Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 1 (4): 350-361, 2006
ISSN 1819-3579

Transformation of Spoken to Written Forms of (Natural) Languages via Spectral and Pseudo-spectral Methods

[1]Munir A. Chaudhuri and [2]Aadel A. Chaudhuri
[1]University Park Campus, University of Southern California,
Los Angeles, CA 90089
[2]Stanford University School of Medicine, Stanford, California, 94305

Abstract: The concept of wave function is employed to mathematically model the the transformation of a natural language from its spoken to both written forms–syllable (symbol) based and alphabet based writings. The fundamental steps to solving this problem are concerned with the solution to the wave equation subjected to initial and boundary conditions and derivation of these solutions via Fourier series and Sinc series expansion techniques for symbol and alphabet-based writing, respectively, which, in turn, through ideal and physical sampling, reduce to discrete time (space) Fourier series (DTFT or DSFT) and discrete Fourier transform (DFT).

Key words: Syllable, alphabet, written language, spectral methods, pseudo-spectral methods, sinc function, ideal sampling, physical sampling, reverse cantor set, rational delta function

Introduction

The writing systems of early civilizations, such as the Egyptian and the Hittite, were based on the use of symbols (hieroglyphs), representing a syllable. Egyptian hieroglyphics eventually included about 700 symbols (Anonymous, 1991). Realizing the difficulty encountered by the Hieroglyphic (or Hieratic, a cursive form of the Hieroglyphic) system in rapid processing of documents relating to business transactions, Phoenicians, who were primarily merchants, were the first to introduce the alphabet system of writing between 1700 and 1500 BC. While Phoenicians introduced twenty-two consonants in their writing system, Greeks complemented them with four vowels between 800 and 700 BC. A detailed discussion on this topic is available in Encyclopedia Britannica (1993).

Language evolution modeling has attracted increased attention from linguists, cognitive scientists, neurobiologists, mathematicians and computer scientists in the twentieth century. For example, Nowak and Komarova (2003) have remarked that language is a biological trait that radically changed the performance of one species and the appearance of the planet. Understanding how human language came about is one of the most interesting tasks for evolutionary biology. These authors have discussed how natural selection can guide the emergence of some basic features of human language, including arbitrary signs, words, syntactic communication and grammar. They further demonstrate how natural selection can lead to the duality of patterning of human language: sequences of phonemes forming words, sequences of which, in turn, forming sentences. Cangelosi (2001) has discussed different types of models for the evolution of communication and language. In particular, this study shows how evolutionary computation techniques, such as Artificial Life, can be employed to investigate the

Corresponding Author: Munir A. Chaudhuri, University Park Campus, University of Southern California, Los Angeles, CA 90089 Tel: (801) 243-4477

emergence of syntax and symbols from simple communication signals. Additionally, computational models of syntax acquisition and evolution is also discussed in the researcch.

Redford *et al*. (2001) have developed a computational model of emergent syllable systems, based on a set of functional constraints on syllable systems and the assumption that language structure emerged through a cumulative change over time. The constraints have been derived from general communicative factors as well as from the phonetic principles of perceptual distinctiveness and articulatory ease. Their model has generated mock vocabularies optimized for the given constraints. More information on this class of papers is available in the Language Evolution Modeling newsletter (Zuidana, 2001).

According to Chomsky and Halle (1968), the phonology, the description of sound change, in a spoken language system can be compared with links in a chain. The chain can be smoothed out as one continuous piece. The effects of this phonetic system would create some low-pass filtering of signal variables and represent them as time functions. The filter of the time functions would be dependent on the language or speaker. The C/D model due to Fujimura (1992) depicts how a mixture of symbols and numbers can be assessed within the phonetic implementation process. More importantly, Fujimura's (1992) C/D model shows "where and how the continuous nature of phonetic variables" can result from "phonologically significant articulatory variables" such as tongue movement. Many papers on experimental phonetics for linguistics, the phonetics-phonology interface, speech recognition, digital signal processing of speech and aero-acoustic modeling of speech are available on the web and will not be discussed further here in the interest of brevity (Fujimura, 2003; Anonymous, 2003).

The above review of the linguistic literature suggests that although substantial progress has been reported on the topic of empirical modeling of language and speech, a rigorous set theoretic basis for such modeling does not exist. The primary objective of the present study is to bridge this long-standing mathematical gap. To the authors' knowledge, the afore-mentioned linguistic literature reveals an absence of advanced mathematical techniques for mathematical representation of syllable- and alphabet-based writing, which will form the primary focus of the present research. In this study, the syllable is approximated to be ideally sampled version of morpheme (represented by a symbol in the case of Hieroglyphics), while the phonetic alphabet is the physically sampled version of the corresponding phoneme. In regards to the relationship between morphemes and syllables, the above assumption is justified on the ground that in ancient Egyptians' writings, morphemes were represented as syllables and vice versa. The assumption pertaining to the relationship between phonetic alphabets and phonemes is justified on the ground of ancient Phoenicians' pioneering usage to the same effect. The classical language prevalent in ancient Northern India, Sanskrit and its derivatives, Hindi, Bengali, Marathi, Punjabi, etc., also constitute a good example, where this assumption is justified. In Sanskrit based languages, the letters are categorized according to the primary organ giving rise to the phoneme. For example, the Sanskritic counterparts for /k/, /kh/, /g/, /gh/ and throaty /ng/ are called throat letters (kantha varna), softer (like French) versions of /t/, /th/, /d/, /dh/ and /n/ are called dental letters (danta varna), while /p/, /ph/, /b/, /bh/ and /m/ are called lip letters (oshtha varna) and so on. The notation '/ /' denotes a phoneme. In many (probably most) languages, there are more phonetic alphabets than written alphabets or letters, a...z. For example, in English 26 letters are used to represent 40 phonetic alphabets (i.e., phonemes). The fundamental steps to solving this problem are concerned with the (i) concept of wave function as a solution to the wave equation subjected to initial and boundary conditions for phonetic/linguistic modeling and (ii) derivation of these solutions via Fourier series and Sinc series expansion techniques. A rigorous set theoretic basis for mathematical modeling of a generic alphabet based written language is provided through application of the reverse Cantor set based rational deltafunction, first introduced by Chaudhuri (2005). Details are given in Appendix A1-A3.

Appendix: Reverse Cantor Based Rational Delta Function

A1: Construction of a Rational delta Function Using the Reverse Cantor Set

The Reverse Cantor set yields a sequence of delta (δ) functions, called δ-sequences, which are defined on a set of rational numbers (e.g., fraction p/q where p and q are integers) and no irrational or imaginary numbers. Thus, the process of starting from a unit interval [-1/2, 1/2], removing the two end-thirds in subsequent operations or iterations (n) and taking the limit as $n\to\infty$, yields a rational-type δ-function, $\lim_{n\to\infty} \delta_n(x)$. Furthermore, denoting the interval $[-(1/2)3^{-n}, (1/2)3^{-n}]$ to be I_n, $\delta_n(x)$ is defined as follows:

$$\delta_n(x) = 3^n \text{ for } x \in I_n(x) \tag{A1a}$$

$$\delta_n(x) = 0 \text{ for } x1 \notin I_n(x) \tag{A1b}$$

The reverse Cantor set discussed above can be graphed as shown in Fig. 2. Graphically, $\delta_n(x)$, which physically represents the density of each line segment, is represented by the spike protruding upwards from the middle of the x axis (x = 0).

A delta function is a representation of the orthonormality of the eigenfunctions, which comprise the possible solutions to the wave equation in a particular situation. For example, the Dirac delta function can be graphically represented by a single line stemming from the origin (x = 0) of the x-axis and approaching infinity. The physical representation of this type of delta function is a solution which is unbounded (hence the height of the graph approaches infinity) and an array of possible wave function eigensolutions which theoretically span the entire real (number) line. This is represented by the width of the graph (which approaches zero), thus showing that the possible eigenvalues in this situation are continuous, with solutions having little spacing between each other. Therefore, the Dirac delta function is a representation of the orthonormality of eigenfunctions in the case in which the wave function does not decay to zero within a finite region (Schiff, 1968).

The Dirac delta function is obviously a limiting case for representing the orthonormality of eigenfunctions associated with a continuous spectrum. The other extreme situation occurs in the case of a localized wave function where the corresponding eigenvalues are discrete. The orthonormality of eigenfunctions representing the solution of the wave equation in this case, involves the Kronecker delta function. The graph of this function is a line of unit length (support) that attains a unit height within the domain of definition, $I_0(x)$ and zero outside. The meaning for the graph is that possible eigensolutions in this case must have integers as indices and the wave does not extend beyond a finite region (Schiff, 1968) for a more detailed explanation of the Kronecker delta function).

A2: Reverse Cantor Set Based Rational Delta Function for Alphabetization of a Natural Language

While the supports of δ functions in the two extreme cases (i.e., Kronecker and Dirac) are either the unit interval or zero, their reverse Cantor set counterparts are all fractions and thus bridge the gap between these two extremes. This phenomenon leads to the possibility that the reverse Cantor set is a candidate for representing the orthonormality of eigenfunctions for more complicated situations where the wave functions of two or more alphabets interact. Upon hypothesizing that this is indeed the case, we must find a numerical method that will allow us to use the reverse Cantor set to obtain solutions to alphabetization problem.

The reverse Cantor set δ-sequence is a representation of the orthonormality of eigenfunctions of the wave equation for the wave functions of the alphabetized form of a natural language. For a syllable based written language with a discrete frequency spectrum, n is small and the reverse Cantor δ function closely resembles the Kronecker δ function. On the other extreme, for a spoken language with a continuous frequency spectrum, the wave function decays to zero very slowly and the reverse Cantor set with a large n-value can be used as a representation of the orthonormality of the eigenfunctions. The orthonormality of eigenfunctions comprising the wave function for a alphabetized form of written language can be best represented using the reverse Cantor set $\delta_n(x)$ with an intermediate value of n. The value of n can be adjusted depending on the type of alphabets (e.g., Hebrew, Latin, Greek, Arabic, Sanskrit, English, French, German, etc.) used to write a particular natural language.

A3: Method for Normalization of the Eigenfunctions for Alphabetization Based on Reverse Cantor Set Delta Function

The wave function, Ψ(x), which is solution to the wave equation, can be expanded in the reverse Cantor set based rational delta function, $\delta_n(x)$ as follows:

$$\psi(x) = \int_{Rat,n} \psi(x')\delta_n(x - x')dx', \tag{A2}$$

where, $\int_{Rat,n}$ denotes a pseudo-integral on rationals. The eigenfunction whose continuous analog given in the form

$$u_k(x) = \overline{C} \exp(ikx), \tag{A3}$$

satisfies the orthonormality relations both in the position as well as momentum (propagator) space with respect to the reverse Cantor set-based rational delta function, $\delta_n(x)$, as follows:

$$\int_{Rat,n} u_k^*(x')u_k(x)dk = \delta_n(x - x'), \tag{A4a}$$

$$\int_{Rat,n} u_{k'}^*(x)u_k(x)dx = \delta_n(k - k'). \tag{A4b}$$

Substitution of Eq. A4a into Eq. A2 yields

$$\psi(x) = \int_{Rat,n} \psi(x') \int_{Rat,n} u_k^*(x')\ u_k(x)dk\ dx' = \int_{Rat,n} A_k^d u \to (x)dk, \tag{A5}$$

where, the Discrete Fourier Transform (DFT), A_k^d, because of Eq. A 4b is given by

$$A_k^d = \int_{Rat,n} u_k^*(x')\ \psi(x')\ dx'. \tag{A6}$$

Since the orthonormality relations, given by Eq. A4a and A4b employ the reverse Cantor set based rational delta function, $\delta_n(x)$ instead of $\delta(x)$, Eq. A5 and A6 must be modified as follows. Identifying $I_n(x) = \Delta_n$, n = 0, 1, 2,... the sampling interval on the x-axis (one-dimensional case), the wave function ψ, given by Eq. A2, can be expressed in the form of discrete Fourier transform as given below:

$$A_k^d = \int_{Rat,\,n} u_k^*(x')\ \psi(x')\ dx' = \int_{Rat,\,n} \psi(x')\ \exp\{2\pi i\nu x'/(n\Delta_n)\}dx', \tag{A7}$$

where

$$k = 2\pi\nu/(n\Delta_n), \tag{A8}$$

with n being the total number of sampling points and $\nu \in Z$ the set of integers. It then immediately follows (Press *et al.*, 1992) that the discrete Fourier transform, A_k^d, of the wave function, $\psi(x)$, can be further approximated to its series-discretized form as follows:

$$A_k^d = \int_{Rat,\,n}^{\psi}(x) \exp\{2\pi i\nu x/(n\Delta n)dx = \Delta n \sum_{j=0}^{n-1} \Psi_j \exp\ (2\pi i\nu j/n) = \Delta n\ \overline{A}_k^d\} \tag{A9}$$

in which

$$\overline{A}_k^d = \sum_{j=0}^{n-1} \psi_j \exp\ (2\pi i\nu j/n) \tag{A10}$$

The inverse discrete Fourier transform, which recovers the sampled $\psi_j = \psi(x)$, from discrete Fourier transform, A_k^d, is, in the one-dimensional case, given by

$$\psi_j = \frac{1}{n}\sum_{j=0}^{n-1} \overline{A}_k^d \exp(-2\pi i\nu j/n) \tag{A11}$$

In what follows, the concept of wave function is employed to mathematically model the transformation of a natural language from its spoken to both written forms–syllable or symbol based written language, e.g., Hieroglyphic (or Hieratic) and alphabet based written language. The fundamental steps to solving this problem are concerned with the solution to the wave equation subjected to initial and boundary conditions and derivation of these solutions for syllables and phonemes via Fourier series and Sinc series expansion techniques for symbol and alphabet-based writings, respectively. For example, the name, Tutankhamon (also known as Tutankhamen) of the boy king, who ruled Egypt from 1333 to 1323 BC, is comprised of three symbols in the Hieroglyphic (or Hieratic) system–Tut,

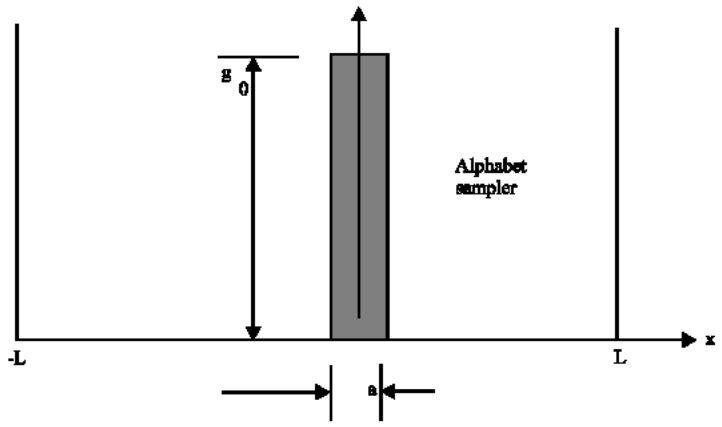

Fig. 1: Alphabet sampler inside a morphemic well

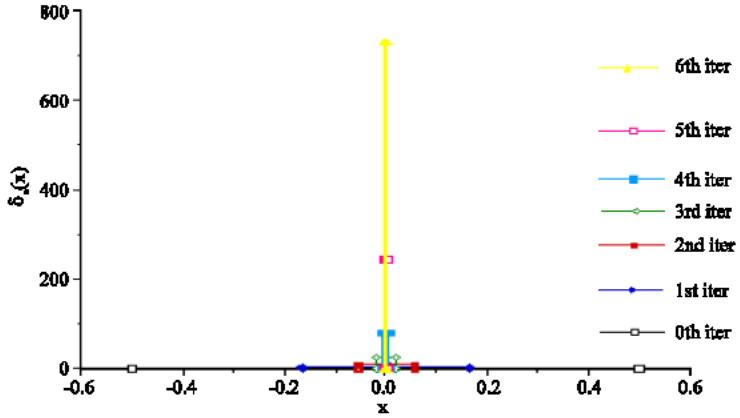

Fig. 2: A graphical representation of the reverse Cantor set based delta function (sequence) $\delta_n(x)$ vs. x

Ankh and Amon. The syllable based written language can be represented by a Fourier series, which has a discrete frequency spectrum. In the other extreme, a spoken language can be represented by a Fourier transform, characterized by a continuous frequency spectrum. The alphabetical form of the written language lies somewhere in between, characterized by a frequency spectrum neither entirely discrete nor completely continuous. This will be represented by a Sinc series. The Fourier and Sinc series, in turn, through ideal and physical sampling, respectively, reduce to discrete time (space) Fourier series (DTFT or DSFT) for symbol based writing and discrete Fourier transform (DFT) for alphabet based writing.

Materials and Methods

Wave Equation for Alphabetization and Syllabication of a Natural Language

Since speech is generated by the vibration of the human vocal chord, movement of tongue and its touching various parts of the mouth and involvement of lips and nose and the generated sound propagates in the air in the form of a wave with velocity c, it is logical to start with the wave equation of the form:

$$\frac{1}{\omega^2}\frac{\partial^2 u}{\partial t^2} = \frac{1}{k^2}\frac{\partial^2 u}{\partial x^2}, \tag{1}$$

in which u is the displacement, t is time and x is spatial coordinate. The angular frequency, $\omega = 2\pi f$, f being the frequency, while the wave number, $k = 2\pi/\lambda$, λ being the wavelength. The wave velocity is given by

$$c = \frac{\omega}{k}, \tag{2}$$

The initial condition is given by

$$u(x,0) = g(x); \quad \frac{\partial u}{\partial t}(x,0) = 0. \tag{3}$$

In addition, the boundary conditions, which depend on the form of a written language, must be specified.

On applying the separation of variables,

$$u(x,t) = \phi(t)\,\psi(x), \tag{4}$$

Eq. 1 yields the following:

$$\frac{1}{\omega^2\phi(t)}\frac{\partial^2 \phi}{\partial t^2} = \frac{1}{k^2\psi(x)}\frac{\partial^2 \psi}{\partial x^2} = C, \tag{5}$$

in which C is a constant. In order to obtain the desired solution in the form of $e^{i(kx - \omega t)}$ and $e^{-i(kx - \omega t)}$ (or equivalently sine and cosine), C must be negative, which is taken to be equal to -1, without any loss of generality. This operation yields the following two ordinary differential equations (ODE) in x and t, respectively:

$$\frac{\partial^2 \psi}{\partial x^2} + k^2\psi = 0, \tag{6}$$

$$\frac{\partial^2 \phi}{\partial t^2} + \omega^2\phi = 0. \tag{7}$$

It may be noted that the separated wave equation in x, given by Eq. 6, is consistent with the underlying hypothesis of the present research that a spoken natural language (speech, word form) can be mathematically represented in the syllabified and alphabetized written forms of the same. This justifies the concept of the spatial wave function, $\Psi(x)$, wherein x is given in terms of phoneme/alphabet units. The wave number k represents the corresponding eigenvalue.

Syllabic Form of Writing

In the syllabic form of writing, a symbol (representing a syllable) is abruptly terminated before a new syllable starts. This is idealized by one dimensional well (in analogy to quantum mechanics). The boundary conditions are specified at the start (x = 0) and end (x = L) of the syllable, where L is given in terms of phoneme units. These are given by

$$\psi(0) = \psi(L) = 0. \tag{8}$$

The solution to the ODE (6) subjected to the boundary condition (8) can then be easily obtained as follows:

$$k_n = \frac{n\pi}{L}; \quad \lambda_n = \frac{2L}{n}; \tag{9}$$

$$\psi(x) = A_n \sin\left(\frac{n\pi x}{L}\right), \tag{10}$$

which is the solution for a standing wave. The solution to the ODE (7) can also be easily obtained as follows:

$$\phi(t) = B_n \cos(\omega_n t) + \overline{B}_n \sin(\omega_n t). \tag{11}$$

The complete solution is given by

$$u(x,t) = \psi(x)\,\phi(t) = \sum_{n=1}^{\infty}\left\{D_n \cos(\omega_n t) + \overline{D}_n \sin(\omega_n t)\right\} \sin\left(\frac{n\pi x}{L}\right), \tag{12}$$

in which

$$D_n = A_n B_n; \quad \overline{D}_n = A_n \overline{B}_n. \tag{13}$$

After applying the initial conditions, the constant coefficients can be obtained as follows:

$$\overline{D}_n = 0, \qquad \sum_{n=1}^{\infty} D_n \sin\left(\frac{n\pi x}{L}\right) = g(x). \tag{14}$$

D_n can easily be determined by expanding g(x) in the form of Fourier sine series in x. The syllabic form of the written language needs to capture the initial condition (14) of the spoken language. Here the computed wave function is localized within the box (well), while in a speech, the wave function is extended all the way to "infinity".

Phonemic Analysis of Speech Sounds and Alphabetic Form of Writing

Morpheme is the smallest part of a word that has meaning of its own. Morphemes may be words, prefixes, suffixes or endings that show inflection (Barnhart and Barnhart, 1991). In the word carelessness, the morphemes are care, -less and -ness. A morpheme, consisting of alphabets, "does not necessarily consist of phonemes, but all morphemes are statable in terms of phonemes" (H.A. Gleason,

Jr., quoted by Barnhart and Barnhart, 1991). In the present work, morpheme/syllable is assumed to be comprised of a number of phonemes (approximated by phonetic alphabets). A phoneme acts as a low-pass filter in the frequency domain. It follows that the corresponding alphabet acts as a long-wave filter or low wave number (k) filter/sampler in the wave number domain.

The characteristics of an ideal low-pass filter, in the frequency and time domains is discussed in standard texts e.g., Gajic (2003). An ideal low-pass filter transfer function is given by Gajic (2003).

$$H(i\omega) = \begin{cases} e^{-i\omega t_d}, & |\omega| \le \omega_0 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

The phase of the ideal filter is assumed to change linearly in frequency, which corresponds to the time shift of the filter input signals by t_d, known as the time delay. The impulse response of the ideal low-pass filter is obtained as follows. Using the time domain Fourier equivalent of a rectangular frequency domain pulse of unit height given by the Fourier transform pair

$$p_{2\omega_0}(\omega) = \begin{cases} 1, & |\omega| \le \omega_0, \\ 0, & \text{otherwise} \end{cases} \leftrightarrow \frac{2\omega_0}{2\pi} \sin c\left(\frac{2\omega_0}{2\pi} t\right), \tag{16}$$

the ideal low-pass filter transfer function is given by Gajic (2003)

$$H(i\omega) = p\omega_0(\omega)\exp^{(-i\omega t_d)} \leftrightarrow \frac{\omega_0}{\pi} \sin c\left(\frac{\omega_0}{\pi}(t - t_d)\right) = h(t), \tag{17}$$

in which the time shift property of the Fourier transform has been utilized and the Sinc function is defined as:

$$\sin c(t) = \frac{\sin(\pi t)}{\pi t}. \tag{18}$$

The characteristics of the same ideal low-pass filter, in the wave number and space domains can be described in a similar manner. An ideal low-pass filter space transfer function can be written as follows:

$$G(ik) = \begin{cases} g_0\ e^{-ikx_d}, & |k| \le k_0 \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

The phase of the ideal filter is again assumed to change linearly in wave number, which corresponds to the space shift of the filter input signals by x_d, to be termed here as the space shift (delay). The spatial impulse response of the ideal low-pass filter is then obtained as follows. Using the space domain Fourier equivalent of a rectangular wave number domain pulse of height, g_0, given by the Fourier transform pair

$$P_{2k0}(k) = \begin{cases} g_0, & |k| \le k_0, \\ 0, & \text{otherwise} \end{cases} \leftrightarrow g_0 \frac{2k_0}{2\pi} \sin c\left(\frac{2k_0}{2\pi} x\right), \tag{20}$$

the ideal low-pass filter spatial transfer function can now be written follows:

$$G(ik) = P_{k0}(k)e^{-ikx_d} \leftrightarrow g_0 \frac{k_0}{\pi} \sin c\left(\frac{k_0}{\pi}(x - x_d)\right) = g(x), \tag{21}$$

in which the spatial shift property of the Fourier transform has been utilized and the Sinc function is defined as before. Generalizing the above, g(x) can be expressed in standard sinc series (Stenger, 1993; Lund and Bowers, 1992) as follows:

$$g(x) = \frac{1}{h}\sum_{j=-\infty}^{\infty} g(jh)\ \sin c\left(\frac{x - jh}{h}\right), \tag{22a}$$

$$g(jh) = \int_{-\infty}^{\infty} g(x)\sin c\left(\frac{x - jh}{h}\right) dx, \tag{22b}$$

in which

$$h = x_d = \frac{\pi}{k_0};$$

and $g_0 = g(0)$ corresponding to $j = 0$.

Sampling with an Ideal/Physical Sampler and the Discrete Time Fourier Transform (DTFT)/Discrete Fourier Transform (DFT)

Shannon's Sampling Theorem (Gajic, 2003) states that a continuous-time band-limited signal, such as the Sinc function, sinc(t), with bandwidth frequency f_{max} can be uniquely reconstructed from its sampled values at sinc (jT_s), $j = 0, \pm1, \pm2, \pm3,...$, if the sampling frequency $f_s = 1/T_s$ satisfies

$$f_s = 1/T_s \geq 2\ f_{max} \tag{23}$$

The frequency $2f_{max}$ is called the Nyquist frequency and the frequency interval $[-f_{max}, f_{max}]$ is called the Nyquist interval. This theorem can be easily extended to the space and wave number domains.

Standard texts on digital signal processing consider two types of samplers: ideal sampler and the physical sampler (Gajic, 2003). It is shown that an ideal sampler gives rise to discrete time Fourier transform (DTFT), while a real or physical sampler produces discrete Fourier transform (DFT). Applying ideal sampling operation to syllable (sound) based written language, such as Hieroglyphic or Hieratic, these written symbols can be mathematically represented by DSFT (discrete space Fourier transform), the spatial counterpart of the DTFT. This is expressed by the spatial counterpart of Eq. (9.16) of Gajic (2003), which suggests that the wave number spectrum of the ideal sampled "signal" (here the symbols) is periodic.

As has been commented by Gajic (2003), the DTFT or in our case DSFT is the Fourier transform of a discrete time or space "signal", useful for representing a syllable based written language and obtained by sampling a continuous-time or -space "signal" by an ideal sampler. It has been shown (Gajic, 2003) that DTFT and its inverse as defined in Eq. 9.23 and 9.27 of that book can be derived

using the discrete-time or -space Fourier series. By analogy to the continuous-time or -space Fourier series, the discrete-time or -space Fourier series are applicable to discrete-time or space periodic signals. A discrete-time or -space aperiodic signal can be considered in the limit as a discrete-time or -space periodic "signal" whose period tends to infinity and thus DTFT or DSFT tends to the corresponding DFT (discrete Fourier transform), discussed in standard treatises (Gajic, 2003; Press *et al*., 1992).

Application of the physical sampling to the phonemic signal analysis produces an alphabet-based written language. Alphabet serves as the physical or real sampler of narrow width, 'a'. The alphabet sampler inside the morphemic "well" model is shown in Fig. 1. Assuming the number of alphabets inside a morpheme to be N, the length of the syllable, L = Na, where a is the width of the phoneme/alphabet. The boundary conditions are given by Eq. (8) at x = ± L.
A rectangular pulse is defined as follows:

$$P_a(x) = \begin{cases} g_0, & |x| \le a/2; \\ 0, & \text{elsewhere} \end{cases} \tag{24}$$

Since in the present alphabetic physical sampler inside the morphemic box (well) problem, the wave function is permitted neither to extend all the way to "infinity", nor to be localized in length scale of the order of "sampler width", a (<< L), the computed eigenfunctions cannot be orthonormalized with respect to either the Dirac or Kronecker δ functions. Consequently, these eigenfunctions need to be orthonormalized with respect to the reverse Cantor set based rational δ function, discussed in the Appendix. Therefore, the correct solution must be of the form of discrete Fourier transform, as given by Eq. (A10) and (A11).

$$\psi_j = \frac{1}{n}\sum_{j=0}^{n-1} \bar{A}_k^d \exp(-2\pi i v j/n), \tag{25a}$$

$$A_k^d = \Delta_n \sum_{j=0}^{n-1} \psi_j \exp(2\pi i v j/n) = \Delta n\ \bar{A}_k^d, \tag{25b}$$

where Ψ_j is the physically sampled spatial wave function. A_k, with k being the eigenvalue, need to be numerically evaluated by using the DFT (or FFT, DWT, etc.) technique and must satisfy the various boundary conditions of the problem. This scheme is currently being implemented and numerical results will be reported in future.

The fast Fourier transform (FFT) is an algorithm that computes in O (N $\log_2$N) operations as compared to the DFT's O (N^2). The discrete wavelet transform (DWT) is a more recently developed fast computational tool that linearly operates on a data vector, the length of which is an integral power of 2 and transforms it into a numerically different vector without altering the length (Daubechies, 1992). Like the FFT, the DWT is invertible and orthogonal, its inverse transform, when viewed as a large matrix, being the transpose of the transform. Both the FFT and DWT can, therefore, be viewed as rotations in function space, from the input space to a new domain. In the case of FFT, the rotated domain has basis functions in the form of standard sines and cosines, while in the wavelet domain the basis functions are somewhat novel, with names like "mother functions" and "wavelets". The details of the FFT and DWT algorithms are available in Press *et al*. (1992) and will not be repeated here.

Conclusions

The concept of wave function is employed to mathematically model the transformation of a natural language from its spoken to both written forms – syllable (symbol) based and alphabet based writings. The fundamental steps to solving this problem are concerned with the solution to the wave equation subjected to initial and boundary conditions and derivation of these solutions for syllables and phonemes via Fourier series and Sinc series expansion techniques for symbol and alphabet-based writings, respectively. These, in turn, through ideal and physical sampling, reduce to discrete time (space) Fourier series (DTFT or DSFT) for symbol based writing and discrete Fourier transform (DFT) for alphabet based writing.

In this work, the syllable is approximated to be ideally sampled version of morpheme (represented by a symbol in the case of Hieroglyphics), while the phonetic alphabet is the physically sampled version of the corresponding phoneme. In regards to the relationship between morphemes (symbols) and syllables, the above assumption is justified on the ground that in ancient Egyptians' writings, morphemes were represented as syllables and vice versa. The assumption pertaining to the relationship between phonetic alphabets and phonemes is justified on the ground of ancient Phoenicians' pioneering usage to the same effect.

A rigorous set theoretic basis for mathematical modeling of a generic alphabet based written language is provided through application of the reverse Cantor set based rational delta function, first introduced by Chaudhuri (2006). Alphabetization bridges the gap between the two situations that arise in phonetic/linguistic research, namely the syllable based written language with discrete eigenvalues and the spoken language with a continuous spectrum of eigenvalues.

Most important, this novel rational delta function, $\delta_n(x)$, provides a rigorous set theoretic basis for permitting the resulting computed wave function to be expressed in the form of the discrete Fourier transform (DFT) in the Fourier domain and recover the sampled wave function in the physical domain by employing the inverse discrete Fourier transform (IDFT). The relatively straightforward transition to faster techniques, such as the fast Fourier transform (FFT), Sinc and discrete wavelet transform (DWT), will be the subject of future research.

This research helps bridge the long-standing gap between phonetic/linguistic research and modern development in computational spectral and pseudo-spectral methods, such as DFT, FFT, Sinc and DWT. The ease of the present method and its relative accuracy will help make challenging language problems numerically solvable.

Acknowledgement

The authors wish to thank Dr. J.A. Laursen and an anonymous reviewer for their helpful suggestions on an earlier version of the manuscript

References

Anonymous, 1991. World Book Encyclopedia, World Book, Inc., London.

Anonymous, 2003. Journal of the International Phonetic Association (JIPA), Cambridge University Press, Cambridge, UK.

Barnhart, C.L. and R.K. Barnhart, 1991. World Book Dictionary, World Book, Inc., London.

Cangelosi, A., 2001. Evolution of communication and language using signals, symbols and words. IEEE Transactions in Evolutionary Computation, pp: 93-101.

Chaudhuri, A.A., 2005. Construction of a rational delta function using the reverse Cantor set and its application to quantum mechanics via pseudo-spectral Methods. J. Applied. Sci. Res., pp: 352-361.

Chomsky, N. and M. Halle, 1968. Sound Pattern of English, Harper and Row, New York.

Daubechies, I., 1992. Wavelets, S.I.A.M. (Soc. Ind. Appl. Math.), Philadelphia.

Encyclopedia Britannica, Writing, Vol. 29, 15th Edn., Encyclopedia Britannica, Inc., London (1993).

Fujimura, O., 1992. Phonology and phonetics: A syllable-based model of articulatory organization. J. Accoust. Soc. Japan (E), pp: 39-48.

Fujimura, O., 2003. Syllable structure constraints: A C/D model perspective. Internet.

Gajic, Z., 2003. Linear Dynamic Systems and Signals, Prentice-Hall (Pearson Education, Inc.), Upper Saddle River, NJ.

Lund, J. and K.L. Bowers, 1992. Sinc Methods for Quadrature and Differential Equations, SIAM, Philadelphia.

Nowak, A.M. and N. L. Komarova, 2001. Towards an evolutionary theory of language. Trends Cogn. Sci., pp: 288-295.

Press, W.H., S.A. Teukolsky, W.T. Vettering and B.P. Flannery, 1992. Numerical Recipes in C, 2nd Edn., Cambridge University Press.

Redford, M.A., C.C. Chen and R. Miikkulainen, 2001. Constrained emergence of universals and variation in syllable systems. Lang. Speech, pp: 27-56.

Stenger, F., 1993. Numerical Methods Based on Sinc and Analytic Functions, Springer Verlag, New York.

Schiff, L.I., 1968. Quantum Mechanics, 3rd Edn., McGraw-Hill, New York.

Zuidana, W.H., 2001. Language evolution modeling: Abstracts of recent papers.