EURASIP Journal on Applied Signal Processing 2005:8, 1174–1184


# Voice Morphing using 3D Waveform Interpolation Surfaces and Lossless Tube Area Functions

**Yizhar Lavner**

*Department of Computer Science, Tel-Hai Academic College, Upper Galilee 12210, Israel*
*Email: yizhar_l@kyiftah.org.il*

*Signal and Image Processing Lab (SIPL), Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel*

**Gidon Porat**

*Signal and Image Processing Lab (SIPL), Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel*
*Email: gidon.porat@intel.com*

*Received 31 December 2003; Revised 2 February 2005; Recommended for Publication by Mark Kahrs*

Voice morphing is the process of producing intermediate or hybrid voices between the utterances of two speakers. It can also be defined as the process of gradually transforming the voice of one speaker to that of another. The ability to change the speaker's individual characteristics and to produce high-quality voices can be used in many applications. Examples include multimedia and video entertainment, as well as enrichment of speech databases in text-to-speech systems. In this study we present a new technique which enables production of a given number of intermediate voices or of utterances which gradually change from one voice to another. This technique is based on two components: (1) creation of a 3D prototype waveform interpolation (PWI) surface from the LPC residual signal, to produce an intermediate excitation signal; (2) a representation of the vocal tract by a lossless tube area function, and an interpolation of the parameters of the two speakers. The resulting synthesized signal sounds like a natural voice lying between the two original voices.

**Keywords and phrases:** voice morphing, prototype waveform interpolation, lossless tube area function, speech synthesis.

## 1. INTRODUCTION

Voice morphing is the process of producing intermediate or hybrid voices between the utterances of two speakers. It can also be defined as the process of smoothly changing speech identity between two speakers [1], or gradually transforming the voice of a given speaker to that of another [2, 3]. The ability to change the speaker's individual characteristics and produce high-quality voices can be used in many applications. For example, in multimedia and video entertainment, voice morphing is similar to its visual counterpart: while seeing a face gradually changing from one person's to another's, we can simultaneously hear the voice progressively changing, as well. Another potential application is forensic voice identification: creating a voice bank of different pitches, rates, and timbres to assist in recognition of a suspect's voice. In a similar manner, producing a databank of different voices which are intermediates between several given speech recordings can enhance the possibility of synthesizing a given utterance more naturally. When prerecorded speech data is taken from different speakers, a natural-sounding new message, created by concatenation of speech segments taken from these speakers, may be achieved using speech morphing [1]. Speech and audio morphing can also be a valuable tool for voice and speaker perception research, for example, in an attempt to control the emotional content of speech [2]. Within a text-to-speech (TTS) synthesis framework, voice morphing also offers the opportunity to generate a variety of voices from a database containing only a small number of speakers. This is potentially advantageous since the voice creation process for a TTS system is quite time consuming, and it can also considerably reduce the memory requirements for storing TTS voices.

A successful procedure for voice morphing requires a representation of the speech signal in a parametric space, using a suitable mathematical model that allows interpolation between the characteristics of the two speakers. In other words, for the speech characteristics of the source speaker's voice to change gradually to those of the target speaker, the pitch, duration, and spectral parameters must be extracted from both speakers. Natural-sounding synthetic intermediates, with a new voice timbre, can then be produced.

Several studies have explored the subject of speech or voice morphing to date. Slaney et al. [3] used a representation of separate time-aligned spectrograms for pitch and spectral envelope, using MFCC, and modified the spectrograms separately to achieve an audio morph. Short vowels were used to demonstrate the resulting morphs. Morphing between a woman's vowel and a short note of an oboe was also used. A similar approach used smooth spectrographic representations to interpolate between utterances with different emotional contents [2]. In another study, real-time morphing was applied to a singing voice, using an interpolation of a source voice with a target voice, based on a sinusoidal model [4]. The same method of sinusoidal analysis was reported in [5]. In the digital music industry, the *Morpheus* synthesizer by E-mu [6] introduced a 14-pole dynamically variable filter which could model different resonant characteristics, perform spectral morphing-like effects between different musical samples, and interpolate between them in real time.

In this study we present a new technique which enables the production of a desired number of intermediate voices between the original voices of two speakers, or the production of one voice signal that changes gradually in time from one speaker to another. The latter means that, at the beginning of the utterance, the voice characteristics are those of one speaker, and the voice is perceived as belonging to that speaker. The voice is gradually modified towards the characteristics of another speaker, so that, by the end of the utterance, it is perceived as belonging to the second speaker. This technique is based on two components. One is the creation of a 3D prototype waveform interpolation (PWI) surface from the residual error signal which is obtained by LPC analysis to produce a new intermediate excitation signal. The second component is a representation of the vocal tract by a lossless tube area function, and interpolation of the parameters of the two speakers.

The morphing algorithm consists of two main stages: analysis and synthesis. In the analysis stage, the residual error signal is estimated, along with the vocal tract parameters. The residual signal is then used to create a PWI surface for each speech utterance. In the synthesis stage, a new residual error signal is recovered from a PWI surface interpolated from the two original surfaces. The area functions of the two speakers are also interpolated, producing a hybrid area function, from which a new vocal-tract filter is computed. The residual signal is then transferred through this filter to yield a morphed speech signal. Thus, we use an excitation signal the dynamics of which are comprised of both excitation waves and pitch period contours, along with a vocal tract with an interpolated structure.

This study resembles other studies on voice conversion or voice transformations [7, 8, 9, 10, 11, 12, 13, 14], but it is significantly different. Voice conversion modifies the utterances of one speaker so that his/her voice will sound like another (target) voice, by matching the source voice to the statistical properties of the target voice. In these studies, different methods are used to represent the relationships between the source and the target speakers, and most of the studies are concentrated on the spectral envelope data of short segments of speech. The spectral envelopes are characterized by one of several possible representations, such as HNM (8), Cepstrum or log-area ratio [8], LSF [9, 11], LPC [12], and formant frequencies [13]. For example, in [7] a Gaussian mixture model (GMM) is used, where the conversion is performed in the context of the harmonic + noise model (HNM), using a continuous probabilistic model of the source envelopes. Conversion using GMM is also utilized in [12], with joint density estimation for the spectral conversion using LPC analysis, while the pitch of the source has been modified to match the average pitch of the target. The residual LPC in each pitch period in the latter study was left intact. In other studies, the conversion is performed using codebook mapping with vector quantization [9, 13], or with artificial neural networks [14].

The aim of speech morphing, as it is proposed and used in [1, 2, 3], and as it has been carried out in the current study, is to produce intermediate voices between two given utterances that will be perceived as lying between the two original voices. In the other morphing type, gradual morphing, the morphed sound should be perceived as one object that smoothly changes into another sound [3]. The morphing algorithm presented here is shown to produce high-quality morphing sounds that are perceived as highly natural and smooth.

The paper is organized as follows. In Section 2.1, the PWI technique is introduced, and in Section 2.2, the idea of using it for speech morphing is presented. The basic morphing algorithm is described in Section 2.3. A detailed computation of the characteristic waveforms function, with the construction of corresponding PWI surface, and the interpolation between two such surfaces are all presented in Section 2.3.1. The procedure for extracting a new intermediate residual error signal is presented in Section 2.3.2. Subsequently, the calculation of the new vocal tract model and the synthesis of the morphed speech are described. In Section 2.4, subjective tests for evaluating the naturalness and the intelligibility of the morphing voices are presented. Finally, we discuss the advantages and disadvantages of the algorithm in contrast to previous studies.

## 2. PROCEDURE AND RESULTS

In this section, a method for decomposing the speech signals of two speakers and recombining the components is presented. The components are the excitation signal, represented by the residual error signal from an LPC analysis, and the vocal tract parameters, represented by the area coefficients of a lossless tube model. The method is designed so that the resulting speech will be characterized perceptually as an interpolated version of the voices of the two speakers.

### 2.1. Prototype waveform interpolation

PWI is a speech coding method described in [15, 16, 17, 18]. This method is based on the fact that voiced speech is quasiperiodic and can be considered as a chain of pitch cycles. Comparing consecutive pitch cycles reveals a slow evolution in the pitch-cycle waveform and duration, that is, each

pitch cycle has a close similarity to its neighbors. The slow change in the shape and duration of the pitch cycle suggests that extracting the cycles' waveform at regular time intervals should be sufficient in order to reconstruct the signal from the sampled cycles by interpolation. The interpolation procedure is carried out by constructing a 3D surface from the speech or the residual error waveforms. The coding procedure can be applied for both the speech signal and its residual error function, derived from the LPC analysis. A detailed description of this coding technique for bit-rate reduction can be found in [16]. The representation of a speech signal's residual error function in the form of 3D surfaces has been found useful for voiced speech morphing. The creation of such a surface is described below.

### 2.2. PWI-based speech morphing

Prototype waveform interpolation is based on the observation that during voiced segments of speech, the pitch cycles resemble each other, and their general shape usually evolves slowly in time (see [16, 17, 18]). The essential characteristics of the speech signal can, thus, be described by the pitch-cycle waveform. By extracting pitch cycles at regular time instants, and interpolating between them, an interpolation surface can be created. The speech can then be reconstructed from this surface if the pitch contour and the phase function (see Section 2.3.1) are known.

The algorithm presented here is based on the source-filter model of speech production [19, 20]. According to this model, voiced speech is the output of a time-varying vocal-tract filter, excited by a time-varying glottal pulse signal. In order to separate the vocal-tract filter from the source signal, we used the LPC analysis [21], by which the speech is decomposed into two components: the LPC coefficients containing the information of the vocal tract characteristics, and the residual error signal, analogous to the derivative of the glottal pulse signal. In the proposed morphing technique, we used the PWI to create a 3D surface from the residual error signal which would represent the source characteristics for each speaker. Interpolation between the surfaces of the two speakers allows us to create an intermediate excitation signal. In addition to the fact that the information of the vocal tract (see Section 2.3.3) is manipulated separately from the information of the residual error signal, it is also more advantageous to create a PWI surface from the residual signal than to obtain one from the speech itself. In this domain, it is relatively easy to ensure that the periodic extension procedure (see below) does not result in artifacts in the characteristic waveform shape [16]. This is due to the fact that the residual signal contains mainly excitation pulses, with low-power regions in between, and thus, allows a smooth reconstruction of the residual signal from the PWI surface with minimal phase discontinuities.

In the proposed algorithm, the surfaces of the residual error signals, computed for each voiced phoneme of two different speakers, are interpolated to create an intermediate surface. Together with an intermediate pitch contour and an interpolated vocal-tract filter, a new voiced phoneme is produced.

### 2.3. The basic algorithm

The morphing algorithm consists of two main stages—analysis and synthesis. As most of the speaker's individuality is contained in the voiced portion of speech [22], and because, in preliminary experiments, it was found that morphing the unvoiced sections yielded low-quality utterances, the algorithm is applied on the voiced segments only. The unvoiced segments were left intact, and concatenated with the interpolated voiced segments. Concatenation of voiced and unvoiced segments was performed by overlapping and adding two adjacent frames, about 20 milliseconds each, one from the voiced phoneme, and the other from the unvoiced one. The two frames are overlapped after each of them is multiplied by a half-left or half-right Hanning or linear windows, to yield a new frame that gradually changes from the voiced segment to the unvoiced segment or vice-versa. Since this operation may shorten the utterance, time-scale compensation is carried out for each vocal segment by extending the PWI surface as necessary.

The unvoiced segments are taken according to the morphing factor ($\alpha$): from the first speaker where $0 \leq \alpha < 0.5$, and from the second one where $0.5 \leq \alpha < 1.0$. The basic block diagram of the algorithm is shown in Figure 1.

In the analysis stage, the voiced segments of both speech signals are marked and each section in one of the voices is associated with the corresponding section in the other. The segmentation and mapping of the speech segments are done semiautomatically. First, a simple algorithm for voiced/unvoiced segmentation is applied, which is based on 3 parameters: the short-time energy, the normalized maximal peak of the autocorrelation function in the range of 3–16 milliseconds (the possible expected pitch period duration), and the short-time zero-crossing rate (Figure 2). The output of the automatic voiced/unvoiced segmentation is a series of voiced segments for each of the two voices. However, due to the imperfection of the segmentation algorithm, and the dissimilarity of the characteristics of the two voices, and as accurate mapping between the corresponding voiced sections of the two voices is crucial for the success of the algorithm, a manual correction mode has been added to refine the preliminary segmentation. The manual mode allows for making adjustments to the edges of the segments, splitting segments, joining segments, and adding new segments or deleting ones. For the demarcation of phoneme boundaries, the user can be assisted by the graph of the 2-norm of the difference between the MFCCs of adjacent frames (10 milliseconds each, see Figure 2).

It was found that, by applying manual segmentation as a refinement of the automatic segmentation, it is possible to reach accurate mapping with only small adjustments.

A pitch detection algorithm is applied to both speakers' utterances. The pitch detection algorithm is based on a combination of the cepstral method [23] for coarse-pitch period detection, and the cross-correlation method [24] for refining the results. Pitch marks are obtained, and after preemphasis, linear prediction coefficients are calculated for each voiced phoneme (either on the whole phoneme as one windowed

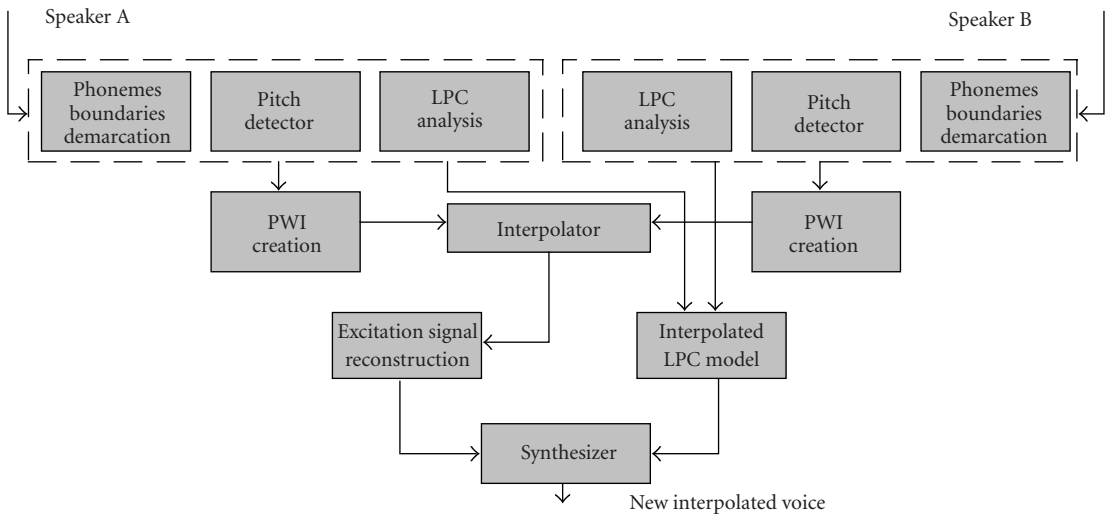


Figure 1: A basic diagram of the algorithm.


Figure 2: The semiautomatic segmentation of the speech signal. A simple algorithm for voiced/unvoiced segmentation is applied. The upper graph shows the speech signal with the v/uv boundaries. In the 4 graphs below the 4 parameters on which the decision is based are depicted the short-time energy, the zero-crossing rate, the autocorrelation, and the MFCC difference coefficient. The user can correct the segmentation manually with an interactive graphical user interface (see text).

frame, or pitch synchronously) to create the vocal-tract filter and the residual error function for each segment. The prototype waveform surfaces are then created from the residual error functions (see Section 2.3.1). In the synthesis stage, a new residual error signal is recovered from a PWI surface interpolated from the two original ones (as described in Section 2.3.2). The two speakers' area functions are also interpolated, producing an intermediate area function, from which a new vocal-tract filter is computed (see Section 2.3.3). The new residual signal is then used to excite the interpolated vocal-tract filter to yield an intermediate speech signal. The final morphed speech signal is created by concatenating the new vocal phonemes, in order, along with the unvoiced phonemes and silent periods of the source or of the target.

### 2.3.1. Computation of the characteristic waveform surface

The characteristic waveform surface, which represents the residual error signal derived from the voiced sections [16], is a two-dimensional signal that represents a one-dimensional signal, and is constructed as follows: let $u(t, \phi)$ be the characteristic waveform, where $t$ denotes the time axis, and $\phi$ is a phase variable whose values are in the range $[0, 2\pi]$. The prototype waveforms are displayed along the phase axis, where each prototype is a short segment from the residual signal with a length of one pitch period. Each prototype is considered as a periodic function, with a period of $2\pi$. The time axis of the surface displays the waveform evolution. A one-dimensional signal can be recovered from $u(t, \phi)$ by using a specific $\phi(t)$, so that

$$r(t) = u(t, \phi(t)), \tag{1}$$

where $\phi(t)$ is calculated using the signal pitch period function or pitch contour, $p(t)$ by

$$\phi(t) = \phi(t_0) + \int_{t_0}^{t} \frac{2\pi}{p(t)} dt. \tag{2}$$

A typical prediction error signal and its surface are shown in Figure 3.

In the proposed solution, the surface for each phoneme is created separately. The construction of such a surface is detailed in using the following procedure (Figure 4).

(1) Pitch detection is applied in order to obtain an instantaneous pitch value, $p(t)$, which will track the pitch cycle change through time. At any given point in time, the pitch cycle is determined by a linear interpolation of the pitch marks obtained by the pitch detector.

(2) A rectangular window with duration of one pitch period multiplies the residual error function around a sampling time $t_i$, with a step update of 2.5 milliseconds, to create a prototype waveform. In order to smooth the surface along the time axis, a low-order Savitzky-Golay filter is applied to the error function.

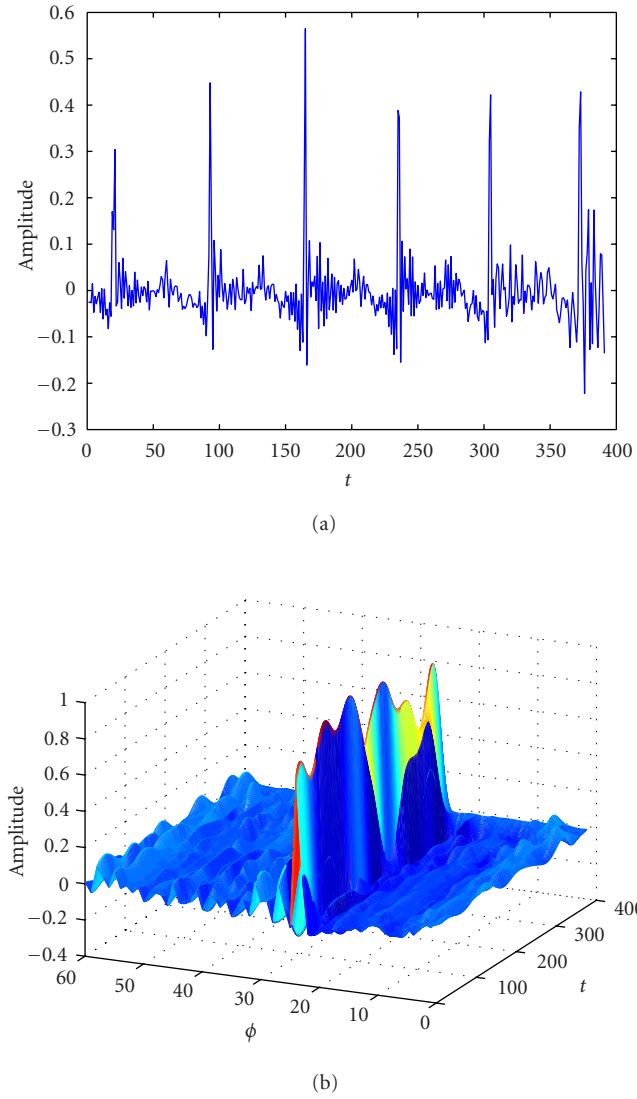


Figure 3: Creating the characteristic waveform surface. (a) A typical residual error signal, derived from the speech signal using linear prediction analysis. (b) The surface is constructed by plotting the prototype waveform along the $\phi$-axis at intervals of 2.5 milliseconds, after alignment and interpolation of the prototypes.

(3) For the reconstruction of the signal from the surface, it is extremely important to maintain similar and minimum energy values at both ends of the prototype waveform (which actually represent the same point due to the $2\pi$ periodicity along the phase axis). Therefore a shift of $\pm\Delta$ samples ($\Delta_{\max}$ was set to be 1 millisecond) is allowed for the location of the window's center in the construction of the PWI surface.

(4) Because the pitch cycle varies in time, each prototype waveform will be of different length. Therefore all prototypes must be aligned along $\phi = [0 - 2\pi]$ and must have the same number of samples.

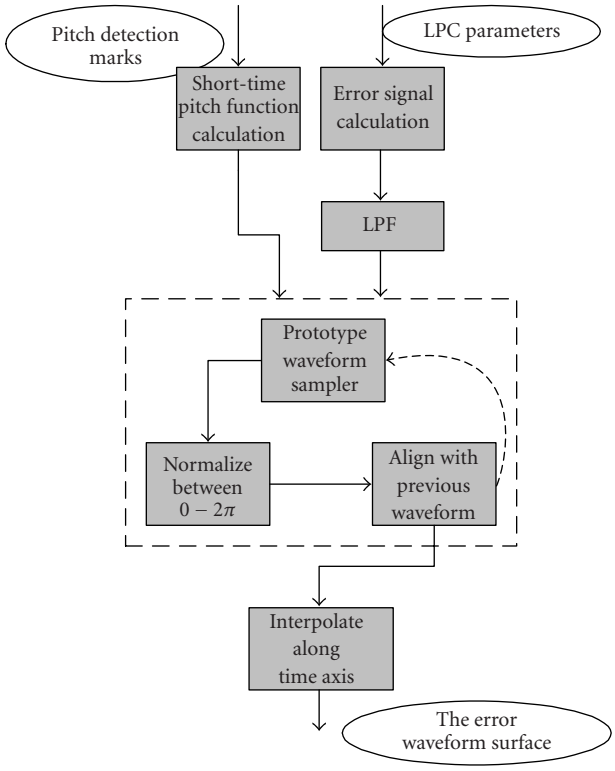


Figure 4: A schematic description of the PWI surface construction.

(5) Once a prototype waveform is sampled according to step (3) above, a cyclical shift along the $\phi$-axis is performed to obtain maximum cross-correlation with the former prototype waveform, thus creating a relatively smooth waveform surface when moving across the time axis.

(6) In order to create a surface that reflects the error's pitch cycle evolution over time, an interpolation along the time axis is performed.

### 2.3.2. Construction of the intermediate residual error signal

Let $u_s(t,\phi)$, $u_t(t,\phi) : \{t : [0 - T_s, T_t]; \phi : [0 - 2\pi]\}$ be the PWI of the source and target speakers, respectively. As described earlier, the new intermediate waveform surface will be an interpolation of the two surfaces (Figure 5). Therefore,

$$u_{\text{new}}(t,\phi) = \alpha \cdot u_s(\beta \cdot t, \phi) + (1-\alpha) \cdot u_{t-\text{aligned}}(\gamma \cdot t, \phi),$$
$$u_{\text{new}}(t,\phi) : \{t : [0 - T_{\text{new}}];\ \phi : [0 - 2\pi]\}, \tag{3}$$

where $\beta = T_{\text{new}}/T_s$, $\gamma = T_{\text{new}}/T_t$, and $\alpha$[1] is the relative part of $u_s(t,\phi)$.

[1]The factor may be invariant; if so, the voice produced will be an intermediate between the two speakers, or it may vary in time (from $\alpha = 1$ at $t = 0$ to $\alpha = 0$ at $t = T$), yielding a gradual change from one voice to the other.

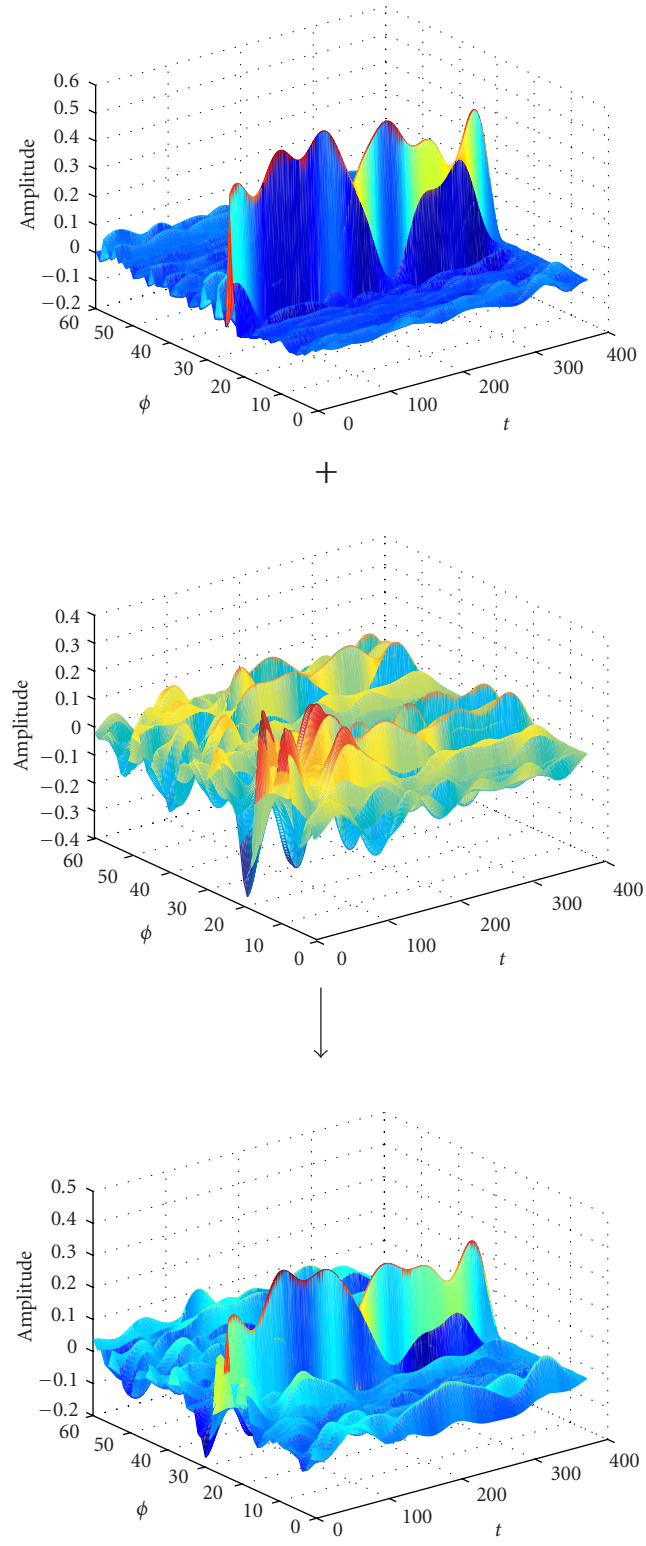


Figure 5: Residual PWI surfaces of the two speakers (upper and intermediate figures, respectively) and the resulting interpolated surface (bottom figure). It can be noticed that the interpolated surface combines characteristics of both speakers. The $\phi$-axis represents samples of the normalized pitch period, in a scale between 0 and $2\pi$.

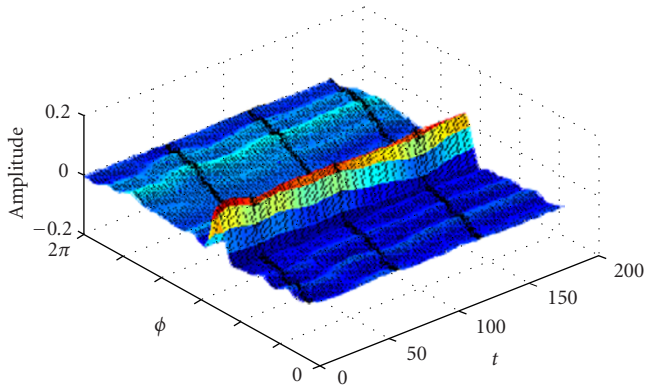


Figure 6: The new PWI surface is constructed by interpolation of two PWI surfaces of the residual error signals of two speakers. The new reconstructed residual can be recovered by moving on the surface along the phase axis according to the phase $\phi(t)$ determined by the new pitch contour. (The tracking is represented by the black line along the surface.)

In order to maximize the cross-correlation between the two surfaces, the target speaker's surface is shifted along the $\phi$-axis, and is referred to as $u_{t-\text{aligned}}(t,\phi)$, where

$$u_{t-\text{aligned}}(t,\phi) = u_t(t,\phi+\phi_n), \tag{4}$$

$$\phi_n = \arg\max_{\phi_e}\left\{\frac{\int_{\phi=0}^{2\pi}\int_{t=0}^{T_{\text{new}}} u_s(t,\phi)\cdot u_t(t,\phi+\phi_e)\,dt\,d\phi}{||u_s||\cdot||u_t(t,\phi+\phi_e)||}\right\}, \tag{5}$$

$$||u|| = \sqrt{\int_{\phi=0}^{2\pi}\int_{t=0}^{T_{\text{new}}} u(t,\phi)\cdot u(t,\phi)\,dt\,d\phi}, \tag{6}$$

where $\phi_e$ is the correction needed for the surfaces to be aligned.

The last step (after creating $u_{\text{new}}(t,\phi)$) is reconstructing the new residual error signal from the waveform surface. The reconstruction is performed by defining $e_{\text{new}}(t) = u_{\text{new}}(t,\phi_{\text{new}}(t))$, for all $t = [0 : T_{\text{new}}]$, where $\phi_{\text{new}}(t)$ is created by the following equation:

$$\phi_{\text{new}}(t) = \int_{t_0}^{t}\frac{2\pi}{p_{\text{new}}(t')}dt'. \tag{7}$$

$p_{\text{new}}(t)$ is calculated as an average of the source's and target's short-time pitch contour functions, as shown in the following equation:

$$\begin{aligned} p_{\text{new}}(t) &= \alpha\cdot p_s(\beta\cdot t) + (1-\alpha)\cdot p_t(\gamma\cdot t),\\ \beta &= \frac{T_{\text{new}}}{T_s},\\ \gamma &= \frac{T_{\text{new}}}{T_t}, \end{aligned} \tag{8}$$

where the factors $\beta$ and $\gamma$ are scaling factors for the new time axis $[0, T_{\text{new}}]$, and $\alpha$, as in (3), is a weighting factor between the pitch of the source and the pitch of the target. Figure 6 shows the derivation of a new residual error signal using the track on the PWI surface determined by $\phi_{\text{new}}(t)$.

#### 2.3.3. New vocal tract model calculation and synthesis

It is well known that the linear prediction parameters (i.e., the coefficients of the predictor polynomial A(z)) are highly sensitive to quantization [19]. Therefore their quantization or interpolation may result in an unstable filter and may produce an undesirable signal. However, certain invertible nonlinear transformations of the predictor coefficients result in equivalent sets of parameters that tolerate quantization or interpolation better. An example of such a set of parameters is the PARCOR coefficients ($k_i$), which are related to the areas of lossless tube sections modeling the vocal tract [20], as given by the following equation:

$$A_{i+1} = \left(\frac{1-k_i}{1+k_i}\right)\cdot A_i. \tag{9}$$

The value of the first area function parameter ($A_1$) is arbitrarily set to be 2. A new set of LPC parameters that defines a new vocal tract is computed using an interpolation of the two area vectors (source and target). This choice of the area parameters seems to be more reasonable, since intermediate vocal tract models should reflect intermediate dimensions [25]. Let the source and target vocal tracts be modeled by $N$ lossless tube sections with areas $A_i^s, A_i^t : \{i : [1-N]\}$, respectively. The new signal's vocal tract will be represented by

$$A_i^{\text{new}} = \alpha\cdot A_i^s + (1-\alpha)\cdot A_i^t \quad \forall i : [1,2,\ldots,N]. \tag{10}$$

After calculating the new areas, the prediction filter is computed and the new vocal phoneme is synthesized according to the following scheme (Figure 7):

(1) compute new PARCOR parameters from the new areas by reversing (9);
(2) compute the coefficients of the new LPC model from the new PARCOR parameters;
(3) filter the new excitation signal through the new vocal-tract filter to obtain the new vocal phoneme.

When temporal voice morphing is applied, informal subjective listening tests performed on different sets of morphing parameters have revealed that in order to have a "linear" perceptual change between the voices of the source and the target, the coefficient $\alpha(t)$ (the relative part of $u_s(t,\phi)$) must vary nonlinearly in time (like the one in Figure 8). When the coefficient changed linearly with time, the listeners perceived an abrupt change from one identity to the other. Using the nonlinear option, that is, gradually changing the identity of one speaker to that of another, a smooth modification of the "source" properties to those of the "target" properties was achieved. In another subjective listening test (see below) performed on morphing from a woman's voice to a man's voice, and vice versa, both uttering the same sentence, the morphed sound was perceived as changing smoothly and naturally from one speaker to the other. The quality of the morphed voices was found to depend upon the specific speakers, the differences between their voices, and the content of the utterance. Further research is required to accurately evaluate the effect of these factors.

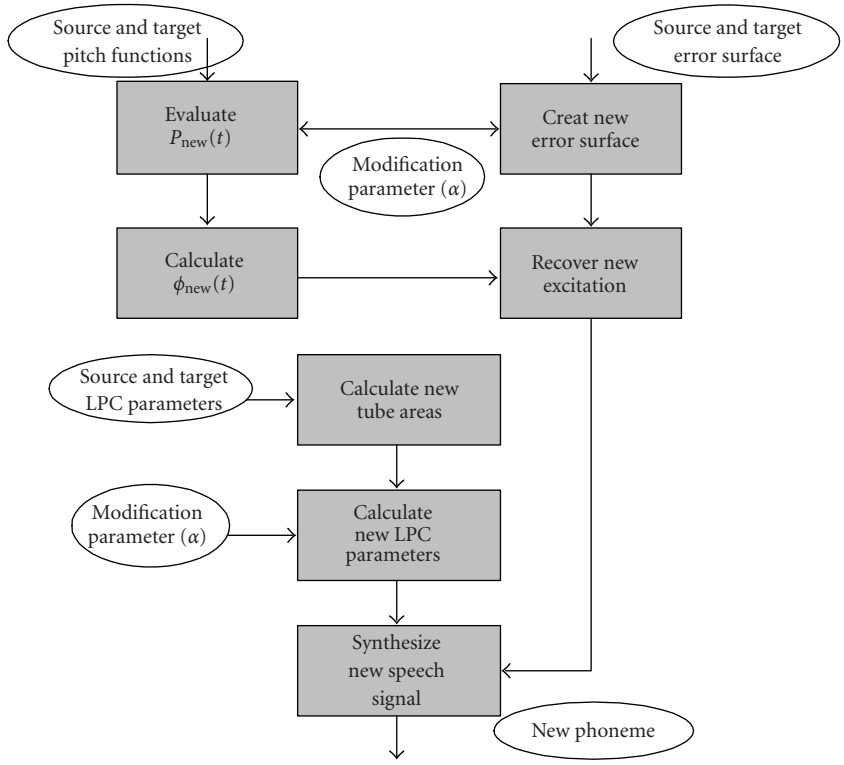


Figure 7: A block diagram of the procedure for synthesizing the new phoneme.

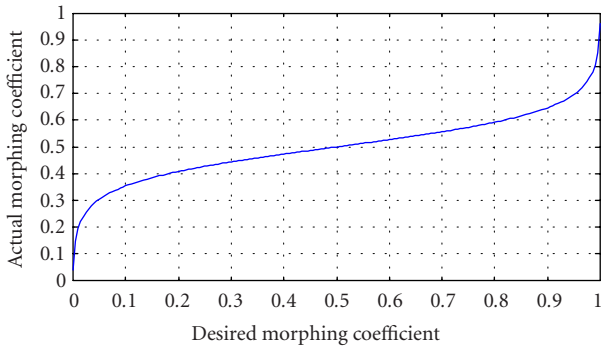


Figure 8: An example of a nonlinear morphing function, obtained empirically, and used to achieve a linear perceptual change between the first speaker and the second speaker, when time-varying morphing was performed.

### 2.4. Evaluation of the algorithm

Evaluation of the algorithm was carried out using three subjective listening tests. The naturalness, intelligibility, identity change, and smoothness of the morphing algorithm were examined in these psychoacoustic tests. In all tests, six listeners participated (three males and three females, all without hearing impairments, and inexperienced in speech morphing). The sentences for the tests were taken from TIMIT database, and from BGU Hebrew database. The speech stimuli were played using high-quality loudspeakers. In all tests, the stimuli were played in random order. The listeners were free to play each stimulus more than once, without any limitation. In the first test, the listeners had to decide if there was a change in the identity of the speaker for each of the three sentences, on a 1–5 scale, where 1 meant that one identity was perceived along the utterance and 5 meant that there was more than one speaker. The listeners had to rate the smoothness of the utterance, as well, on a similar 1–5 scale, where 5 meant smooth utterance and 1 meant abrupt change or changes in the utterance. Three types of stimuli were used: the original speech, morphed speech (in which the identity changed gradually from one speaker to another during the sentence), and concatenated speech of two speakers, at a fixed point [1]. The aim of this test was to evaluate if the identity change was perceived, and to assess the effect of the algorithm on the smoothness of the gradual morphed utterances. The results of this experiment for one of the sentences ("We were different shades, and it did not make a bit of difference among us," taken from [9]) are depicted in Figure 9. As expected, it is readily seen that the original utterance was perceived as a smooth speech without identity change, while the abrupt change in identity in the concatenated utterance was perceived and rated accordingly, that is, as more than one identity and with an abrupt change. The change of identity was perceived in the morphed signal, as well (average score 3.5, std. = 0.96), since the sentence was long enough to notice the modification. Nevertheless, it was also perceived as relatively smooth (average score 3.5, std. = 1.12). The other

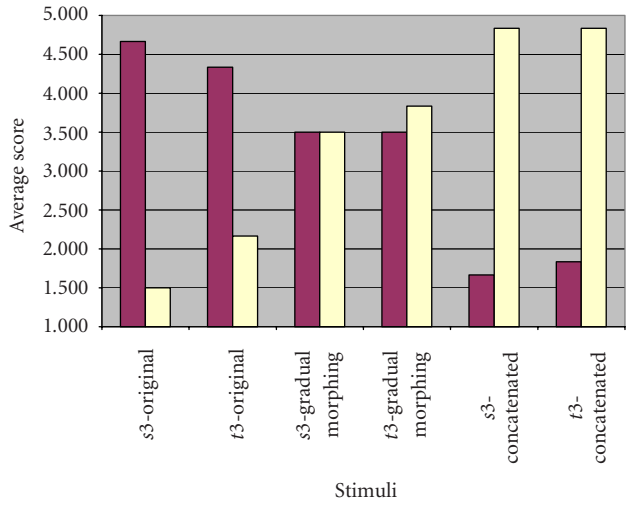


Figure 9: Evaluation of smoothness and identity change in an utterance in which a gradual morphing is produced (smoothness—left column, identity change—right column).

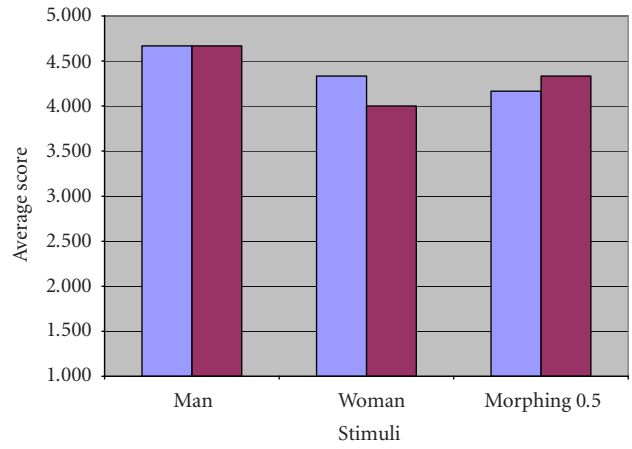


Figure 10: Results of subjective evaluation for naturalness (left bars) and intelligibility (right bars) for 3 stimuli: a man (left) and a woman (middle) voices, and a morphing between the two voices (right).

two utterances yielded similar results. In the second test, the listeners had to evaluate the naturalness and intelligibility of a given sentence on a 1–5 scale. The listeners attended to 3 stimuli: two original sentences, one uttered by a male speaker, and the other by a female speaker, and a morphed sentence with a morphing factor of 0.5, which is a hybrid signal between the two originals. The results are summarized in Figure 10. It is clear from this figure that the morphed signal was perceived as highly natural and clear by most of the listeners (naturalness average score: 4.3, std. = 0.74, intelligibility average score: 4.17, std. = 0.69). In the third experiment, the listeners had to rate the identity change and the naturalness of two sentences on a 1–5 scale. Each sentence was repeated as a cyclostationary morph (see [3]) 16 times with various morphing factors, from 0 to 1. The average naturalness was 3.25, and the identity change was rated with an average of 3.3. Naturalness in this case was not perfect (although it is quite high), and can be explained by the fact that the voice timbers in this test were distant, and the hybrid voice that was produced could be perceived as uncommon. In summary, the results of the tests show that the morphed signals are perceived in most cases as highly natural, highly intelligible, with a relatively smooth change from one speech voice to another.

## 3. CONCLUSIONS

In this study, a new speech morphing algorithm is presented. The aim is to produce natural sounding hybrid voices between two speakers, uttering the same content. The algorithm is based on representing the residual error signal as a PWI surface, and the vocal tract as a lossless tube area function. The PWI surface incorporates the characteristics of the excitation signal, and enables reproduction of a residual signal with a given pitch contour and time duration, which includes the dynamics of both speakers' excitations. It is known [16] that PWI surfaces can be exploited efficiently for speech coding, and therefore, they allow for higher compression of the speech database. The area function was used in an attempt to reflect an intermediate configuration of the vocal tract between the two speakers [25]. The utterances produced by the algorithm were shown to be of high quality and to consist of intermediate features of the two speakers.

There are at least two modes in which the morphing algorithm can be used. In the first mode the morphing parameter is invariant, meaning, for example, taking a factor of 0.5, and receiving a morphed signal with characteristics which are between the two voices for the whole duration of the articulation. In the second mode, we start from the first (source) speaker (morphing factor = 0), and the morphing factor is changed gradually along the duration of the sentence, so its value is 1 at the end of the sentence. The same morphing factor was used for both the excitation and the vocal tract parameters.

In this time-varying version of the algorithm, that is, when morphing gradually from one voice to another over time, smooth morphing was achieved, producing a highly natural transition between the source and the target speakers. This was assessed by subjective evaluation tests, as previously described.

The algorithm described here is more capable of producing longer and more smooth and natural sounding utterances than previous studies [1, 3]. One of the advantages of the proposed algorithm is that, in addition to the interpolation of the vocal tract features, interpolation is also performed between the two PWI surfaces of the corresponding residual signals, and thus captures the evolution of the excitation signals of both speakers. In this way, a hybrid excitation signal can be produced, that contains intermediate characteristics of both excitations. Thus, a more natural morphing between utterances can be achieved, as has been demonstrated. Furthermore, our algorithm performs the interpolation of

the residual signal regardless of the pitch information, since the pitch data is normalized within the PWI surface. Therefore, the morphed pitch contour is extracted independently, and can be manipulated separately. In addition, the current approach enables morphing between utterances with different pitches, between male and female voices, or between voices of different and perceptually distant timbres. Kawahara and his colleagues [26, 27] implemented a morphing system based on interpolation between time-frequency representations of the source and the target signals. It appears that the STRAIGHT-based morphing system (see [2, 26]) was able to produce intermediate voices of higher clarity than our algorithm, but the need to assign multiple anchor points for each short segment, using visual inspection and phonological knowledge, is a noticeable disadvantage of that system, which can make it difficult to use for morphing between long utterances.

Further research is needed to improve the quality of the morphed signals, which are natural sounding (see http://spl.telhai.ac.il/speech/), but are somewhat degraded compared to the originals.

## ACKNOWLEDGMENTS

We would like to thank Professor David Malah, the Head of SIPL, for proposing to apply PWI to voice morphing, and for his valuable discussions and comments. We also thank Dima Ruinskiy for his valuable assistance, for programming part of the morphing system, and for preparing the semiautomatic segmentation, and Yefim Yakir for preparing part of the figures, and for technical support. The authors thank the reviewers for their helpful comments. This study was partly supported by Guastella Fellowship of the Sacta-Rashi Foundation, and the JAFI project.

## REFERENCES

[1] M. Abe, "Speech morphing by gradually changing spectrum parameter and fundamental frequency," Tech. Rep. SP96-40, The Institute of Electronics, Information and Communication Engineers (IEICE), July 1996.

[2] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 256–259, Hong Kong, China, 2003.

[3] M. Slaney, M. Covell, and B. Lassiter, "Automatic audio morphing," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 1001–1004, Atlanta, GA, USA, May 1996.

[4] P. Cano, A. Loscos, J. Bonada, M. de Boer, and X. Serra, "Voice morphing system for impersonating in karaoke applications," in *Proc. International Computer Music Conference (ICMA '00)*, pp. 109–112, Berlin, Germany, August 2000.

[5] E. Tellman, L. Haken, and B. Holloway, "Timbre morphing of sounds with unequal numbers of features," *Journal of the Audio Engineering Society (AES)*, vol. 43, no. 9, pp. 678–689, 1995.

[6] D. Rossum, "The "armadillo" coefficient encoding scheme for digital audio filters," in *proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '91)*, pp. 129–130, New Paltz, NY, USA, October 1991.

[7] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[8] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Communication*, vol. 16, no. 2, pp. 139–151, 1995.

[9] L. M. Arslan, "Speaker Transformation Algorithm using Segmental Codebooks (STASC)," *Speech Communication*, vol. 28, no. 3, pp. 211–226, 1999.

[10] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '98)*, vol. 1, pp. 285–288, Seattle, WA, USA, May 1998.

[11] A. Kain and M. Macon, "Personalizing a speech synthesizer by voice adaptation," in *Proc. 3rd ESCA/COCOSDA International Speech Synthesis Workshop*, pp. 225–230, Jenolan, Caves, Australia, November 1998.

[12] A. Kain and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 2, Salt Lake City, Utah, USA, May 2001.

[13] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '88)*, pp. 655–658, New York, NY, USA, April 1988.

[14] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," *Speech Communication*, vol. 16, no. 2, pp. 206–216, 1995.

[15] O. Lev (Fellah) and D. Malah, "Low bit-rate speech coder based on a long-term model," in *Proc. 22nd IEEE Convention of Electrical and Electronics Engineers in Israel*, Tel-Aviv, Israel, December 2002.

[16] W. B. Kleijn and J. Haagen, "Waveform interpolationfor coding and synthesis," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 5, pp. 175–207, Elsevier Science B. V., Amsterdam, the Netherlands, 1995.

[17] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 386–399, 1993.

[18] W. B. Kleijn and J. Haagen, "Transformation and decomposition of the speech signal for coding," *IEEE Signal Processing Lett.*, vol. 1, no. 9, pp. 136–138, 1994.

[19] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*, chapter 7, Macmillan Publishing, New York, NY, USA, 1993.

[20] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, chapter 8, Signal Processing Series. Prentice-Hall Publishing, Englewood Cliffs, London, 1978.

[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

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

[23] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[24] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, no. 1, pp. 40–48, 1991.

[25] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of the acoustic speech waveforms," *IEEE Transaction Audio Electroacoustic*, vol. AV-21, no. 5, pp. 417–427, 1973.

[26] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[27] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," in *Proc. European Union Activities in Human Language Technologies (EUROSPEECH '99)*, vol. 6, pp. 2781–2784, Budapest, Hungary, September 1999.

**Yizhar Lavner** received a Ph.D. degree from the Technion – Israel Institute of Technology in 1997. He joined the Department of computer science, Tel-Hai Academic College, Upper Galilee, Israel, in 1997, where he is a Senior Lecturer. He is teaching in SIPL (Signal and Image Processing Lab), Electrical Engineering Faculty, the Technion, Haifa, Israel), since 1998. His research interests include speech and audio signal processing and genomic signal processing.

**Gidon Porat** received his B.S. degree (cum laude) in electrical engineering from the Technion – Israel Institute of Technology in January 2003. As a part of his graduation requirements, Porat has researched voice morphing in the Electrical Engineering Faculty's Signal and Image Processing Lab in the Technion. Porat was a recipient of a Best Student's Paper Award at the IEEE 22nd Convention of Electrical & Electronics Engineers, Israel, in December 2002. Since his graduation, Porat has been employed by Intel Corporation as a member of the Mobile Platform Group Chip Design Team. Porat takes part in the development of high-speed, low-power circuit designs for the next-generation CPUs for mobile computers.