Unsupervised Speech Morphing between Utterances of any Speakers

Hartmut R. Pfitzinger

Institute of Phonetics and Speech Communication
University of Munich, Germany
hpt@phonetik.uni-muenchen.de

Abstract

A new approach to speech morphing is presented which avoids the extraction of fundamental and formant frequencies as well as the detection of phone or syllable boundaries. All prominent spectral and temporal features of the source and target utterances are automatically related and interpolated. The method consists of three main parts: LPC-based source-filter decomposition, separate interpolation, and composition of the morphed speech signal. The paper focuses on the alignment and interpolation problems on three speech signal layers: the timing structure on a phone- and syllable-level, the shape of the frequency spectrum including formants and other spectral properties, and the micro-timing of the source signal. Particularly, the source signal alignment and interpolation is described since it is most crucial for the resulting quality of the modified speech signal. The new morphing procedure was applied to utterances taken from the freely available CMU_ARCTIC speech corpus and assessed by a perceptual MOS experiment. Preliminary results indicated an excellent acoustic quality of the morphed speech signals.

1. Introduction

When in the 1941 movie version of the famous story "Dr. Jekyll and Mr. Hyde" by R.L. Stevenson the first visual metamorphosis effect was instantiated, it attracted great attention and inspired a number of film-makers to apply and improve this effect. But it took 47 years, until the 1988 movie "Willow", to apply the first fully digital visual morphing effect. And until now acoustical morphing and especially speech morphing is still far from being a standard application.

Before describing our new high-quality speech morphing method we should clarify its usefulness in speech science: Important findings are based on investigating human interaction with unreal stimuli (see e.g. McGurk and MacDonald 1976). Speech morphing techniques are well suited to create those unreal speech stimuli which might be perceived as real.

Another application of speech morphing in stimulus generation is to repeat well-established experiments concerning categorical perception: instead of e.g. creating formant transitions for the well-known /ba/-/da/-/ga/-experiments by hand (Liberman, Delattre, and Cooper 1952), it would be interesting to use speech morphing between naturally spoken syllables, thus retaining all unknown cues and interpolating not only those cues known to be crucial.

Finally, in concatenative speech synthesis, where each desired voice has to be recorded and processed before being available for synthesizing new utterances, morphing between a small number of male and female target speakers defining the corners of a target 'voice space' may be sufficient to create a large number of new, natural, and unfamiliar voices (Kain and Macon 1998).

Visual morphing algorithms decompose pictures into edges and textures, which then are interpolated separately before being reconstructed. According to our understanding of speech morphing, visual edges correspond to the evident formant structure and segmental timing structure of the time-frequency envelope of speech while visual textures correspond to the excitation signal with its fundamental frequency and its spectral properties.

1.1. Speech Morphing vs. Voice Conversion

Speech morphing and voice conversion are obviously different: In the former case the source and target signals should be sufficiently similar to become reasonably aligned and interpolated for achieving new signals. In the latter case a source-target relationship is learned from a number of (not necessarily similar) utterances from two speakers, which is then used to convert unseen signals of the source speaker towards the target speaker.

Glottal source conversion (Mokhtari, Pfitzinger, and Ishi 2003; Mokhtari, Pfitzinger, Ishi, and Campbell 2004) or vocal tract warping alone are neither speech morphing nor voice conversion since modifying only one component retains the characteristics of the source speaker in the other component of the source-filter model.

1.2. Two schools of speech signal manipulation

Speech signals can be modified in numerous ways (Moulines and Laroche 1995) but if both the vocal tract and the glottal properties have to be changed at the same time, two very different strategies are commonly used:

1. Pitch-Synchronous Fourier Spectrum Modification
Fourier Transformation (DFT/FFT) yields a time-frequency representation of the speech signal which can be modified and re-converted into a speech signal. A constant analysis frame size (Moulines and Laroche 1995) or a pitch-synchronous window technique is used which allows for efficiently changing F0 (Simon 1983; Tillmann, Schiefer, and Pompino-Marschall 1984; Abe 1996; Kawahara and Matsui 2003). However, due to unavoidable errors introduced by Fourier-spectrum modification, the resynthesized speech often suffers from audible artifacts.

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

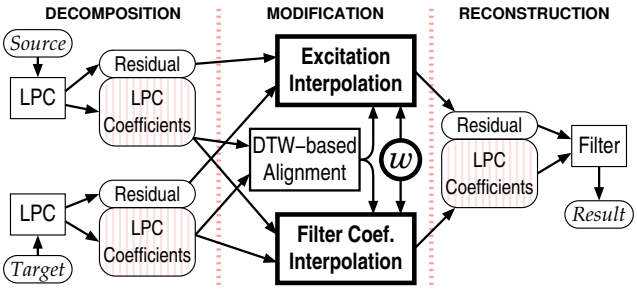

Figure 1: Block diagram of the proposed morphing technique. The only user-adjustable parameter ω determines the ratio between the source and target signal.

2. Source-Filter Decomposition

Despite the fact that the glottis and the vocal tract are interacting (Childers and Wong 1994), a decomposition via Linear Predictive Coding (LPC) into a quasi-excitation signal and quasi-stationary filter coefficients enables meaningful modifications. Chappell and Hansen (2002, p. 352 and Fig. 4) recommend interpolation of the residual signal and the Line Spectrum Pair (LSP) representation separately. Many recent approaches to speech morphing and voice conversion still use source-filter decomposition (Valbret, Moulines, and Tubach 1992; Mizuno and Abe 1995; Ribeiro and Trancoso 1996; Rinscheid 1996; Orphanidou, Moroz, and Roberts 2003).

2. Method

The proposed speech morphing procedure is outlined in Fig. 1. While the overall concept is straightforward, two components are crucial for the resulting speech signal quality, namely the vocal tract interpolation (section 2.3.) and the excitation signal interpolation (section 2.4.).

2.1. Source-Filter decomposition

The speech signals of the source and the target speaker are sampled at 16 kHz, pre-emphasized, and LPC-transformed to autoregressive filter coefficient polynomials of order 18 in steps of 10ms at a window size of 20ms. Then the amplitude-normalized quasi-excitation signals are estimated by inverse filtering (see left part of Fig. 1).

2.2. Time-alignment of two utterances

Since the durations of syllables and phones usually vary greatly among speakers, even in repeated utterances of the same sentence, the first step of the modification process (see middle part of Fig. 1) consists in aligning corresponding frames of both utterances.

This is done by standard Dynamic Time Warping (DTW) using the local path constraints shown in the right panel of Fig. 4. The autoregressive filter coefficients are transformed into statistical Z-scores of 32 frequency bands each with a bandwidth of 1 Bark and equally-spaced on a Bark scale between 200 Hz and 8 kHz. Preliminary observations showed that this spectral representation has reasonable inter-speaker alignment properties when using Euclidean distance measure.

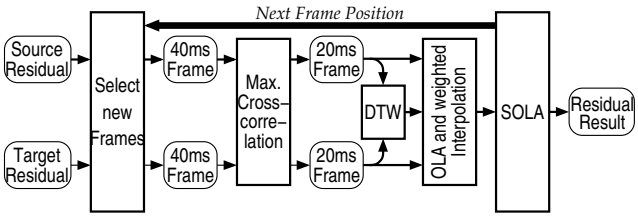

Figure 2: Detailed block diagram of the excitation interpolation method (see section 2.4.).

2.3. Interpolation of two frequency spectra

In (Pfitzinger 2004) we introduced a new DFW-based spectral smoothing technique which is useful not only for reducing discontinuities in Concatenative Speech Synthesis but also for the interpolation of corresponding frames taken from two different speakers.

It uses Dynamic Programming to find the best alignment of the derivatives of two amplitude-normalized Fourier spectra of LP-polynomials. This spectral representation shows properties well-suited for DP alignment since its overall slope is flat and each pole is characterized by a zero-crossing as well as a large local negative slope. A search space restriction to $\pm$500 Hz along the frequency axis is sufficient for coping with male-female formant frequency differences. The morphed spectrum is achieved by weighted linear interpolation between the aligned spectral lines of the source and the target frames. Finally, the spectrum is transformed back to autoregressive filter coefficient polynomials.

Line spectrum pair (LSP) interpolation, which is known to yield reasonable results (Chappell and Hansen 2002), suffers from increased formant bandwidths and reduced formant amplitudes in an interpolation range of 30–70% between the source and the target spectrum.

Our spectral interpolation method is computationally more expensive than LSP-interpolation but it aligns all formants with small bandwidths perfectly, while poles with large bandwidths are moved reasonably to support the overall spectral shape. Thus it creates interpolated spectra not by merely cross-fading which is inappropriate in a speech morphing task, but by more correctly shifting the acoustic resonance characteristics.

2.4. Interpolation of two excitation signals

As mentioned above, the quality of any speech morphing technique strongly depends on the amount of distortion, aliasing, and other artifacts introduced by the particular excitation signal interpolation. Fig. 2 outlines the main stages of our new interpolation procedure which are described in the following sections.

2.4.1. Pre-processing of the residual signals

Both LPC-derived residual signals (Fig. 1) are FIR-high-pass filtered at 100 Hz (-12dB/oct) followed by integration to compensate for lip radiation and thereby obtain quasi-excitation signals. We use the term ‘quasi’ to express that these signals are not necessarily identical to the real excitation signals.

Accepted after full paper review

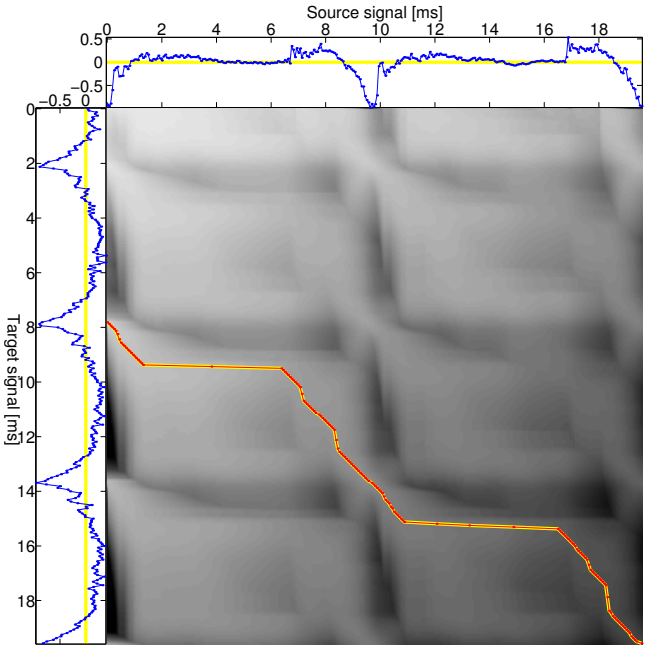

Figure 3: Alignment of 20ms-stretches of amplitude-normalized excitation signals from the male source and the female target.

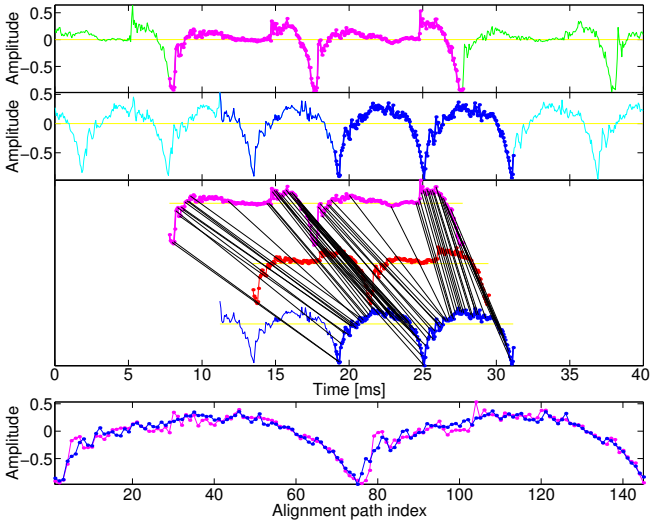

Figure 5: *1st and 2nd panel:* 20ms-frames (*dark colors*) automatically selected for DP-alignment from 40ms-stretches of excitation signals. *3rd panel:* Aligned parts of both 20ms-frames, aligned samples (indicated by lines joining aligned samples), and an interpolated excitation signal at a 50%-ratio. *4th panel:* Corresponding excitation samples plotted according to the lines joining aligned samples.

2.4.2. Frame selection

Initially, the very first 40ms of both excitation signals are selected for further processing. Then, feedback information from the final SOLA-based frame-concatenation stage (section 2.4.6.) together with the DTW-based alignment-path (section 2.2.) determines the center of the next 40ms-frames.

2.4.3. Endpoint alignment by cross-correlation

To guarantee a good match of the endpoints of two frames, two-dimensional short-term cross-correlations (0.4ms window size≈7 samples) are estimated over 20ms-stretches of the 40ms-frames. These are right-justified and Bartlett-windowed to favour an endpoint near to the 30ms positions and thus yield centered 20ms-frames (Fig. 5).

2.4.4. Dynamic Time Warping of the micro-structure

Fig. 4 shows two local path constraint methods we compared to align excitation signals. The method with seven predecessors restricted the slope of the alignment path to a ratio of 4:1 or 1:4 which corresponds to a maximum f0 difference of two octaves and thus should account for most situations where two subjects produced the same sentence. Furthermore this method favours compression or expansion of a glottal cycle rather than skipping part of the cycle. But informal perception results showed that skipping fractions of cycles (see Fig. 3) is an advantageous property of the method with only three predecessors, reducing the amount of jitter introduced by source signal modification.

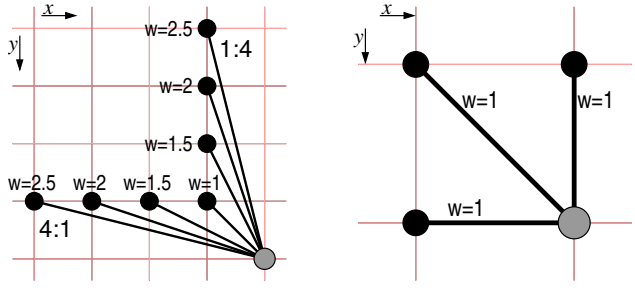

Figure 4: Two different local path constraints tested in the DP alignment. *Left:* The alignment slope is restricted between 4:1 and 1:4. *Right:* No slope restrictions but high preference for diagonal alignment because all weights are set to 1.

2.4.5. Overlap-and-Add and weighted interpolation

An OLA (*overlap and add*) method applied to very short stretches (so called *granulars*) of both signals according to the micro-alignment path (provided by the preceding stage) followed by weighted interpolation yielded the morphed excitation frames (see third panel of Fig. 5). As the fourth panel of Fig. 5 shows, the stretches of both excitation signals are reasonably aligned.

It is noteworthy that while the closing phase is changed extremely, the steepness of the negative slope remains untouched. The voice quality which is remarkably sensitive to this slope seems to be comparable between both speakers. Informal listening tests confirmed that both speakers used a modal voice.

2.4.6. Synchronous Overlap-and-Add (SOLA)

The last 7ms of the already finished new excitation signal are compared with the new frame via AMDF (*average magnitude difference function*) to find the optimal concatenation point. Our AMDF halves marginal distances thus favouring the minimization of window-central mismatches.

The previous and the new signal are then concatenated using a linear cross-fading over 2ms to avoid any click artifact. Usually, the preceding processing steps yield frames which fit together quite well. Finally, the position of the

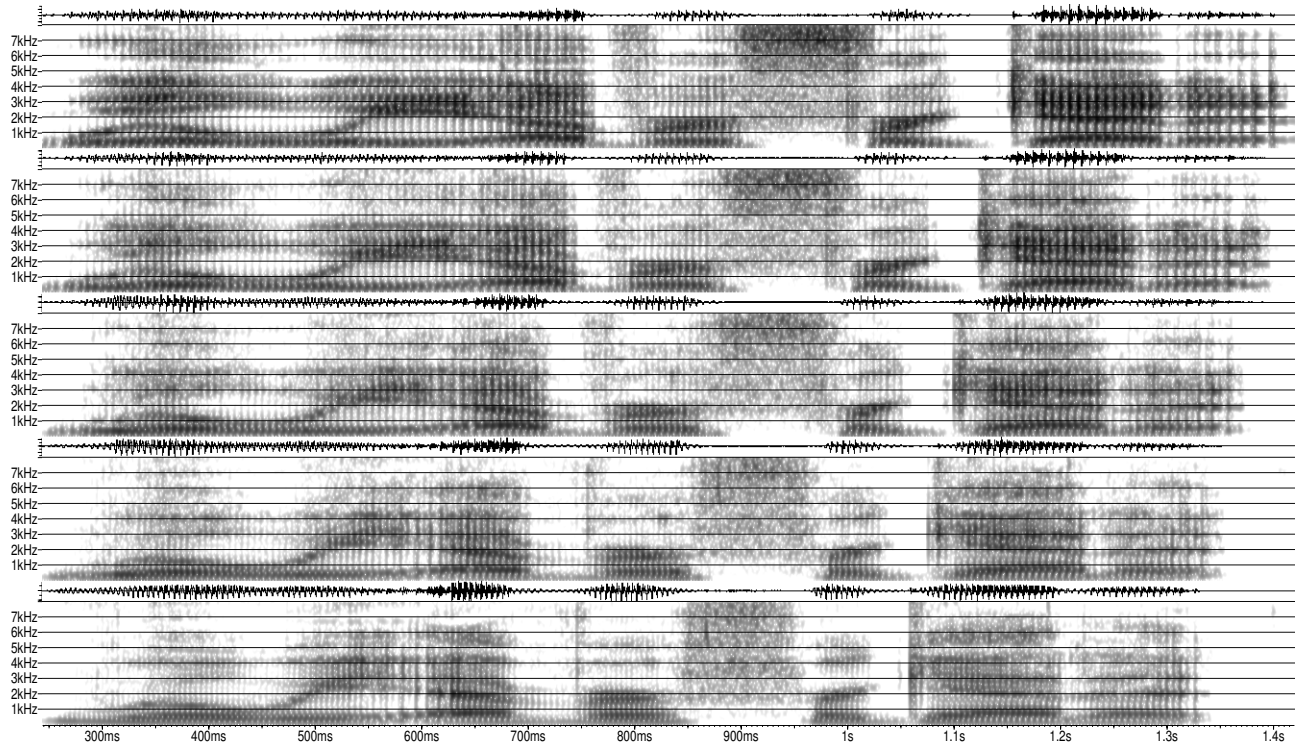

Figure 6: Oscillograms and sonagrams of the utterance *a0005* (*"Will we ever forget it."*) taken from the CMU_ARCTIC speech corpus. *Top panel:* Male speaker *bdl*. *Bottom panel:* Female speaker *slt*. *2nd–4th panel:* 75%, 50%, and 25% morphing between the speakers.

new last sample guides the selection of the next 40ms-frames (see section 2.4.2.) until the whole residual signals are processed.

2.5. Source-Filter reconstruction

Finally, the separately interpolated excitation signal and vocal tract filter polynomials are composed via autoregressive filtering and then de-emphasized to obtain the morphed speech signal (see right part of Fig. 1).

3. Experiment

Our new morphing method was applied to utterances spoken by two speakers taken from the freely available CMU_ARCTIC speech corpus.[1] The signal shown in Fig. 6 is available on the web[2] and we strongly encourage everybody to gain an impression of the achieved sound quality by listening to the example.

It is remarkable that even the crossing of the second and third formant which can be observed in the bottom panel of Fig. 6 at ca. 500ms is evenly morphed to the more ambiguous formant complex of the male speaker in the top panel.

3.1. Preliminary Quality Assessment

To test the acoustic quality of our new speech morphing method we conducted a Mean Opinion Score (MOS) experiment where ten subjects were asked to assess the playback quality of randomly presented stimulus utterances on a scale from 1 (=*excellent*) to 6 (=*poor*). These utterances consisted of five sentences taken from the CMU_ARCTIC speech corpus which were spoken by a male (*bdl*) and a female speaker (*slt*). Additionally, for each of these pairs 11 intermediate utterances were estimated using our speech morphing method. Each utterance was presented twice giving a total of 130 stimuli ($5 \cdot (1 + 1 + 11) \cdot 2 = 130$).

3.2. Results

Fig. 7 shows the results of this experiment. Each errorbar is based on 100 judgements (10 subjects times 5 utterances times 2 repetitions). An ANOVA revealed none of the independent variables to be significant. Since only ten subjects took part in the experiment we regard these results as preliminary. But they surely indicate that the acoustic quality of the morphed stimuli 2–12 is only marginally reduced compared with the original stimuli 1 and 13.

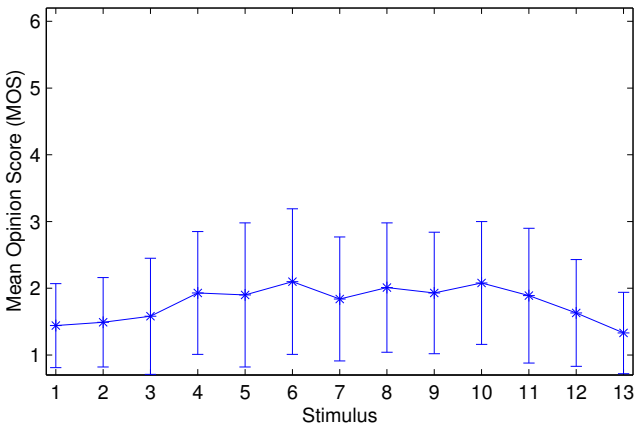

Figure 7: Mean Opinion Score (MOS) results of the acoustic quality of the stimuli (1=*excellent*, 6=*poor*).

[1] http://festvox.org/cmu_arctic/

[2] http://www.phonetik.uni-muenchen.de/∼hpt/morphing/

4. Discussion

The proposed morphing method extensively uses Dynamic Programming techniques to find the alignments of the speech signals on three different layers: *i)* the timing structure on a phone- and syllable-level, *ii)* the shape of the frequency spectrum including formants and other spectral properties, and *iii)* the micro-timing of the source signal, where the latter mostly influences the resulting speech signal quality. A further decomposition of the excitation signal into harmonic and noise components (Childers 1995; Yegnanarayana, d'Alessandro, and Darsinos 1998) is currently under investigation.

In our view, full voice conversion techniques are still unusable in any demanding application. The problem to be solved consists not only in finding the best alignments and artifact-free time-scale and frequency-scale inter- or extrapolation methods. It consists mainly in modeling the speaker's behaviour to express himself. His use of fundamental frequency, of momentarily changing voice qualities, of the reduction or insertion of phones and even syllables, of allophonic variants, etc., is not only determined by the phones, syllables or words he produces.

The context which determines the speaker's specific choice is wider than only the adjacent constituents: it includes the speaker's dialect, sociolect, mood, and intention. Supposing that listeners would accept the syntactic, semantic, and pragmatic structure of the utterance (even though they have a very good idea about which words a known speaker would probably say or not), it is a very easy task for them to recognize inconsistencies in the resulting prosodic and dialectal properties of the target utterance. In fact, listeners unconsciously and continuously analyse and collect extralinguistic information about the speaker which is conveyed by the speech signal. They can perfectly detect inconsistencies about the speaker. Therefore, the poor speech signal quality of current voice conversion techniques is generally far from being considered acceptable.

In contrast, speech morphing is a much simpler task because both the source as well as the target utterances are given. All speech segments and features which must be present in the morphed speech signal are already present in at least one of the two utterances. Thus, morphing methods don't have to *generate any new* excitation signal, phone, or syllable. Naturally, if the number of syllables or the syllable structure differs, the alignment procedure is not guaranteed to produce a realistic result (e.g. [a:s] vs. [sa:] where an appropriate excitation signal morphing is simply impossible). But our morphing procedure is flexible enough to also yield reasonable results in the case of two linguistically different utterances (e.g. *keen* [ki:n] vs. *clean* [kli:n] or *morning* and *corner* (Slaney, Covell, and Lassiter 1996)) and thus is — as originally intended — a practical tool for creating new stimuli for categorical speech perception experiments.

5. Conclusions

A useful solution to the problem of speech morphing was presented. The resulting speech signal quality is extremely high and shows only very few artifacts. Our solution might not be optimal in a computational sense since it is considerably slower than other methods. But this is mainly caused by the higher degrees of freedom which our method provides and which we claim to be necessary for a convincing solution to the problem of unsupervised morphing between any two utterances.

6. Acknowledgements

I would like to thank Denis Burnham for encouraging me to work again on speech morphing, Nick Campbell for his patience with me (since I finished some central algorithms during my stay at his lab), Parham Mokhtari and Carlos Toshinori Ishi for countless speech signal discussions, and JST/CREST and BMW Group Research and Technology Pty Ltd, Munich for their financial support.

References

Abe, M. (1996). Speech morphing by gradually changing spectrum parameter and fundamental frequency. In *Proc. of ICSLP '96*, Volume 4, Philadelphia, pp. 2235–2238.

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

Arslan, L. M. and D. Talkin (1997). Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum. In *Proc. of EUROSPEECH '97*, Volume 3, Rhodes; Greece, pp. 1347–1350.

Chappell, D. T. and J. H. L. Hansen (2002). A comparison of spectral smoothing methods for segment concatenation based speech synthesis. *Speech Communication 36*, 343–374.

Childers, D. G. (1995). Glottal source modeling for voice conversion. *Speech Communication 16*(2), 127–138.

Childers, D. G. and C.-F. Wong (1994). Measuring and modeling vocal source-tract interaction. *IEEE Transanctions on Biometrical Engineering 41*(7), 663–671.

Erkelens, J. S. and P. M. T. Broersen (1998). LPC interpolation by approximation of the sample autocorrelation function. *IEEE Transactions on Speech and Audio Processing 6*(6), 569–573.

Gillett, B. (2003). Transforming voice quality and intonation. Master's thesis, Centre for Speech Technology Research (CSTR), University of Edinburgh.

Goncharoff, V. and M. Kaine-Krolak (1995). Interpolation of LPC spectra via pole shifting. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP95)*, Volume 1, Detroit, pp. 780–783.

Iwahashi, N. and Y. Sagisaka (1994). Speech spectrum transformation by speaker interpolation. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP94)*, Volume 1, Adelaide, pp. 461–464.

Iwahashi, N. and Y. Sagisaka (1995). Speech spectrum conversion based on speaker interpolation and

Accepted after full paper review

multi-functional representation with weighting by radial basis function networks. *Speech Communication 16*(2), 139–151.

Kain, A. and M. W. Macon (1998). Spectral voice conversion for text-to-speech synthesis. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP98)*, Volume 1, Seattle, pp. 285–288.

Kawahara, H. and H. Matsui (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP03)*, Volume 1, Hong Kong, pp. 256–259.

Liberman, A. M., P. C. Delattre, and F. S. Cooper (1952). The role of selected stimulus-variables in the perception of the unvoiced stop consonants. *American Journal of Psychology 65*, 497–516.

Mareüil, P. B. d., P. Célérier, and J. Toen (2002). Generation of emotions by a morphing technique in English, French and Spanish. In *Proc. of the 1st Int. Conf. on Speech Prosody*, Aix en Provence, pp. 187–190.

McGurk, H. and J. MacDonald (1976). Hearing lips and seeing voices. *Nature 264*, 746–748.

Mizuno, H. and M. Abe (1995). Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt. *Speech Communication 16*(2), 153–164.

Mokhtari, P., H. R. Pfitzinger, and C. T. Ishi (2003). Principal components of glottal waveforms: Towards parameterisation and manipulation of laryngeal voice-quality. In *Proc. of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (Voqual'03)*, Geneva, pp. 133–138.

Mokhtari, P., H. R. Pfitzinger, C. T. Ishi, and N. Campbell (2004). Laryngeal voice quality conversion by glottal waveshape PCA. In *Proc. of the Spring 2004 Meeting of the Acoustical Society of Japan*, Atsugi, pp. 341–342.

Mori, H. and H. Kasuya (2003). Speaker conversion in ARX-based source-formant type speech synthesis. In *Proc. of EUROSPEECH '03*, Volume 4, Geneva, pp. 2421–2424.

Moulines, E. and J. Laroche (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication 16*(2), 175–205.

Narendranath, M., H. A. Murthy, S. Rajendran, and B. Yegnanarayana (1995). Transformation of formants for voice conversion using artificial neural networks. *Speech Communication 16*(2), 207–216.

Orphanidou, C., I. M. Moroz, and S. J. Roberts (2003). Voice morphing using the generative topographic mapping. In *Proc. of the Int. Conf. on Computer, Communication and Control Technologies*, Volume 1, pp. 222–225.

Paliwal, K. K. (1995). Interpolation properties of linear prediction parametric representations. In *Proc. of EUROSPEECH '95*, Volume 2, Madrid, pp. 1029–1032.

Pfitzinger, H. R. (2004). DFW-based spectral smoothing for concatenative speech synthesis. In *Proc. of ICSLP '04*, Volume 2, Korea, pp. 1397–1400.

Rentzos, D., S. Vaseghi, Q. Yan, C.-H. Ho, and E. Turajlic (2003). Probability models of formant parameters for voice conversion. In *Proc. of EUROSPEECH '03*, Volume 4, Geneva, pp. 2405–2408.

Ribeiro, C. M. and I. M. Trancoso (1996). Application of speaker modification techniques to phonetic vocoding. In *Proc. of ICSLP '96*, Volume 1, Philadelphia, pp. 306–309.

Rinscheid, A. (1996). Voice conversion based on topological feature maps and time-variant filtering. In *Proc. of ICSLP '96*, Volume 3, Philadelphia, pp. 1445–1448.

Simon, T. (1983). Manipulation of natural speech signals according to the speech parameters of different speakers. Forschungsberichte (FIPKM) 17, Institut für Phonetik und Sprachliche Kommunikation der Universität München.

Slaney, M., M. Covell, and B. Lassiter (1996). Automatic audio morphing. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP96)*, Volume 2, Atlanta, pp. 1001–1004.

Tillmann, H. G., L. Schiefer, and B. Pompino-Marschall (1984). Categorical perception of speaker identity. In M. P. R. v. d. Broecke and A. Cohen (Eds.), *Proc. of the Xth Int. Congress of Phonetic Sciences (Utrecht 1983)*, Volume IIB, Dordrecht, pp. 443–448.

Turk, O. and L. M. Arslan (2002). Subband based voice conversion. In *Proc. of ICSLP '02*, Volume 1, Denver, pp. 289–292.

Valbret, H., E. Moulines, and J. P. Tubach (1992). Voice transformation using PSOLA technique. *Speech Communication 11*(2–3), 175–187.

Verhelst, W. and J. Mertens (1996). Voice conversion using partitions of spectral feature space. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP96)*, Volume 1, Atlanta, pp. 365–368.

Ye, H. and S. Young (2003). Perceptually weighted linear transformations for voice conversion. In *Proc. of EUROSPEECH '03*, Volume 4, Geneva, pp. 2409–2412.

Ye, H. and S. Young (2004). High quality voice morphing. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP04)*, Montreal, Canada.

Yegnanarayana, B., C. d'Alessandro, and V. Darsinos (1998). An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transactions on Speech and Audio Processing 6*(1), 1–11.

Accepted after full paper review