Voice morphing by 3-D waveform interpolation surface and lossless tube area function

G. Porat (1), Y. Lavner (1,2)

(1) Signal and Image Processing Lab (SIPL), Faculty of Electrical Engineering, Technion IIT, Haifa, Israel
(2) Computer Science, Tel-Hai Academic College, Upper Galilee, Israel.

1. Introduction

Voice morphing is the process of gradually transforming the voice of a given speaker to that of another. 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 just like its visual counterpart: while seeing a face gradually changing from one person's to another's, we can simultaneously hear the voice changing as well. Another application could be in forensic voice identification: creating a voice-bank of different pitches, rates, and timbres, to assist in recognition of the suspect's voice. In this study we present a new technique, which enables the production of N intermediate voices that gradually change between voices of two speakers, or one voice signal that changes gradually. This technique is based on two components. One is creating a 3-D prototype waveform interpolation (PWI) surface from the residual error signal, which is estimated from 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 two speakers' parameters.

2. The challenge

In order to perform voice morphing successfully, the speech characteristics of the source speaker's voice must change gradually to those of the target speaker's; therefore, the pitch, duration, and spectral parameters must be extracted from both speakers. Then, natural-sounding synthetic intermediates have to be produced.

3. Solution method

3.1 Prototype Waveform Interpolation: PWI is a speech coding method described in [1], [2]. It is based on the fact that voiced speech is quasi-periodic and can be considered as a chain of pitch cycles. The slow change in the shape and duration of the pitch cycle suggests that sampling these cycles at regular time intervals should be sufficient in order to reconstruct the signal later. This coding procedure can be applied for both the speech signal and its residual error function from the LPC analysis. The exact *coding* technique is described in [1] and is *not* a part of the offered solution. However, presentation of a speech signal's error function in the form of 3-D surfaces was found useful for *voiced* speech morphing. The creation of such a surface will be briefly described later.

3.2 PWI speech morphing: The ability to represent a speech signal with a prototype 3-D surface (a surface from which the speech signal can be reconstructed, and which holds the key to the speech signal's pitch evolution in time), and the fact that the spectrum of the error signal is relatively flat, thus eliminating the effects of the formants, leads to the assumption that little degradation will occur while interpolating it, as opposed to the speech signal itself.

In the offered solution, the surfaces of one *voiced phoneme's error signal* of two different speakers are interpolated to create an intermediate one. Together with an intermediate short-time pitch function and an interpolated vocal-tract filter, an intermediate voiced phoneme is produced.

3.3 The basic algorithm: The morphing algorithm consists of two main stages – analysis and synthesis – and is only applied on the voiced speech segments, due to the fact that they carry most of the speaker's individuality [3]. The basic block diagram can be seen in figure 1.

In the analysis stage (1st and 2nd layers), the voiced segments of both speech signals are marked and associated with their parallels (mapping between two speakers). Pitch marks are made and linear prediction coefficients are calculated for each voiced phoneme, to create the vocal-tract filter and the residual error function. The prototype waveform surfaces are then created from the residual error functions (as described in 3.3.1). In the synthesis stage (3rd and 4th layers), a new residual error signal is recovered from a PWI surface interpolated from the two original ones (as described in 3.3.2). The two speakers' area functions are also interpolated, producing an intermediate area function, from which a new vocal-tract filter is computed (described in 3.3.3). The new residual signal is then transferred through the interpolated vocal-tract filter to yield an intermediate speech signal. The final and new speech signal is created by concatenating the new vocal phonemes, in order, along with the source's/target's unvoiced phonemes and silent periods.

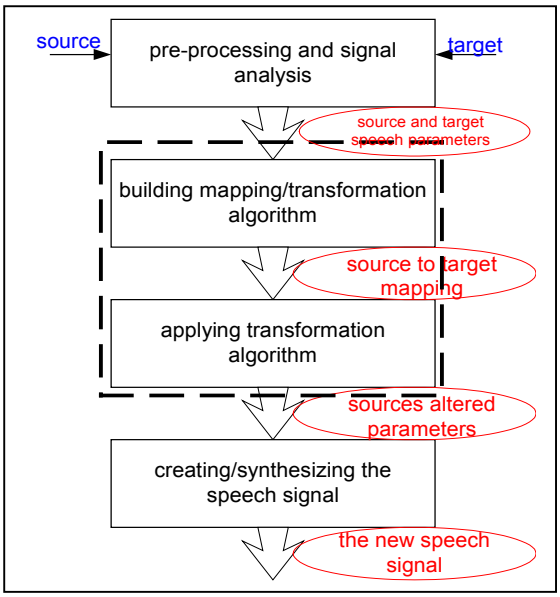

Figure 1 - Basic block diagram

3.3.1 Computation of characteristic waveform surface [1]:

The surface is displayed along (t) – the time axis and (ϕ) – the phase axis. The prototype waveforms are displayed along the phase axis, while the time axis displays the waveform evolution. A typical prediction error signal and its surface are shown in figure 2. In the process of creating the characteristic waveform surface, segments of the linear prediction residual signal are aligned to create a 3D surface. These segments are sampled every 2.5 msec. and are one pitch cycle long. These segments are called "prototype waveforms." The surface is created for each phoneme separately, as follows:

1. Pitch detection is applied in order to create a short-time pitch cycle function, $p(t)$, that will track the pitch cycle change through time. At any given point in time, the pitch cycle is determined by an interpolation of the pitch marks obtained by the pitch detector.
2. A rectangular window with duration of one pitch period multiplies the error function around a sampling time t_i, which changes at a rate of 2.5 msec, to create a prototype waveform. In order to get maximum "smoothness" of the surface along the time axis, a low-pass filter is applied to the error function.
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 2π periodicity along the phase axis). Therefore a shift of $\pm\Delta$ samples (about 1 msec.) is allowed for the location of the window's center in the prototype waveform surface construction.
4. Because the pitch cycle varies over time, each prototype waveform will be of different length. Therefore all prototypes must be aligned between $\phi=[0-2\pi]$ and have the same number of samples.
5. Once a prototype waveform is sampled and aligned between $\phi=[0-2\pi]$, a cyclical shift around the ϕ axis is needed to create the best 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

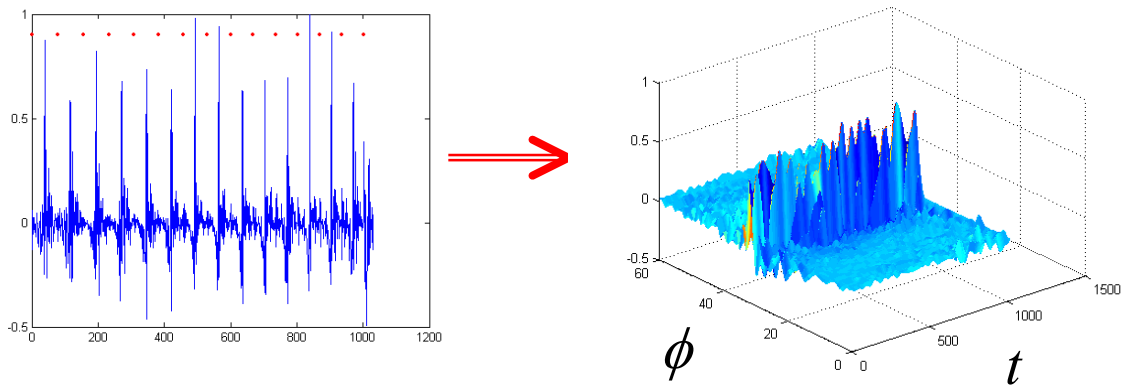

Figure 2 - An error function (left) with its waveform surface (right)

3.3.2 The Interpolator Morphing System:

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[1] waveform surface will be an interpolation of the two surfaces. Therefore:

$u_{new}(t,\phi)=\alpha\cdot u_s(\beta\cdot t,\phi)+(1-\alpha)\cdot u_{t-aligned}(\gamma\cdot t,\phi)$; $u_{new}(t,\phi):\{t:[0-T_{new}];\phi:[0-2\pi]\}$ In order to maximize the cross-correlation between the two surfaces, the target speaker's surface is shifted along the ϕ axis, and is referred to as $u_{t-aligned}(t,\phi)$, where $u_{t-aligned}(t,\phi)=u_t(t,\phi+\phi_n)$;

$$1)\ \phi_n=\arg\max_{\phi_e}\{\frac{\int_{\phi=0}^{2\pi}\int_{t=0}^{Tnew}u_s(t,\phi)\cdot u_t(t,\phi+\phi_e)dtd\phi}{\|u_s\|\cdot\|u_t(t,\phi+\phi_e)\|}\};\ \|u\|=\sqrt{\int_{\phi=0}^{2\pi}\int_{t=0}^{Tnew}u(t,\phi)\cdot u(t,\phi)dtd\phi}$$

[1] The factor α could be invariant, and then the voice produced will be an intermediate between the two speakers, or it could vary in time, (from α=1 at t=0 to α=0 at t=T), yielding a gradual change from one voice to the other.

where $\beta = T_{new} / T_s ; \gamma = T_{new} / T_t ; \alpha$ is the relative part of $u_s(t,\phi)$, and ϕ_e is the correction needed for the surfaces to be aligned. The last step (after creating $u_{new}(t,\phi)$) is reconstructing the new residual error signal from the waveform surface. The reconstruction is performed by defining: $e_{new}(t) = u_{new}(t, \phi_{new}(t)), \forall t = [0 : T_{new}]$, where $\phi_{new}(t)$ is created by the following equation:

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

$p_{new}(t)$ is calculated as an average of the source's and target's short-time pitch functions, as shown in equation 3:

3) $$p_{new}(t) = \alpha \cdot p_s(\beta \cdot t) + (1-\alpha) \cdot p_t(\gamma \cdot t) ; \ \beta = T_{new} / T_s ; \gamma = T_{new} / T_t$$

3.3.3 New LPC creator and the synthesizer: It is well known that prediction parameters (i.e., the coefficients of the predictor polynomial A(z)) are highly sensitive to quantization [4], because they are usually small and are not defined in a linear space. Therefore their quantization may result in an unstable filter and an unrecognizable speech signal. However, certain invertible nonlinear transformations of the predictor coefficients result in equivalent sets of parameters that tolerate quantization better. An example of such a set of parameters are the PARCOR coefficients (k_i), which are related to the areas of the lossless tubes, as shown in the equation 4:

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

A new set of LPC parameters that define a new vocal tract are computed using an interpolation of the two area vectors (source and target). Let the source and target vocal tracts be modeled by N lossless tubes with areas $A_i^s, A_i^t : \{i : [1-N]\}$, respectively. Then the new signal's vocal tract will be represented by:

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

After calculating the new areas, the prediction filter is computed and the new vocal phoneme is synthesized according to the following schema:

1. Calculate new PARCOR parameters from the new areas by reversing equation 4.
2. Calculate new LPC from PARCOR.
3. Filter the new error signal through the new vocal tract filter to obtain the new vocal phoneme.

Hearing tests performed on different sets of morphing parameters revealed that in order for one to hear a "linear" change between the source's and the target's voices, the coefficient, $\alpha(t)$ (the relative part of $u_s(t,\phi)$) has to vary in time nonlinearly. An example of such a function is plotted in figure 3. The quality of the morphed voices was found to depend upon the specific speakers, the difference between their voices, and the content of the utterance. Further research is required to evaluate this dependency.

3.3.4 Conclusions: In this study a speech morphing algorithm was presented. The algorithm is based on a representation of a PWI surface for the residual error signal, and a lossless tube area function for the vocal tract. The utterances produced by the algorithm were shown to consist of intermediate features of the two speakers. Although most of the speaker's individuality is concentrated in the voiced speech segments, degradation could be noticed, when interpolating between two speech signals that differ greatly in their unvoiced speech segments, such as heavy breathing, long periods of silence, etc.

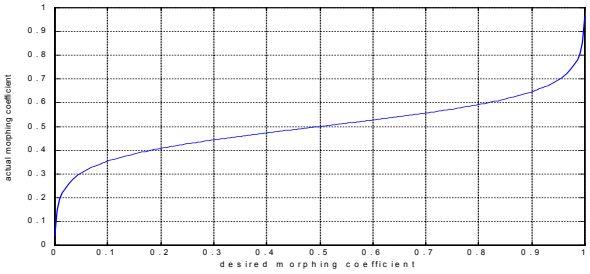

Figure 3 - "Morphing transfer function"

4. References

[1] Fellah Orit, "Low bit rate speech coding based on a long term model", Master thesis, Technion, Haifa, (2000).

[2] Kleijn, W.B. and Haagen, J., "Waveform Interpolation for Coding and Synthesis",Chapter 5 in W.B. Kleijn and K. K. Paliwal, Eds., "*Speech Coding and Synthesis*", Elsevier Science B.V., (1995).

[3] Hisao Kuwabara, Yoshinori Sagisaka, "Acoustic characteristics of speaker individuality : control and conversion", *Speech Communication* 16 (1995) 165-173.

[4] Deller, J.R., Proakis, J.G. and Hansen J.H.L. *Discrete Time Processing of Speech Signals*. New York: Macmillan, (1993).

Acknowledgment: We would like to thank Prof. David Malah for the idea of applying PWI to voice morphing, and for his valuable discussions and comments.

This study was partly supported by a Guastella Fellowship of the Sacta-Rashi Foundation, and the JAFI project to promote higher education in the Eastern Galilee.