign corresponds to $\Gamma < 1$ (shifting formants to the left, see Figs. 3(b), 4(b) and (d)). The modified log spectrum envelope is determined by (1) using the non-linear function of a circle arc (4). Let us return the linear function of (2), where the formant modification factor γ can be expressed by the constant ratio $f/\eta(f)$. Similarly in the non-linear frequency scale mapping, the formant modification factor as a function of frequency is given by the ratio $f/\eta(f)$. It is shown in Fig. 3(c) for $\Gamma > 1$, corresponding to shifting formants to the right, and in Fig. 3(d) for $\Gamma < 1$ corresponding to shifting formants to the left.

4.2. Second method

In the second proposed method the formant modification factor is constant in an interval which is considered as perceptually relevant. The formant modification factor $f/\eta(f)$ is equal to the constant γ in the interval $\langle f_s/2m, f_s/2n\rangle$, and it is given by

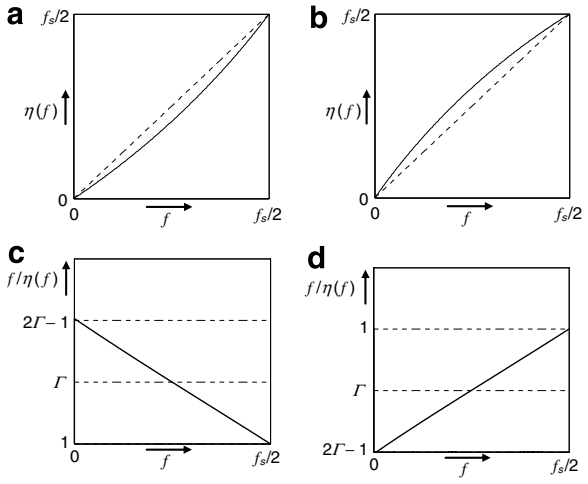

Fig. 3. Non-linear frequency scale mapping (first method): mapping function $\eta(f)$ as a circle arc for shifting formants to the right (a) and to the left (b), derived formant modification factor $f/\eta(f)$ for shifting formants to the right (c) and to the left (d).

parabolas in the marginal intervals $\langle 0, f_s/2m\rangle$, and $\langle f_s/2n, f_s/2\rangle$ (see Fig. 5(a) and (b)). Then, for the interval $\langle f_s/2m, f_s/2n\rangle$ it holds

$$\frac{f}{\eta(f)} = \gamma. \tag{6}$$

In the frequency range $\langle 0, f_s/2m\rangle$ the formant modification factor is represented by a parabola with the vertex $[f_s/2m, \gamma]$ and the point $[0, 1]$ lying on it. Its equation is

$$\frac{f}{\eta(f)} = a_1 f^2 + b_1 f + c_1, \tag{7}$$

where

$$a_1 = \frac{4m^2(1-\gamma)}{f_s^2},$$
$$b_1 = \frac{4m(\gamma-1)}{f_s}, \tag{8}$$
$$c_1 = 1.$$

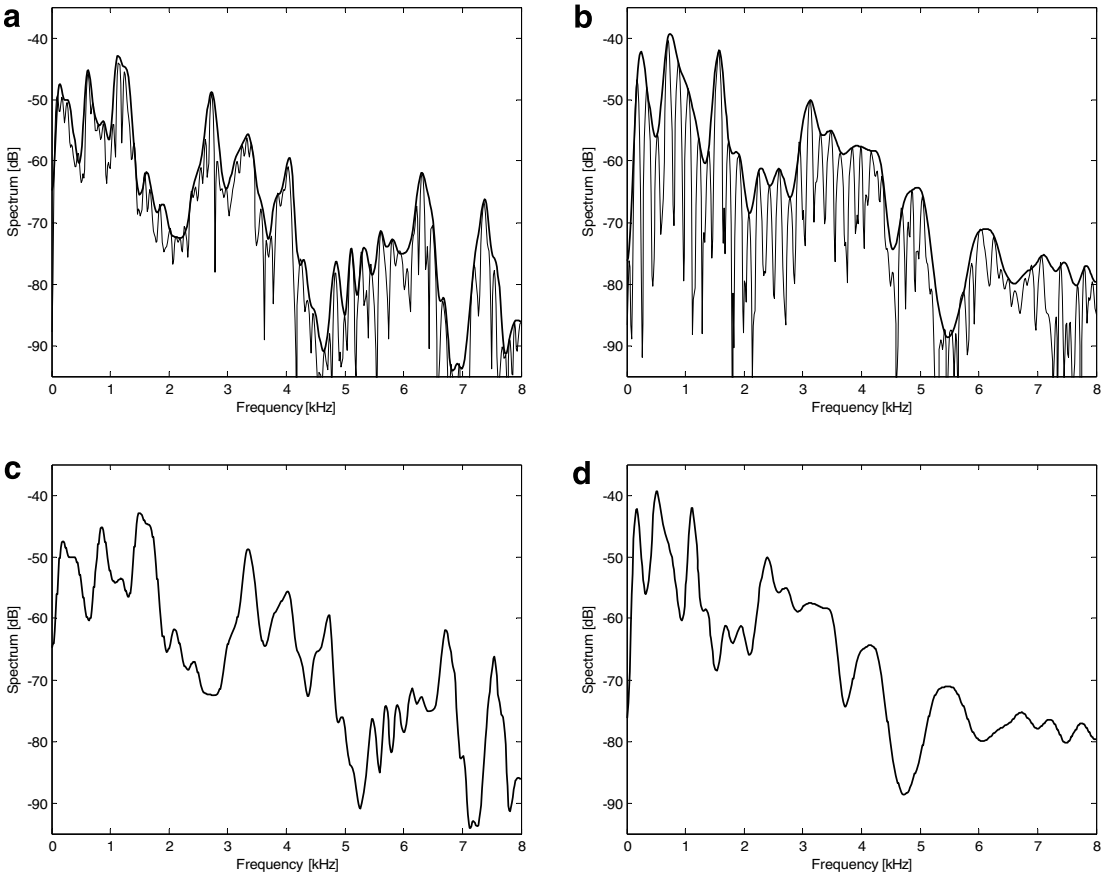

Fig. 4. Demonstration of non-linear frequency scale mapping (first method): original male (a) and female (b) speech spectrum and its envelope, non-linear modification of male (c) and female (d) speech spectrum envelope (20% mean format shift).

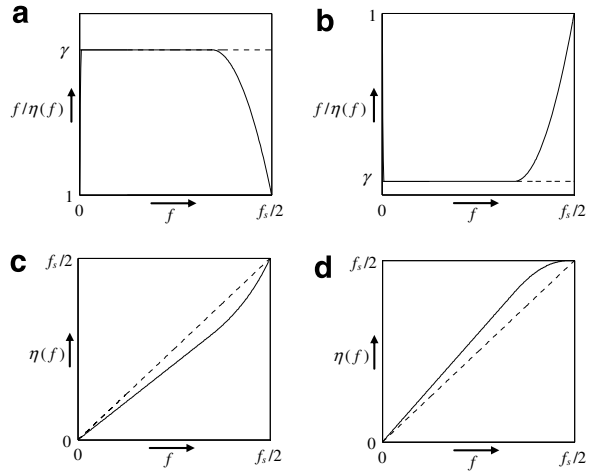

Fig. 5. Non-linear frequency scale mapping (second method): formant modification factor $f/\eta(f)$ constant in a given interval for shifting formants to the right (a) and to the left (b), derived frequency scale mapping function $\eta(f)$ for shifting formants to the right (c) and to the left (d).

In the frequency range $\langle f_s/2n, f_s/2\rangle$ it is represented by a parabola with the vertex $[f_s/2n, \gamma]$ and the point $[f_s/2, 1]$ lying on it. Its equation is

$$\frac{f}{\eta(f)} = a_2 f^2 + b_2 f + c_2, \tag{9}$$

where

$$a_2 = \frac{4n^2(1-\gamma)}{f_s^2(n-1)^2},$$
$$b_2 = \frac{4n(\gamma-1)}{f_s(n-1)^2}, \tag{10}$$
$$c_2 = \gamma + \frac{1-\gamma}{(n-1)^2}.$$

For $\gamma > 1$ both parabolas open downwards; for $\gamma < 1$ they open upwards. We have chosen the interval $\langle 80\text{ Hz}, 5.5\text{ kHz}\rangle$ as a perceptually relevant frequency interval with the constant formant modification factor. For the sampling frequency $f_s =$ 16 kHz it means $m = 100$ and $n = 16/11$.

The non-linear frequency scale mapping function $\eta(f)$ corresponding to the formant modification factor $f/\eta(f)$ in the frequency interval $\langle f_s/2m, f_s/2n\rangle$ is given by relation

$$\eta(f) = \frac{f}{\gamma}. \tag{11}$$

In the frequency range $\langle 0, f_s/2m\rangle$ its equation is

$$\eta(f) = \frac{f}{a_1 f^2 + b_1 f + c_1}, \tag{12}$$

and in the interval $\langle f_s/2n, f_s/2\rangle$ it is given by

$$\eta(f) = \frac{f}{a_2 f^2 + b_2 f + c_2}. \tag{13}$$

Fig. 5(c) shows the resulting curve $\eta(f)$ for $\gamma > 1$ corresponding to shifting formants to the right (compare Fig. 6(a) and (c)); Fig. 5(d) shows the curve for $\gamma < 1$ corresponding to shifting formants to the left (compare Fig. 6(b) and (d)).

For illustration of the described linear and non-linear transformation methods the original speech spectra and their envelopes are given in Figs. 2(a), (b), 4(a), (b) and 6(a), (b). Speech spectrum of a 24-ms frame of a vowel "A" sampled at 16 kHz spoken by a male and a female voice in the same phonetic context is used. For determination of the speech spectrum envelopes, the log spectrum is computed from the frame of pre-emphasized speech signal weighted by the Hamming window. This log spectrum is divided into the intervals of a pitch width. In each of the intervals, the local maxima are found. These local maxima located at pitch harmonics represent the samples of the spectrum to be interpolated (Přibilová and Vích, 2004). Using the inverse B-spline filtering (Unser, 1999), the B-spline coefficients are determined.

The resulting spectrum envelope is computed by spline interpolation as a convolution of the B-spline coefficients with B-splines of a third degree.

Then, the mean format modification factor of 1.2 is used for spectrum envelope transformation from male to female, and that of 1/1.2 for transformation from female to male as shown in Figs. 2(c), (d), 4(c), (d) and 6(c), (d).

As a result of used mapping methods Fig. 6(c) shows compression of the spectrum envelope near the half of the sampling frequency while in Fig. 2(c) the spectrum envelope near the half of the sampling frequency is lost as a result of linear expansion of the whole spectrum envelope. The resemblance of the mapping curves in Figs. 1(d) and 5(d) in the vicinity of half the sampling frequency results in similar transformed spectrum envelopes in Figs. 2(d) and 6(d). It can be seen that the position of the most important male formants of Figs. 2(a), 4(a) and 6(a) is closer to that of the transformed spectra in Figs. 2(d), 4(d) and 6(d) than to the original female ones in Figs. 2(b), 4(b) and 6(b). Similarly, the position of the most important female formants of Figs. 2(b), 4(b) and 6(b) is closer to that of the transformed spectra in Figs. 2(c), 3(c) and 4(c) than to the original male ones in Figs. 2(a), 4(a) and 6(a).

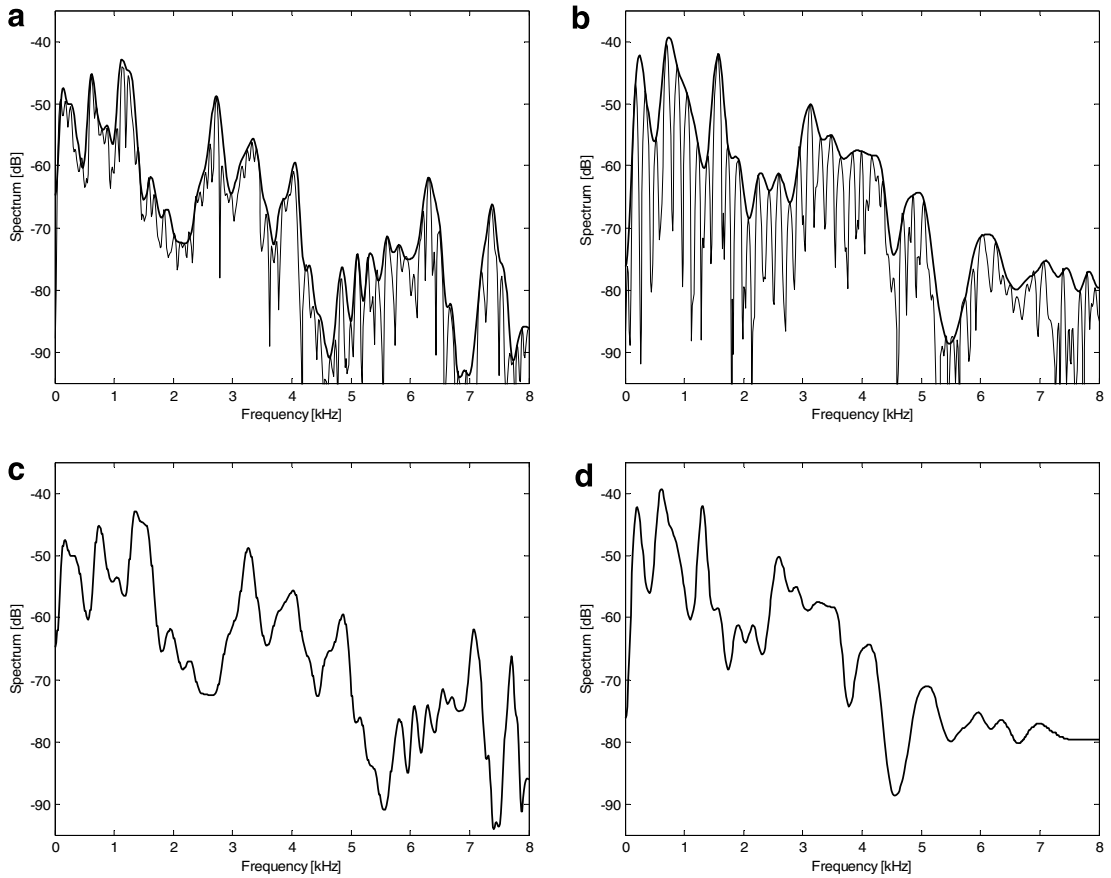

Fig. 6. Demonstration of non-linear frequency scale mapping (second method): original male (a) and female (b) speech spectrum and its envelope, non-linear modification of male (c) and female (d) speech spectrum envelope (20% mean format shift).

5. Cepstral speech parameters modification

Original cepstral description of a speech database can be transformed to a cepstral description of a new speaker using the non-linear modification methods described in Section 4. Block diagram of this transformation is shown in Fig. 7. The truncated cepstrum represents an approximation of a log spectrum envelope

$$E(f) = c_0 + 2\sum_{n=1}^{N_C} c_n \cos(n \cdot 2\pi f). \quad (14)$$

The first cepstral coefficient c_0 corresponding to the energy $\exp(c_0)$ is set to one, as the energy is supplied by the prosody generator of the TTS system. The first or the second non-linear frequency scale mapping method is used to get a modified log spectrum envelope $E'(f)$. It is further inverse Fourier transformed and truncated to get modified cepstral parameters. The procedure is performed for each speech frame of every speech unit in the speech inventory.

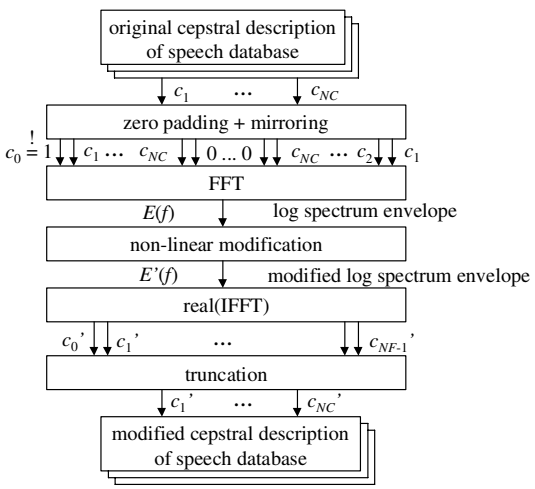

Fig. 7. Modified cepstral parameters computation.

6. Multi-voice TTS system

The multi-voice realization of the TTS system with cepstral description based on the source-filter speech model or the harmonic speech model with cepstral description is shown in Fig. 8. The text to be synthesized is entered as the input sequence of phonemes. It is converted to the combination of diphones and phone units through the prosody generator. Each of these units has its cepstral description in the speech database. For the real-time TTS synthesis based on the source-filter speech model (see Vích, 2000; Imai, 1978) the minimum-phase cepstral coefficients are used to model the vocal tract transfer function. For unvoiced speech, the synthesis filter is excited by the noise generator. For voiced speech, a superposition of the impulse generator output and the high-pass filtered noise according to the spectral flatness measure S_F is used. For the harmonic speech model (see Madlová, 2002; McAulay and Quatieri, 1995) the synthetic speech is generated as a sum of sine waves with given frequencies, amplitudes, and phases. The frequencies are the multiples of the pitch frequency F_0. The amplitudes and the phases are given by sampling the frequency response of the vocal tract model given by the cepstral coefficients. The spectral flatness measure is transformed to the maximum voiced frequency and the phases above this frequency are randomized. For unvoiced speech, the maximum voiced frequency is set to zero, so the phases are randomized in the whole bandwidth. For both speech models (source-filter, harmonic) the energy is supplied from the prosody generator. The user can modify the speech rate and the volume. According to the language (Czech, Slovak) prosody rules are applied. The resulting synthetic speech is generated pitch-synchronously.

Table 1
Pitch frequency and formant modification factor setting

Voice type	$F_{0\,\mathrm{basic}}$ (Hz)	min F_0 (Hz)	max F_0 (Hz)	Mean γ
Male	105	75	130	1
Young male	160	75	175	1.1
Female	230	175	275	1.2
Childish	300	275	430	1.35

We have used the original male voice database and three derived databases with given formant modification factors γ according to Table 1. Modified cepstral description of speech database for every new voice was computed off-line as described in Section 5 using both non-linear frequency scale mapping methods. The basic pitch frequency and the corresponding cepstral description of the speech database are selected and attached for synthesis by the choice of a voice. The basic pitch frequency values and possible pitch frequency ranges for each of the synthetic voices are also shown in Table 1.

7. Listening tests

Three types of listening tests have been performed for evaluation of synthetic male, and derived female, childish and young male voices:

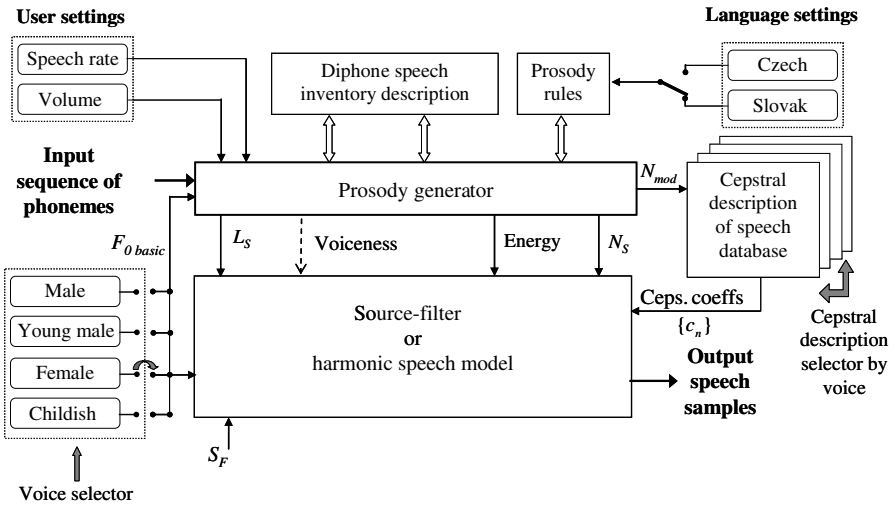

Fig. 8. Multi-voice realization of speech synthesis in the TTS system with cepstral description.

determination of voice type, naturalness of synthetic speech, and better sound of transformed voice. Twenty nine listeners within the age from 23 to 75 years (18 Czechs and 11 Slovaks, 17 men and 12 women) took part in the listening tests. The testing speech corpus consisted of 175 short utterances in the Czech language (with the average time duration of 3 s) synthesized with sampling frequency 16 kHz and the source filter speech model. Every utterance was performed with flat sentence prosody (a type of an unfinished sentence) applied in the block of the prosody generator (see Fig. 8). Every listening test consisted of a set of 10 audio samples selected randomly from the whole testing speech corpus. Each listener could hear the utterance for optional number of times before making an evaluation. The listening tests were performed at normal noise conditions (mainly in the office room), with the desktop speakers as the computer acoustic output.

In the first test, determination of voice type, the listeners chose the voice category from the five items: male, female, childish, young male, or not recognized (see results for male and female listeners in Fig. 9).

The second test, naturalness of synthetic speech, uses the mean opinion score (MOS) providing a numerical indication of the perceived quality with individual listeners' rating: 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. Fig. 10 shows the arithmetic mean of male and female listeners' individual scores.

In the third test, better sound of transformed voice, pairs of audio samples generated by two non-linear conversion methods were compared. Possible resulting answers were chosen from the cat-

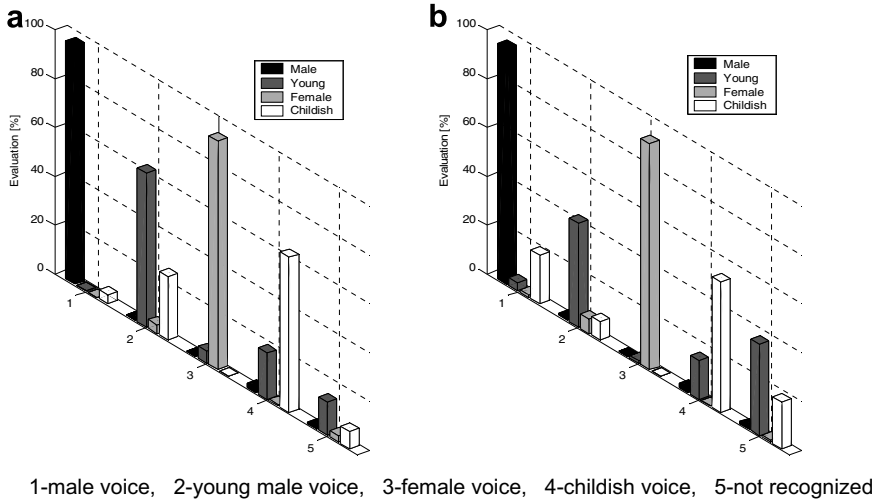

Fig. 9. Voice type determination by male listeners (a) and female listeners (b).

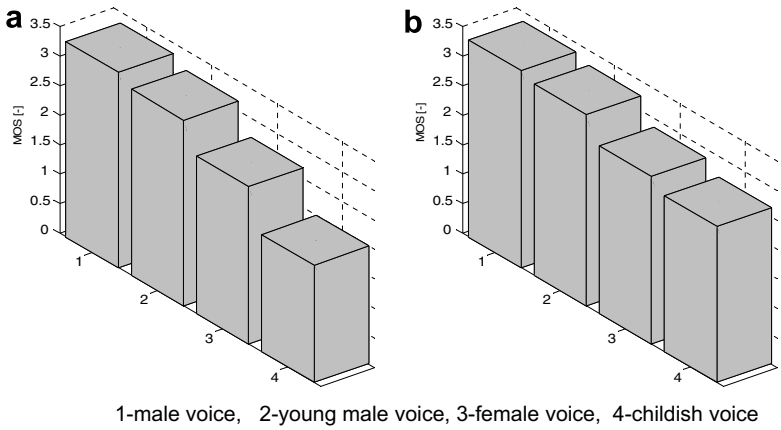

Fig. 10. MOS of synthetic speech naturalness according to male listeners (a) and female listeners (b).

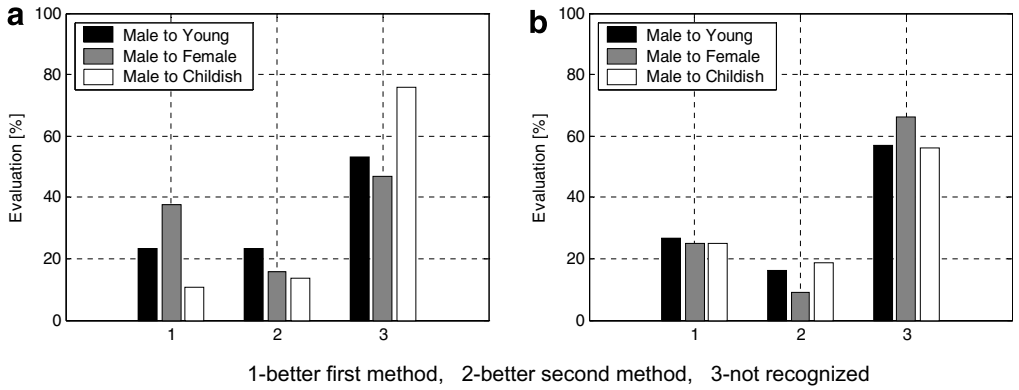

Fig. 11. Better sound of transformed voice according to male listeners (a) and female listeners (b).

egories: better first method, better second method, and not recognized (see results for male and female listeners in Fig. 11).

The developed testing program automatically generated the text protocol about the test. The output test values (correctness of voice determination, MOS, method of higher quality) were stored in separate files for final statistical post-processing and evaluation of the results in a graphical form. The listening tests results have shown a specific difference between evaluation of the male and female listeners. On the other hand, the age and the nationality of the listeners have not explicit influence on the evaluation. Summary results of all the listeners are presented in the numerical form in Tables 2–4.

Table 2
Summary results for voice type determination

Voice type	Accuracy (%)	False or not recognized (%)
Male	97.2	2.8
Young male	54.5	45.5
Female	93.1	6.9
Childish	59.3	40.7

Table 3
Summary results for speech naturalness determination

Voice type	Mean opinion score
Male	3.36
Young male	3.28
Female	2.71
Childish	2.28

Table 4
Summary results for better sound of transformed voice

Type of voice transformation	Better first method (%)	Better second method (%)	Not recognized (%)
Male to young male	23.8	20.4	55.7
Male to female	30.6	12.9	56.5
Male to childish	15.9	16.7	67.5

8. Conclusion

New non-linear methods of formant modification were introduced for voice conversion in the TTS system based on the diphone inventory of a male speaker. New synthetic voices were produced knowing mean formant and pitch shifts between male, female, and childish voices. Formant shifting was performed by frequency scale mapping of the speech spectrum envelope corresponding to original cepstral parameters and the spectrum envelope with shifted formants was transformed back into the modified cepstral parameters corresponding to a different voice.

In the first method, the frequency scale mapping function as a circle arc was proposed. The subsequently derived frequency dependent formant modification factor has gradually decreasing (increasing) character throughout the frequency bandwidth for voice conversion with increasing (decreasing) formants.

In the second proposed method, a constant formant modification factor in a perceptually relevant frequency interval was used. The interval $\langle 80\ \text{Hz}, 5.5\ \text{kHz} \rangle$ has been chosen, however, somewhat different marginal frequencies might have the similar perceptual effect. Although the quadratic

polynomials are used on the margins of this frequency interval, other smooth curves could be used as well. The subsequently derived frequency scale mapping function resembles more the linear mapping function than the circle arc.

Owing to proposed voice modification methods, only new cepstral description files must be generated and the original speech database is applied as one common area for all voices. This is very important for TTS system implementation in small mobile devices (like Pocket PC-s) which have problem with the size of the usable RAM memory. Next limitation follows from the fact that the used special main processor has lower computational performance having influence on a proper function of speech synthesis in real time. The second advantage of the described realization of the multi-voice TTS system consists in a fact that the derived cepstral descriptions for other voices are prepared in advance and synthesis can run with maximum speed as in the case of the original voice. The described modification of the segmental characteristics, the pitch and the formant locations and bandwidths, performed for voice conversion in the context of the Czech and Slovak TTS system, was implemented practically in the special aid for blind and partially sighted people – Braille notetaker based on the Pocket PC with the special Braille keyboard connected via wireless BlueTooth standard (Přibil and Přibilová, 2005).

The listening tests have shown that no audible differences between two non-linear voice conversion methods were perceived (see Table 4). While the synthetic original male and the derived female voices were mostly determined correctly, the detection of the derived young male and childish voices was little worse, as shown in Table 2. Even though the naturalness MOS values were rather dispersed, as they are dependent on individual listeners' opinions, the MOS evaluation decreases with increasing format position and fundamental frequency when compared with the original male voice parameters (see Tables 1 and 3).

Acknowledgements

The work had been done in the framework of the COST 277 Action "Non-linear Speech Processing". It has also been supported by the National Research Program "Targeted Research", Academy of Sciences of the Czech Republic, project number S108040569 and the Ministry of Education of the Slovak Republic (102/VTP/2000, 1/3107/06).

References

Arslan, L.M., 1999. Speaker transformation algorithm using segmental codebooks (STASC). Speech Commun. 28, 211–226.

Duběda, T., Horák, P., Vích, R., 2005. History of speech synthesis in the Czech lands. In: Vích, R. (Ed.), Electronic Speech Signal Processing, Proc. 16th Conf. Joined with the 15th Czech–German Workshop "Speech Processing", Prague, Czech Republic, pp. 364–371.

Duxans, H., Bonafonte, A., 2003. Estimation of GMM in voice conversion including unaligned data. In: Proc. European Conf. Speech Communication and Technology (EUROSPEECH'03), Geneva, Switzerland, pp. 861–864.

Fant, G., 1997. Acoustical analysis of speech. In: Crocker, M.J. (Ed.), Encyclopedia of Acoustics. John Wiley & Sons Inc., pp. 1589–1598.

Gutiérrez-Arriola, J.M., Montero, J.M., Vallejo, J.A., Córdoba, R., San-Segundo, R., Pardo, J.M., 2001. A new multi-speaker formant synthesizer that applies voice conversion techniques. In: Proc. European Conf. Speech Communication and Technology (EUROSPEECH'01), Aalborg, Denmark, pp. 357–360.

Imai, S., 1978. Low bit rate cepstral vocoder using the log magnitude approximation filter. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'78), Tulsa, OK, USA, pp. 441–444.

Iwahashi, N., Sagisaka, Y., 1995. Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks. Speech Commun. 16, 139–151.

Janota, P., 1967. Personal Characteristics of Speech. Academia, Praha.

Janota, P., Vích, R., 1994. Text-to-speech synthesis of Czech and Slovak. In: Jiříček, O. (Ed.), Proc. 31st Conf. Acoustics, Czech Acoustical Society, Prague, Czech Republic, pp. 139–142.

Janota, P., Dohalská, M., Palková, Z., Ptáček, M., 1994. Current situation in the research of automatic generation of the prosodic features with the diphone synthesis of spoken Czech. In: Lundin, F.J. (Ed.). Acta Universitatis Carolinae – Philologica, Phonetica Pragensia, pp. 33–58.

Kain, A., Macon, M.W., 1998. Spectral voice conversion for text-to-speech synthesis. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'98), Seattle, WA, USA, pp. 285–288.

Kuwabara, H., Sagisaka, Y., 1995. Acoustic characteristics of speaker individuality: control and conversion. Speech Commun. 16, 165–173.

Lee, K.-S., Doh, W., Youn, D.-H., 2002. Voice conversion using low dimensional vector mapping. IEICE Trans. Inf. Syst., 1297–1305.

Leutelt, L., Heute, U., 2004. Analysis of inter- and intra-speaker variations for voice conversion. In: Proc. Int. EURASIP Conf. Analysis of Biomedical Signals and Images (BIOSIGNAL'04), Brno, Czech Republic, pp. 30–32.

Madlová, A., 2002. Autoregressive and cepstral parametrization in harmonic speech modelling. J. Electr. Eng. 53, 46–49.

McAulay, R.J., Quatieri, T.F., 1995. Sinusoidal coding. In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, pp. 121–173.

Mizuno, H., Abe, M., 1995. Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. Speech Commun. 16, 153–164.

Mouchtaris, A., Van der Spiegel, J., Mueller, P., 2004. Non-parallel training for voice conversion by maximum likelihood constrained adaptation. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'04), Montreal, Canada, pp. I-1–I-4.

Narendranath, M., Murthy, H.A., Rajendran, S., Yegnanarayana, B., 1995. Transformation of formants for voice conversion using artificial neural networks. Speech Commun. 16, 207–216.

Přibil, J., Přibilová, A., 2005. Czech TTS engine for BraillePen device based on pocket PC platform. In: Vích, R. (Ed.), Electronic Speech Signal Processing, Proc. 16th Conference Joined with the 15th Czech–German Workshop "Speech Processing", Prague, Czech Republic, pp. 402–408.

Přibilová, A., Vích, R., 2004. Non-linear frequency scale mapping for voice conversion. In: Proc. 14th Int. Czech-Slovak Scientific Conf. Radioelektronika, Bratislava, Slovak Republic, pp. 100–103.

Rentzos, D., Vaseghi, S., Turajlic, E., Yan, Q., Ho, C.-H., 2003. Transformation of speaker characteristics for voice conversion. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU'03), St. Thomas, US Virgin Islands, pp. 706–711.

Slifka, J., Anderson, T.R., 1995. Speaker modification with LPC pole analysis. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'95), Detroit, MI, USA, pp. 644–647.

Stevens, K.N., 1997. Models of speech production. In: Crocker, M.J. (Ed.), Encyclopedia of Acoustics. John Wiley & Sons Inc., pp. 1565–1578.

Stylianou, Y., Cappé, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech Audio Process. 6, 131–142.

Sündermann, D., Bonafonte, A., Höge, H., Ney, H., 2004. Voice conversion using exclusively unaligned training data. In: Spanish Society for Natural Language Processing Conference, Barcelona, Spain, pp. 41–48.

Toda, T., Saruwatari, H., Shikano, K., 2001. Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'01), Salt Lake City, UT, USA, pp. 841–844.

Turk, O., Arslan, L.M., 2002. Subband based voice conversion. In: Proc. Int. Conf. Spoken Language Processing (ICSLP'02), Denver, CO, USA, pp. 289–292.

Unser, M., 1999. Splines. A perfect fit for signal and image processing. IEEE Signal Process. Magazine 16, 22–38.

Vích, R., 2000. Cepstral speech model, Padé approximation, excitation, and gain matching in cepstral speech synthesis. In: Proc. Int. EURASIP Conf. Analysis of Biomedical Signals and Images (BIOSIGNAL'00), Brno, Czech Republic, pp. 77–82.

Vích, R., Horák, P., 1996. Text-to-speech conversion, history and present state, Invited paper. In: Proc. 6th National Scientific Conf. with Int. Participation Radioelektronika 96, Brno, Czech Republic, pp. 1–7.

Vondra, M., Vích, R., 2005a. Speech identity conversion. In: Chollet, G., Esposito, A., Faundez-Zanuy, M., Marinaro, M. (Eds.), Nonlinear Speech Modeling and Applications, Advanced Lectures and Revised Selected Papers 3445. Springer, pp. 421–426.

Vondra, M., Vích, R., 2005b. Sound-dependent spectral warping in voice identity conversion. In: Vích, R. (Ed.), Electronic Speech Signal Processing, Proc. 16th Conf. Joined with the 15th Czech–German Workshop "Speech Processing", Prague, Czech Republic, pp. 423–429.

Ye, H., Young, S., 2003. Perceptually weighted linear transformations for voice conversion. In: Proc. European Conf. Speech Communication and Technology (EUROSPEECH'03), Geneva, Switzerland, pp. 2409–2412.