ESTIMATION OF THE PARAMETERS OF A LONG -TERM MODEL FOR ACCURATE REPRESENTATION OF VOICED SPEECH

Yoram Stettiner[1], David Malah[2] and Dan Chazan[3].

[1] Dept. of Electrical Engineering, Technion - Israel Institute for Technology, Technion City, Haifa 32000, Israel.
Also with Nexus Telecommunication Systems Ltd., Korazin 1, Givataim 53583, Israel

[2] Dept. of Electrical Engineering, Technion - Israel Institute for Technology, Technion City, Haifa 32000, Israel

[3] IBM Science and Technology Center, Technion City, Haifa 32000, Israel

ABSTRACT

The paper addresses the problem of estimating the parameters of a long-term model of voiced speech. The model is able to describe slow time-variation (non-stationarity) of speech and hence enables the analysis of a whole phoneme in a single frame. This is of great importance in the separation of close pitch harmonics common in speech separation problems. It also has potential in speech coding and synthesis applications. The model considered is an extension of the model proposed by Almeida and Tribolet [1]. Contrary to [1], which uses a Taylor series approximation and a fixed pitch in the analysis interval, this work presents an efficient iterative algorithm for explicit estimation of the model parameters - including the time-warping function which describes the pitch variation in the analysis frame.

1. INTRODUCTION

Most speech analysis-synthesis systems assume short term stationarity, and hence use relatively short analysis frames in which this assumption is not seriously violated. In some applications, however, using longer frames is beneficial. For example, in the co-channel voiced speech separation problem [10 and references therein], separability may improve with longer frame lengths, provided that the stationarity assumption holds. This is important when the speakers have close spectral harmonics.

With real speech, however, using too long frames degrades separability and quality, unless the non-stationarity is incorporated into the model. In a companion paper [10], we assume a quasi-periodic model which allows the pitch to vary linearly within the analysis frame. In a speech separation system based on the model, we were able to use 60 ms long analysis frames while keeping to a minimum the smearing of harmonics due to varying pitch. However, the time variation of the spectral envelope and the pitch deviations from a linear function, increasingly degraded the performance with frame lengths exceeding 60 ms. We concluded that for effective separation, we need a long term non-stationary model that accounts for variations both in the pitch and the spectral envelope, and can accurately describe the waveform of voiced speech, in which case it may also be useful in speech synthesis and coding.

Martinelli et al. [6] proposed a long term time-varying sinusoidal model, but avoided estimating the time-varying frequencies of the harmonics. Consequently, they resorted to equally spaced harmonics of a fixed pitch. The harmonic coders [2,7], and the prototype waveform interpolator [8], can accurately describe non-stationary voiced speech, but only for durations under 30 ms.

In this work we extend a non-stationary spectral model for voiced speech presented by Almeida and Tribolet [1], and propose an efficient estimation scheme of its parameters. Their original model consists of a linear time-varying filter excited by a time-warped periodic impulse train. Under some reasonable assumptions, this model becomes

$$s(t) = \sum_{k=-\infty}^{\infty} H(t, k\Omega(t))\, e^{jk\phi(t)}$$

$$x(t) = s(t) + \upsilon(t) \qquad (1)$$

where x is the voiced speech signal, s is the model output, υ is a Gaussian noise process representing both additive noise and modeling errors, H is the short time Fourier transform (STFT) of the time varying impulse response of the vocal tract, $\phi(t)$ is an invertible time-warping function,

$\Omega(t) \triangleq \dot{\phi}(t)$ is the fundamental instantaneous frequency, and k is the spectral harmonic index.

Almeida and Tribolet avoided the difficult problem of explicitly estimating the time-warping function by expanding the model into a low order Taylor series and assuming a fixed pitch for the whole analysis frame. Eventually they arrived at what is known as the harmonic coder [2,7], which unlike their original model (1), cannot describe voiced speech for the duration of full phonemes.

Our approach is to begin with (1) and develop an efficient iterative scheme for the explicit estimation of ϕ and H, assuming nothing that may compromise the ability of the model to describe long voiced phonemes. To this end, the model is decomposed so that a strictly periodic signal is time-warped and fed into a time-varying all-pole filter.

2. THE MODEL

We begin by decomposing (1) so that each voiced phoneme is modeled by a double transformation on a strictly periodic signal with period 2π (Fig. 1). The first transform is the inverse time-warp ϕ^{-1}, the output of which is quasi-periodic. The second transform, H_2, is a linear time-varying (all pole) system. The physical and warped time are t and $u = \phi(t)$, respectively. The strictly periodic signal $\tilde{s}(u)$ has the time-invariant Fourier coefficients c_k , where k is the harmonic index. The derivative of the time-warping function $\Omega(t) = \dot{\phi}(t)$ represents the instantaneous pitch. The spectral envelope is divided into a time-invariant and time-varying parts, represented by $\{c_k\}$ and H_2, respectively. The resulting time-varying generalized Fourier coefficients are $d_k(t)$. In Fig. 1, $g(t)$ is a scalar gain function representing the speech loudness.

Let $\tilde{H}_1$ be the spectral envelope of $\tilde{s}(u)$ so that

$$c_k \triangleq \tilde{H}_1(k \cdot 1 \ {}^{rad}\!/_{sec}) = H_1(t, k\Omega(t)) \qquad (2)$$

where H_1 a time varying transfer function, sampled at multiples of the instantaneous pitch. It is shown in [1] that H_1 is the warped version of $\tilde{H}_1$. Similarly, the time varying Fourier coefficients are

$$d_k(t) \triangleq \tilde{H}_2(\phi(t), k) = H_2(t, k\Omega(t)) \qquad (3)$$

In terms of these transfer functions, $s(t)$ is given by

$$s(t) = g(t) \sum_{k=-\infty}^{\infty} H_2(t, k\Omega(t))\, H_1(t, k\Omega(t)) e^{jk\phi(t)} \qquad (4)$$

Finally, the model is

$$s(t) = g(t) \sum_{k=-\infty}^{\infty} d_k(t)\, c_k\, e^{jk\phi(t)} \qquad (5)$$

In summary, the model has 3 vector parameters: c_k - Fourier coefficients of the periodic signal in the warped-time domain.; $\phi(t)$ - Invertible time-warping function; and d_k - Time-varying Fourier coefficients of the linear time-varying transfer function.

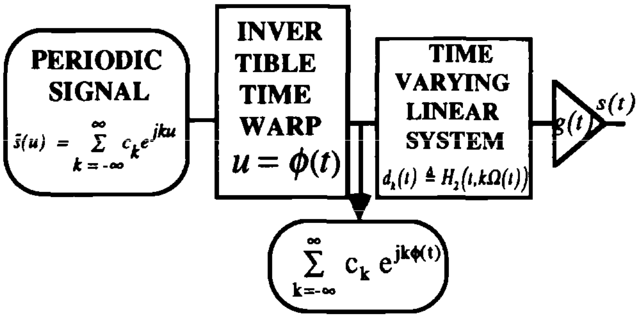

PERIODIC SIGNAL
$\tilde{s}(u) = \sum_{k=-\infty}^{\infty} c_k e^{jku}$
INVER TIBLE TIME WARP
$u = \phi(t)$
TIME VARYING LINEAR SYSTEM
$d_k(t) \triangleq H_2(t, k\Omega(t))$
$g(t)$
$s(t)$
$\sum_{k=-\infty}^{\infty} c_k e^{jk\phi(t)}$

Fig. 1 - Proposed long term model for voiced speech

Clearly, the model (5) is not unique. We will develop later a criterion for a unique selection of $\{c_k\}$ and $\{d_k\}$. The dimensions of the parameters become finite when we assume a finite analysis bandwidth and use discrete time.

3. ML PARAMETER ESTIMATION

With additive Gaussian noise (AGN), Maximum Likelihood (ML) estimation is equivalent to Weighted Least Squares (WLS). We pose the WLS problem using discrete time notation. Given a voiced speech signal $x(t)$ and defining $\underline{x}, \underline{s}, \underline{d}, \underline{\phi}$ and $\underline{g}$ as the discrete time versions of $x(t), s(t), d(t)$ and $g(t)$, respectively, in the analysis frame, the parameter vector is

$$\underline{\theta} \triangleq [\underline{\phi} \quad \underline{c} \quad \underline{d} \quad \underline{g}]^T \qquad (6)$$

we wish to solve

$$\underset{\underline{\theta}}{Min} \ (\underline{x} - \underline{s}(\underline{\theta}))^T Q^{-1} (\underline{x} - \underline{s}(\underline{\theta})) \qquad (7)$$

subject to

$$0 < a < \dot{\phi}(nT) < b \ , \quad n = 0, \dots, N-1 \qquad (7a)$$

where T is the (uniform) sampling period in the t domain, $[0,(N-1)T]$ is the analysis interval, d_k are slowly varying, g is strictly positive and slowly varying, and Q is the covariance matrix of the noise. Since (7) is a highly complex non-linear optimization problem, we propose a sub-optimal iterative approach. First the gain function g is estimated and the signal is gain normalized to have a constant loudness, thus reducing the variance of $\{d_k\}$. As noted earlier, the spectral envelope is determined by both $\{c_k\}$ and $\{d_k\}$. We will now show how to select $\{d_k\}$ such that the application of H_2^{-1} to the gain normalized signal will facilitate the estimation of $\{c_k\}$ and the warping function. Suppose we are given

$$x_2(t) = s_2(t) + v_2(t) \qquad (8)$$

where x_2 and υ_2 are the result of passing the input signal x and noise υ, respectively, through gain and spectral envelope normalization (i.e. after applying the inverses of g and H_2), and

$$s_2(t) \triangleq \sum_{k=-\infty}^{\infty} c_k e^{jk\phi(t)} \tag{9}$$

is the quasi-periodic signal model. Defining $\underline{x_2}$ as the discrete version of $x_2(t)$, and Q_2 as the covariance matrix of υ_2, the WLS estimation problem is

$$\underset{s_0,\phi}{Min}\ \left(\underline{x_2} - \underline{s_0}(\phi)\right)^T Q_2^{-1}\left(\underline{x_2} - \underline{s_0}(\phi)\right) \tag{10a}$$

subject to (7a) and

$$s_0(u) = s_0(u + 2\pi) \tag{10b}$$

The problem in (10) is a complex optimization problem. However, assuming for the moment that s_0 is known and Q_2 is diagonal (we will justify these assumptions later), (10) may be rewritten as

$$\underset{\phi}{Min} \sum_{n=0}^{N-1} \underbrace{w_1(nT)\left|x_2(nT) - s_0(\phi(nT))\right|^2}_{e^2(n,\phi(nT))} \tag{11}$$

where $w_1(nT) = (diag(Q_2^{-1}))_n$. Fortunately, (11) conforms to the Bellman optimality principle [9], and consequently may be posed as a variational problem, solvable by Dynamic Time Warping (DTW) [5,9] - a specialized Dynamic Programming technique - where e^2 is the local distance and the DTW local constraints reflect the constraints of (7a). The computational efficiency of the DTW technique motivates us to select $\{d_k(t)\}$ such that the application of $H_2^{-1}(t)$ to the gain normalized noise will diagonalize the noise covariance matrix Q_2. Another requirement is that the application of $H_2^{-1}(t)$ to the gain normalized signal will result in a time-invariant spectral envelope, to facilitate the estimation of the time-invariant $\{c_k\}$. Provided that the variations of the signal and the noise spectra are not too large - ensured by the way the phonemes are segmented - the above two requirements can be roughly met by using the following procedure:

First, estimation of the spectral envelope evolution of the signal by all pole modeling using any of the numerous available techniques. We preferred the Discrete All Pole algorithm (DAP) [2] since it is more suitable for voiced speech and may be used in a multi-speaker environment. Next, LPC inverse filtering of the gain normalized signal $x_1(t)$ to reduce spectral envelope variations, and finally, filtering by a time invariant noise whitening filter to diagonalize Q_2 as much as possible. In the special case where υ is white, this filter matches the average spectral envelope of the signal.

Now with Q_2 being diagonally dominant, we show how (11) may be solved by an iterative application of a simple comb filter [4] and DTW. To avoid cumbersome indexing, we present the procedure in the continuous time domains t and u. Let us rewrite (11) in the u time domain,

$$\underset{\phi,\ s_0}{Min} \int_{-\infty}^{\infty} \tilde{w}(u,\phi)\left|\tilde{x}_2(u,\phi) - s_0(u)\right|^2 du \tag{12a}$$

where $\tilde{w}(u,\phi) \triangleq w\left(\phi^{-1}(u)\right)\ \dot{\phi}^{-1}(u)$ (12b)

and $\tilde{x}_2(u,\phi) \triangleq x_2\left(\phi^{-1}(u)\right)$ (12c)

The following iterative procedure (Fig. 2), is proposed,

Initialization - Assume $\phi(t) = 2\pi\beta t$, where β is the average pitch of the segment in Hz.

Step A - Given the time-warped signal

$$\tilde{x}_2^{(n)}(u,\phi^{(n)}) \triangleq x_2\left(\phi^{(n)^{-1}}(u)\right) \tag{13a}$$

and the time-warped window,

$$\tilde{w}^{(n)}(u,\phi^{(n)}) \triangleq w\left(\phi^{(n)^{-1}}(u)\right)\ \dot{\phi}^{(n)^{-1}}(u) \tag{13b}$$

find the periodic waveform $s_0^{(n)}$ which minimizes (12a) with respect to s_0. The estimator can be shown [4] to be a comb filter in the u domain,

$$\hat{s}_0^{(n)}(u) = \frac{\sum_{k=-\infty}^{\infty} \tilde{w}^{(n)}(u+2\pi k)\,\tilde{x}_2^{(n)}(u+2\pi k)}{\sum_{k=-\infty}^{\infty} \tilde{w}^{(n)}(u+2\pi k)} \tag{14}$$

Step B - Given the best periodic estimate $s_0^{(n)}(u)$, use DTW to minimize (11) with respect to the warping function $\phi(t)$, subject to (7a), over some neighborhood of $\phi(t) = 2\pi\beta t$, and obtain $\phi^{(n+1)}(t)$

We note that step A is best carried out in the u domain since s_0 in (12a) is independent of $\phi(t)$. Conversely, step B minimizes (11) in the t domain since only s_0 there depends on $\phi(t)$. The procedure is guaranteed to converge to a stationary point of the error function since both steps cannot increase the error. With real voiced speech, convergence is achieved within 3-4 iterations.

4. SIMULATIONS

To demonstrate the capabilities of the model and the estimation scheme, we applied it to a 100 ms long segment (Fig. 3) of the phoneme 'u' having a considerable variation in pitch (160 to 200 Hz). Note the smearing of the higher harmonics in the spectrum due to the varying pitch (Fig. 4). Attempting to fit a strictly periodic signal to the given segment results in a poor match, primarily because the pitch and loudness are varying considerably. Next, the long-term model parameters are estimated and Fig. 3 compares the model to the given segment. The model is evidently capable of tracking the pitch and gain variations. The mismatches in signal amplitude, particularly those at the right side of the figure, imply that the spectral envelope and gain normalizations need to be further improved. Fig. 5 shows the spectrum of the time-warped signal (denoted by $\tilde{x}_2$ in Fig. 2). Comparing to Fig. 4, we note that the smeared harmonics became sharp and narrow . Furthermore, some higher harmonics that were obscured in Fig. 4 become apparent in Fig. 5.

5. CONCLUSION

We propose an efficient estimation scheme for the parameters of a long-term model for voiced speech, and demonstrate some of its capabilities. Preliminary simulations with voiced speech show that the model has potential in accurately describing whole voiced phonemes, even those having several hundreds of milliseconds in duration, subject to appropriate segmentation.

Future efforts will focus on improving the spectral envelope and gain normalizations. A related task is the optimal segmentation of phonemes in terms of minimal modeling errors vs. analysis frame length.

REFERENCES

[1] Almeida L.B. and Tribolet J.M., "Non-stationary spectral modeling of voiced speech",ASSP-31, No. 3, June 1983.

[2] Almeida L.B., "Frequency - varying sinusoidal modeling of speech", ASSP-37, No. 5,May 1989.

[3] El-Jaroudi A. and Makhoul J., "Discrete all-pole modeling", ASSP-39, No. 2, Feb. 1991.

[4] Friedman D.H., "Pseudo maximum likelihood pitch extraction", ASSP-25, No. 3, pp.213-221, June 1977.

[5] Ney H., "A time warping approach to fundamental period estimation", IEEE trans. on Systems, Man and Cybernetics, SMC-12, No. 3, May 1982.

[6] Martinelli G. et al.,"Excitation source identification in long term speech",Signal Processing III: Theories and Applications , pp. 565-568, 1986

[7] McAulay R.J. and Quatieri T.F., "Multirate sinusoidal transform coding at rates from 2.4 kbps", ICASSP-87, 1987.

[8] W. Bastiann Kleijn,"Continous representations in linear predictive coding", ICASSP-91,pp. 201-204, 1991

[9] Dreyfus S.E., "Dynamic programming and the calculus of variations", Academic Press, 1965

[10] D. Chazan, Y. Stettiner and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation", ICASSP-93

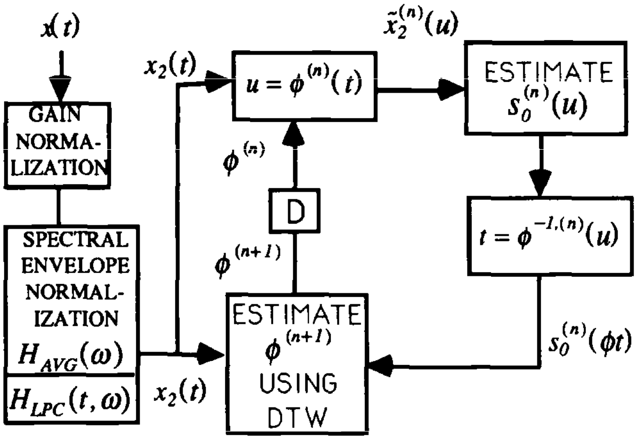


Fig.2 - Iterative scheme for the estimation of the warping function and the periodic excitation

Fig. 4 - Spectrum of the original voiced segement

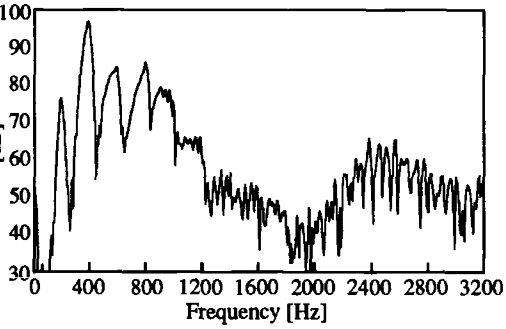


Fig. 5 - Spectrum of time-warped original

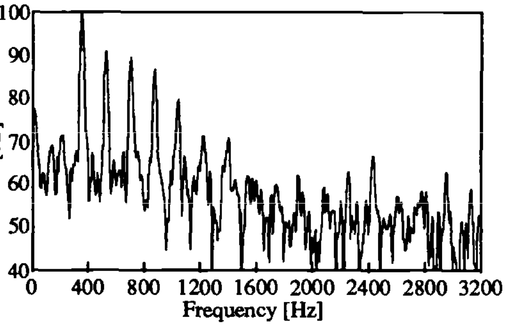
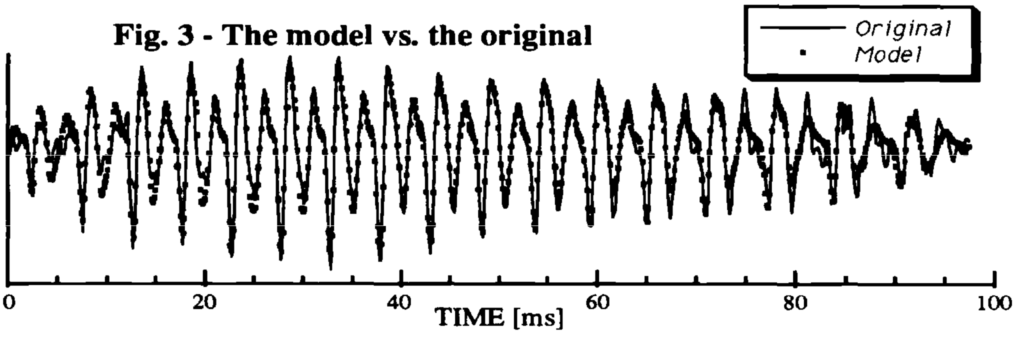


Fig. 3 - The model vs. the original


Original
Model
TIME [ms]