SIGNAL APPROXIMATION IN HILBERT SPACE AND ITS APPLICATION ON ARTICULATORY SPEECH SYNTHESIS

Jun Huang, Stephen Levinson and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering,
University of Illinois, Urbana, IL 61801, USA

ABSTRACT

In this paper, we apply signal approximation theory to estimate the articulatory trajectory for a nonlinear speech synthesizer. First, we analyze the L_2 error bounds of interpolation and LS approximation in the Hilbert space framework. Second, we use two signal approximation techniques, specifically, cubic spline interpolation and the LS approximation to estimate the trajectory of articulatory parameters given the static articulatory parameters of discrete phonemes. In our articulatory speech synthesizer, we assume that the sound propagation inside the human vocal tract is a three-dimensional non-plane wave propagation inside a viscous fluid described by the governing Navier-Stoke equations. Finally, we present some experimental results of the estimated articulatory trajectories of English diphthongs and the synthesized phonemes.

1. INTRODUCTION

Articulatory synthesis is a parametric approach which attempts to produce speech by first understanding how the vocal apparatus changes shape during speech production, then understanding the acoustic problem of how those movements translate into sounds. One important problem in articulatory speech synthesis is to estimate the articulatory movement patterns during speech production. Several techniques exist for the motor control problem of speech production. Saltzman *et al.* proposed a task-dynamic model based on the assumption that the motion of vocal tract variables can be described by damped, second-order linear differential equations [1]. Nelson designed several articulator motor control systems based on some minimum cost principles [3]. In this paper, we applied signal approximation theory in the Hilbert space framework to estimate the dynamic articulatory trajectory for speech synthesis. The advantage of our approach is that it provides a general framework and some quantitative error bounds can be attained. This paper is organized as following. In section 2, we will first review some principles of signal approximation theory and analyze the pointwise error bounds and the L_2 error bounds of (quasi-)interpolation and least squares (LS) approximation methods, respectively. In section 3, we describe the application to dynamic articulatory parameter estimation. The experimental results of articulatory speech synthesis are presented in section 4. Finally, a conclusion is made in section 5.

2. ERROR BOUND ANALYSIS OF SIGNAL APPROXIMATION

Suppose that we are given a vector space of functions X and a set of samples from a function $x(t) \in X$, the objective of signal approximation is to find an element $\hat{x} \in X$ that is the optimal approximation to $x(t)$ according to certain minimum error criterion. In practice, most researchers use short kernel methods such as bilinear interpolation, cubic convolution or polynomial spline interpolation, which are much more efficient to implement, especially in higher dimensions. These methods are all convolution-based in the sense that they use an interpolation model of the form:

$$(I_h s)(x) = \sum_{k \in Z} c_h(k) \varphi(\frac{x}{h} - k) \tag{1}$$

where h is the sampling step and $\varphi(x)$ is the basic interpolation kernel. The expansion coefficients in Eq. (1) typically correspond to the samples of the input function $s(x)$ taken on a uniform grid: $c_h(k) = s(hk)$.

More recently, researchers have proposed a systematic formulation of this class of representations using the Hilbert space framework. Let's define the signal subspace as $V_h = span\{\varphi(\frac{x}{h} - k)\}_{k \in Z}$. A more sophisticated approach for obtaining a representation of the signal $s(x) \in L_2$ in $V_h(\varphi)$ is to determine its minimum L_2-norm approximation (orthogonal projection). This LS approximation is given by:

$$(P_h s)(x) = \sum_{k \in Z} c(k) \varphi(\frac{x}{h} - k) \tag{2}$$

$$c(k) = \frac{1}{h} < s(x), \tilde{\varphi}(\frac{x}{h} - k) > \tag{3}$$

where $< \cdot, \cdot >$ denotes inner product, $\tilde{\varphi}$ is the dual of φ and is defined by:

$$\tilde{\Phi}(\omega) = \frac{\Phi(\omega)}{a_\varphi(\omega)} \tag{4}$$

$$a_\varphi(\omega) = \sum_{n \in Z} |\Phi(\omega + 2n\pi)|^2 \tag{5}$$

where $\Phi(\omega)$ and $\tilde{\Phi}(\omega)$ denote the Fourier transforms of $\varphi(x)$ and $\tilde{\varphi}(x)$, respectively.

In this section, we will briefly compare the L_2 error bounds of interpolation and LS approximation. The detailed proof is not included in this paper, but some relevant information can be found in [6], [7].

2.1. L_2 Error of Interpolation

The L_2 error bound of the interpolator is given as following. If φ is a quasi-interpolant of order L with sufficient decay, then:

$$\forall s \in W_2^L, \|s - I_h s\|_2 \leq C_{\varphi,L} \cdot h^L \cdot \|s^{(L)}\|_2 \quad (6)$$

$$C_{\varphi,L} = \frac{1}{L!} \sup_{x\in[0,1]} \sum_{k\in Z} |x-k|^L |\varphi(x-k)| \quad (7)$$

where $s^{(L)}(x)$ is the L-th derivative of $s(x)$, and W_2^L is the Sobolev space of all functions whose first L derivatives belong to a L_2 space. Here the L_2 is the space of measurable, square-integrable, real-valued functions.

2.2. L_2 Error of the LS Approximation

The Fourier representation of the LS approximation of Eq. (2) is given by:

$$(\hat{P_h} s)(\omega) = \Phi(h\omega) \sum_{k\in Z} \overline{\tilde{\Phi}(h\omega + 2\pi k)} S(\omega + \frac{2\pi}{h}k) \quad (8)$$

The approximation error in the Fourier domain can be decomposed into two components.

$$S(\omega) - (\hat{P_h} s)(\omega) = e_1(\omega) + e_2(\omega) \quad (9)$$

$$e_1(\omega) = [1 - \frac{|\Phi(h\omega)|^2}{a_\varphi(h\omega)}] S(\omega) \quad (10)$$

$$e_2(\omega) = -\sum_{k\neq 0} \frac{\Phi(h\omega)}{a_\varphi(h\omega)} \cdot \overline{\Phi(h\omega + 2\pi k)} \cdot S(\omega + \frac{2\pi}{h}k) \quad (11)$$

where $e_1(\omega)$ and $e_2(\omega)$ denotes the in-band error and out-of-band error, respectively. The L_2 error bound of the LS approximation is described as following. If $\Phi^{(m)}(2\pi k) = 0, \quad k \in Z, \quad k \neq 0$ and $m = 0, \cdots, L-1$, then:

$$\forall s \in W_2^{2L}, \|s - P_h s\|_2 \leq K_{\varphi,2L} \cdot h^{2L} \cdot \|s^{(2L)}\|_2 + K_{\varphi,2L}^{1/2} \cdot h^L \cdot \|S^{(L)}\|_2 \quad (12)$$

$$K_{\varphi,2L} = \frac{1}{(2L)!} \cdot \frac{1}{A} \cdot \sup_{\xi} |\sum_{k\neq 0} (|\Phi|^2)^{(2L)}(\xi + 2\pi k)| \quad (13)$$

where $A = inf_\omega[a_\varphi(\omega)]$ is the lower Riesz bound.

2.3. Comparision

The error bound for the LS approximation consists of two distinct terms that represent the in-band (e_1) and out-of-band (e_2) contribution of the error, respectively. For small values of h, the first part of the error becomes negligible and the bound is dominated by the second $O(h^L)$ term. The corresponding constant $K_{\varphi,2L}^{1/2}$ turns out to be smaller than the constant $C_{\varphi,L}$ in Eq. (6). This is a first indication that there is an advantage in using least squares over interpolation. In addition, we note that in the case of larger values of h, the first term in the LS error bounds becomes dominant and it has the characteristic form of the error of an interpolator of order $2L$. We further observe that the function $\Phi_{2L}(\omega) = \frac{|\Phi(\omega)|^2}{a_\varphi(\omega)}$ represents the frequency response of an interpolator of order $2L$, Therefore under the condition that $\|e_1\| \gg \|e_2\|$, the LS solution of order L should perform as well as the corresponding interpolator with twice the order. This condition typically arises for larger h when the signal is somewhat undersampled. In the following section, we will apply this observation to a specific task where the observed discrete signal is undersampled.

3. DYNAMIC ARTICULATORY PARAMETER ESTIMATION

3.1. Review of Speech Production Models

The conventional source-filter speech production model assumes that speech is generated when sound sources excite a time-varying linear filter. The wave propagation inside the vocal tract is considered as a one-dimensional plane wave travelling through a concatenation of lossless acoustic tubes. This assumption leads to the linear predictive coding (LPC) model which is succefully used in most speech processing scenarios. However, researchers have revealed that the dynamics of speech airflow will create various degrees of turbulence during the production of speech sounds by the human vocal apparatus. This leads to articulatory speech production model which is based on the fundamental physics of the human vocal appratus. In this model, the human vocal tract is treated as a three-dimensional geometry described by several physical parameters.

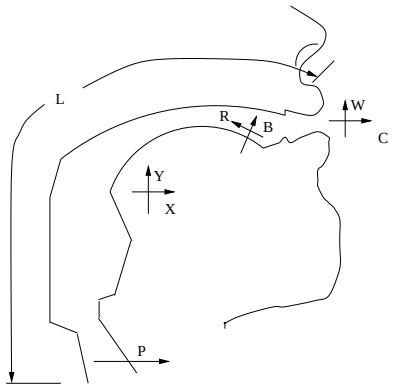

Figure 1. Coker's articulatory model.

Fig. 1 shows an articulatory model proposed by Coker in [4]. In this model, the midsaggital section of the human vocal tract was described by eight physical parameters : tongue body height (Y); anterior/posterior position of the tongue body (X); pharyngeal opening (P); tongue tip height (B); tongue tip curl (R); lip opening (W); lip protrusion (C); and vocal tract length (L). The production of

speech can be described as a combinative process that uses a limited set of units to produce a very large number of linguistically well-formed utterances. Segmental speech units such as phonemes are usually seen as discrete, static and invariant across a variety of contexts. The static articulatory parameters for different phonemes can be estimated based on minimizing the difference between the natural speech spectra and the model spectra computed from the Webster equation [5]. However, the shape of the vocal tract changes continuously during the speech production. These changes in shape are produced by the movements of a number of relatively independent articulators such as velum, tongue, lips and jaw. Once the dynamic articulatory parameters are estimated using certain approximation technique such as interpolation or LS approach, we can convert them into three-dimensional vocal tract shape and use it as the boundary condition for the solution of Navier-Stokes equations.

3.2. Estimation of Dynamic Articulatory Parameters

Suppose we are given N discrete phonemes with their static articulatory parameters denoted by $s_m(n), \quad m = 1, 2, \cdots M, \quad n = 1, 2, \cdots, N$, where M denotes the number of articulatory parameters per phoneme ($M = 8$ in our case). Our objective is to estimate the dynamic articulatory parameters between consecutive phonemes. This problem can be restated: given N discrete observations $s_m(n)$ of M continuous functions $s_m(x)$, we want to construct the functions $\hat{s}_m(x)$ which are an "optimal" approximation to $s_m(x)$ according to a certain criterion. Due to the physical constraints of the human vocal apparatus, all of its parameters such as tongue body height have finite values and their L^{th} derivatives also have finite value. Thus the functions $s_m(x) \in W_2^L$ and we can use the interpolation or LS approximation discussed above to solve the dynamic parameter estimation problem. In this paper, we investigated the cubic spline interpolation and the LS approximation discussed in previous section.

The LS approximation in Eq. (2) can be implemented by first prefiltering the discrete data $s_m(n) = s_m(x)|_{x=nh}$ by an "optimal" filter $P(z)$, which yields the sequence $c(k)$ in Eq. (3), and then by applying the interpolation in Eq. (1).

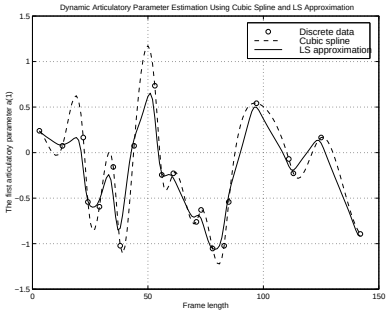

Figure 2. First articulatory parameter trajectory.

Figures (2) - (3) shows the estimated dynamic articulatory parameters (the first and the second articulatory parameter) of English sentence "Where were you while we were away" using the approximation methods discussed above. For each figure, the circle points denotes the static articulatory parameters computed from the method in [5]. For example, the first circle in Fig. (2) denotes the first articulatory parameter (anterior/posterior position of the tongue body) of the first phoneme /w/; the dashed line denotes the approximated continuous articulatory parameters using the cubic spline interpolation method; and the solid line denotes the estimated continuous articulatory parameters using the LS approximation approach. We can see from these figures the LS approach achieve smoother approximation of the articulatory parameter trajectory over time and don't have the "overshooting" problems of cubic spline methods. One hypothesis in human speech production is the "minumum jerk" hypothesis, which says that speech movements are planned to avoid sudden force changes [3]. We can see from these figures that the LS solution tends to be consistent with this hypothesis because its trajectory doesn't have the abrupt changes which exist in some regions of the cubic spline solutions.

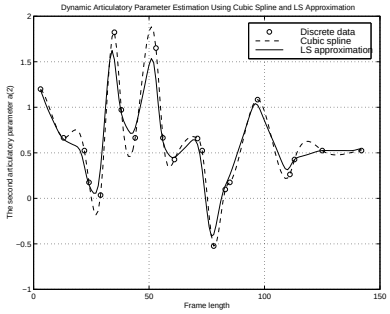

Figure 3. Second articulatory parameter trajectory.

4. EXPERIMENTAL RESULTS

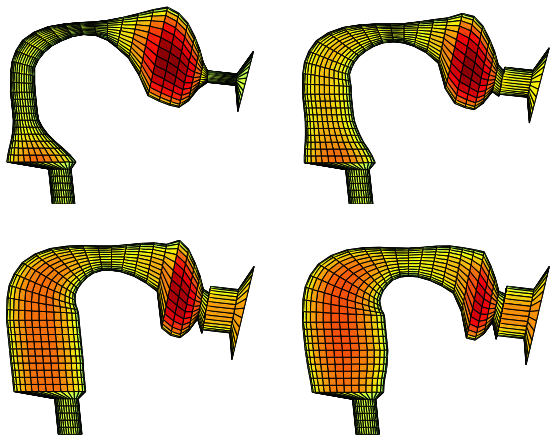

Figure 4. Snapshots of the vocal tract movements of phoneme /EH/ in /W/-/EH/-/R/.

Fig. 4 shows snapshots of the estimated vocal tract movements of phoneme /EH/ in /W/-/EH/-/R/ ("were") using the LS approximation technique described above. We can see from this figure that the LS approximation method can successfully estimate the articulatory trajectory and thus the dynamic vocal tract movements. Although the vocal tract shape of the phoneme /EH/ is relatively static when produced in isolation, we can see from this figure that the vocal tract geometry of /EH/ will change dynamically with respect to its neighbor phonemes, resulting in the acoustic changes which are called *coarticulation* in phonetics.

In our articulatory speech synthesizer, sound propagation inside the vocal apparatus is treated as a three-dimensional non-plane wave propagation inside a viscous fluid described by Navier-Stokes equations. In our synthesizer, the excitation signal was generated by specifying a time varying area at the inlet of the pharyngeal cavity and a sawtooth waveform was used to represent the area variation in time. A K-epsilon turbulence model was used to represent the effect of turbulence. Two different initial conditions were applied at the inlet of the vocal tract. One is based on specifying the volume velocity and the other is based on specifying the partical velocity, respectively. We converted the estimated articulatory trajectory into athree-dimensional vocal tract shape and use it as the boundary conditions for the solution of slightly compressible Reynold-Averaged Navier-Stokes (RANS) equations which can be written in the tensor expression :

$$M^2 \frac{\partial p}{\partial t} + \frac{\partial \nu_i}{\partial x_i} = 0 \tag{14}$$

$$\frac{\partial \nu_i}{\partial t} + \frac{\partial \nu_i \nu_j}{\partial x_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}[\frac{(\gamma + \gamma_t)}{R_e}\frac{\partial \nu_i}{\partial x_j}] \tag{15}$$

where M is the Mach number, ν_i denotes the mean velocity components, p denotes the static pressure, γ denotes the kinematic viscosity, γ_t is the turbulent eddy viscosity coefficient and R_e is the Reynold number.

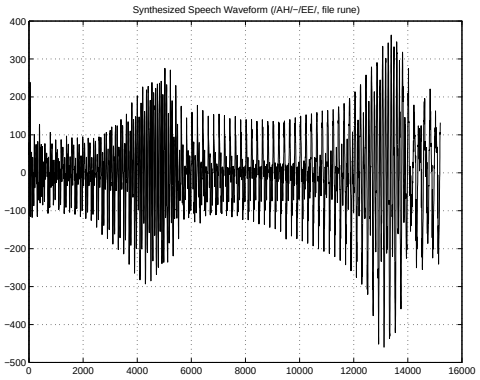

Figure 5. Waveform of the synthesized dipthong /AY/ ("uy" in buy).

Fig. 5 shows the waveform of the synthesized dipthong /AY/ ("uy" in buy).

5. CONCLUSION

It's clear that more knowledge about the speech production process might lead to improved speech synthesis and other speech processing systems. Signal approximation in the framework of Hilbert space [6] provides a powerful tool for articulator motor control because it is more efficient to implement in high dimensions and has more quantitative error bounds than the conventional interpolation technique based on Shannon's sampling theory. The LS approximation approach has a smaller error bound than the interpolation approach, especially in the case of under-sampling. Experimental results showed that the LS approach can successfully estimate the dynamic articulatory trajectory. Navier-Stokes equations describe fluid flow inside human vocal apparatus based on first principles of physics and has the potential to synthesize highly intelligible, natural sounding speech signal. The future work includes construction of the articulatory model based on magnetic resonance imaging (MRI) data and the synthesis of sentences.

ACKNOWLEDGE

This work is supported by a Motorola Partnerships in Research grant. The authors would like to thank Don Davis and Scott Slimon in the General Dynamics Electric Boat Corporation for providing the Navier-Stokes solver, and the NCSA center at University of Illinois for providing the Cray Origin 2000 supercomputer.

REFERENCES

[1] E.L. Saltzman and K.G. Munfall, "A dynamical approach to gestural patterning in speech production," *Haskind Laboratories Status Rerport on Speech production*, SR-99/100, pp. 38-68, 1989.

[2] N. Chomsky and M. Halle, *The Sound Pattern of English*, New York: Harper & Row, 1968.

[3] W.L. Nelson, "Physical principles for economies of skilled movements," *Journal of Biological Cybernetics*, Vol. 46, pp. 135 - 147, 1983.

[4] C. H. Coker, "A model of articulatory dynamics and control," *Proceedings of the IEEE*, Vol. 64, No. 4, pp. 452 - 460, Apr. 1976.

[5] S. E. Levinson and C. E. Schmidt, "Adaptive computation of articulatory parameters from the speech signal," *J. Acoust. Soc. Am.*, 74(4), pp. 1145 - 1154, Oct. 1983.

[6] M. Unser and I. Daubechies, "On the approximation power of convolution-based least squares versus interpolation," *IEEE Trans. on Signal Processing*, Vol. 45, No. 7, pp. 1697 - 1711, July 1997.

[7] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques: Part I – interpolators and projectors," *IEEE Trans. on Signal Processing*, Vol. 47, No. 10, pp. 2783 - 2795, October 1999.