Analysis of the Voiced Speech using the Generalized Fourier Transform with Quadratic Phase

D. Petrinovic, V.Cuperman

Signal Compression Laboratory
University of California, Santa Barbara, ECE Department
davor,vladimir@ece.ucsb.edu

Abstract

One significant problem with sinusoidal modeling of the speech signal is due to the use of standard Fourier Transform for a quasi-periodic signal. The analysis accuracy is severely limited by the lack of stationarity of the analyzed segment, since the analysis is based on the conventional Fourier Transform. An improved analysis technique based on the Generalized Fourier Transform (GFT) with quadratic phase will be discussed in this paper. Speech signal is modeled as a sum of harmonic cosines but with nonlinear phases. A technique for estimation of the time-varying model parameters from the GFT spectrum is proposed. It will be shown that the modeling gain can be improved significantly by inclusion of a single additional parameter in the analysis procedure.

1. Introduction

Parametric modeling of the periodic signals based on sinusoidal representation is a very popular approach in signal compression [1][2][3]. The basic idea is to represent a periodic signal as a sum of sinusoidal signals that are integer multiples of the fundamental frequency of the signal. After analysis, the signal is parametrically represented using a set of the estimated parameters that include: the fundamental frequency, the amplitudes, and the phases of the harmonically related sines. The signal is assumed to consist of a set of fixed frequency sines, hence the estimation of the sine parameters can be performed in the spectral domain using the Fourier analysis.

For pseudo-stationary signals, such as the speech signal, modeling and spectral analysis must be performed on short time basis. Fourier analysis of the signal segment results in accurate parameter estimates only if the signal is really consistent with the initial fixed frequency assumption. To simplify the encoding of the estimated parameters, the width of the analysis window and the update rate of the parameters are usually fixed. The minimum window width is determined by the longest expected pitch period and the typical values for speech analysis are 20 to 30 ms. With such windows, the assumption of the fixed fundamental frequency may not hold any more, especially for high pitch female voices. For this reason, narrower windows are many times preferred, even at the expense of the increased update rate of the model parameters. Synthesis is usually performed by interpolation of the estimated parameters in the interval of L samples between the centers of the two neighboring analysis windows. By decreasing the update rate, the synthesis frame width L is increased. Since the tapering windows are usually applied to improve the spectral estimation of the model parameters, some overlap is required, such that the ratio between L and the analysis window width N is usually kept in the range of 0.5 to 0.75. Therefore, the minimum update rate is determined by the analysis window width.

Ideally, for good signal compression, the update rate should be kept as low as possible. In order to solve the problems related to signal pseudostationarity, we propose a model that reflects signal variability. The idea of frequency/amplitude varying sinusoidal model is not new but such models are usually employed only in the synthesis (e.g. [4][2][3]). Parameter estimation of time-varying models using the spectral analysis based on the conventional Fourier Transform is difficult, since the signal spectrum is distorted by frequency modulation.

A new signal-adaptive transformation will be introduced in this paper to enable accurate estimation of the time-varying model parameters employing a new spectral domain. By using the proposed approach, a wider analysis window can be used, resulting in the same estimation and modeling accuracy as with the conventional models but at the lower update rate of the model parameters.

2. Time-varying harmonic model

Quasi-periodic signals with a time varying period can be modeled as a sum of harmonically related sines, whose frequencies are integer multiples of a variable fundamental frequency $\Omega_0(t)$. This assumption is true as long as the spectral envelope of the signal is changing slowly. This concept of nonstationary modeling was introduced in [1] under the name of the generalized harmonics. A discrete time parametric model $\hat{x}(n)$ can be represented as a sum of generalized harmonics $\hat{x}_m(n)$, with complex amplitudes A_m:

$$\hat{x}(n) = \sum_{m=1}^{M} \hat{x}_m(n)\,, \quad \hat{x}_m(n) = \mathrm{real}(A_m e^{j\phi_m(n)}) \tag{1}$$

$$\text{where,} \qquad A_m = \left|A_m\right| e^{j\phi_{m0}} \tag{2}$$

The number of harmonics M is determined by the fundamental frequency of the signal in the discrete time domain, $\overline{\omega}_0$, i.e. $M<\pi/\overline{\omega}_0$. For sufficiently short analysis windows, a linear frequency model can be assumed [5], such that the instantaneous phase $\phi_m(n)$ of the mth harmonic is expressed as a quadratic function of the time index n:

$$\phi_m(n) = m\overline{\omega}_0 n\left(1 + n\frac{\delta_0}{2}\right) \tag{3}$$

while its initial phase is included in the complex amplitude A_m. It is obvious that this phase function corresponds to the following instantaneous harmonic frequency $\omega_m(n)$:

$$\omega_m(n) = m \cdot \omega_0(n)\,, \qquad \omega_0(n) = \overline{\omega}_0\left(1 + n\delta_0\right) \tag{4}$$

For a symmetric analysis windows, $\overline{\omega}_0$ is equal to the mean fundamental frequency of the model, that is also equal to the instantaneous fundamental frequency in the center of the analysis window (n=0), i.e. $\overline{\omega}_0 = \omega_0(0)$. The model parameters are thus: the mean fundamental frequency $\overline{\omega}_0$, the normalized slope of the fundamental frequency, δ_0, and the complex amplitudes of the harmonics A_m, m=1,..,M. A short hand notation for the model will be $S=\{\overline{\omega}_0, \delta_0, A_m, m=1,.., M\}$. This model includes as well the conventional fixed-frequency sinusoidal model for δ_0=0.

All these parameters will be estimated for each frame, on the analysis interval of N samples centered around the time origin n=0, i.e. $n\in[-(N-1)/2, (N-1)/2]$. For simplicity of the notation, the frame index will be omitted except for equation where an explicit frame index is really needed.

2.1. Analysis based on the time-varying harmonic model

In the conventional sinusoidal modeling, the analysis and parameter estimation are usually performed by assuming the linear phase model (δ_0=0). To improve the modeling accuracy, the frequency variability of the model should be included in the analysis as well. However, estimation of the time-varying model parameters from the Short Time Fourier Transform (STFT) of a quasi-periodic signal is very complex even if the frequency variation model of the signal is known exactly [1]. Almeida and Tribolet proposed a higher order spectral model where each generalized spectral line was described using a set of coefficients instead of single complex amplitude in order to capture the variability of the harmonic amplitude and frequency (phase). The estimation of the model parameters was based on MSE minimization of the modeling error in the Fourier domain.

For signals with variable frequency, the spectral model of a generalized harmonic in the Fourier domain is widened due to the frequency modulation effect. Widening is proportional to the harmonic frequency, such that neighboring high frequency harmonics overlap each other significantly, thus affecting the estimation accuracy of the model parameters. An iterative estimation algorithm was proposed in [5] and [6] by Marques and Almeida and it was shown that for certain convenient choices of the analysis windows, the model parameters can be estimated from STFT. However this algorithm results in proper estimates only for the low-order harmonics, whose initial position can be resolved by the peak-picking algorithm. Spectral line sharpening was proposed in [5] to solve the problem by using the time warping, such that the time-warped signal has almost constant frequency. The STFT based parameter estimation on such warped signal results in much better estimates, since the frequency modulation effect can be reduced by time warping

3. Generalized Fourier Transform with Quadratic Phase

The previous section shows that the Fourier Transform is not the best tool for spectral analysis of quasi-periodic signals. If the signal $x(n)$ can be modeled well as a sum of sinuses with linearly increasing frequency as in (4), then the optimal basis function should be constructed the same way, i.e. like complex exponentials with a constant frequency slope that is proportional to the mean frequency of the basis function. This corresponds to a generalization of the Fourier Transform, since the basis functions are still the complex exponentials but with a quadratic phase functions defined by the normalized frequency slope δ_0. Therefore, the Generalized Fourier Transform (GFT) is an adaptive transformation, parameterized by δ_0, whose discrete time representation, $GFT(x(n))$, is given by the following equation:

$$\Xi(e^{j\omega'},\delta_0) = GFT(x(n)) = \\ = \frac{1}{N}\sum_{n} w(n,\delta_0)\cdot x(n)\cdot e^{-j\omega' n(1+\delta_0 n/2)} \qquad (5)$$

The frequency variable is denoted by ω', $\omega' \in [-\pi, \pi]$, to emphasize the fact that it is different from the DFT frequency ω. Summation range over n in (5) is determined by the analysis window width N, as $n\in[-(N-1)/2, (N-1)/2]$. Only the case of the odd N will be considered since it simplifies the modeling of the signal phase at the center of the analysis frame. The adaptive analysis window $w(n,\delta_0)$ is defined as:

$$w(n,\delta_0) = \left(1+\delta_0 n\right)\sum_{q=0}^{Q}\alpha_q \cos\left(q\frac{2\pi}{N}\left(\frac{\delta_0}{2}n^2 + n - \delta_0\frac{N^2}{8}\right)\right) \qquad (6)$$

The shape of the window is determined by the coefficients α_0 to α_Q given in Table 1. for some commonly used windows.

Table 1. Typical window coefficients

Window	Q	α_0	α_1	α_2
Rectangular	0	1.00	-	-
Pseudo-Hamming	1	0.54	0.46	-
Pseudo-Hann	1	0.50	0.50	-
Pseudo-Blackman	2	0.42	0.50	0.08

Obviously, for δ_0=0, the equation (5) reduces to the conventional DFT, while the adaptive window $w(n,\delta_0)$ in equation (6) becomes one of the classical raised-cosine windows.

3.1. Parameter estimation in the GFT domain

Parameter estimation of the model S can be performed by minimizing the MSE modeling error, $\varepsilon(S)$, between the GFT of the signal $\Xi(e^{j\omega'},\delta_0)$ and GFT of the model $\hat{x}(n)$ in the ω' domain:

$$\varepsilon(S) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\Xi(e^{j\omega'},\delta_0) - \hat{\Xi}(e^{j\omega'},S)\right|^2 d\omega' \qquad (7)$$

The GFT of the model with parameters S, denoted by $\hat{\Xi}(e^{j\omega'},S)$ is simply a sum of GFTs of each of the model harmonics $\hat{\Xi}_m(e^{j\omega'},\delta_0)$ which can be found from the equation (5) by substituting $x(n)$ with $\hat{x}_m(n)$.

$$\hat{\Xi}(e^{j\omega'},S) = \sum_{m=0}^{M}\hat{\Xi}_m(e^{j\omega'},\delta_0) \qquad (8)$$

It is important to emphasis that the normalized slope of the GFT and the normalized slope of the model are chosen to be equal.

Under the assumption that the normalized slope δ_0 and the mean fundamental frequency $\overline{\omega}_0$ are known, the complex amplitude of the m^{th} model harmonic minimizing (7) can be determined as a normalized correlation between the GFT of the signal and GFT of the m^{th} harmonic of the model with unit amplitude, denoted with $\hat{\Xi}_m(e^{j\omega'},\delta_0,A_m=1)$, i.e.:

$$A_m = \frac{1}{2}\frac{\int\limits_{a_m}^{b_m} \Xi(e^{j\omega'},\delta_0)\cdot\hat{\Xi}_m^*(e^{j\omega'},\delta_0,A_m=1)\,d\omega'}{\int\limits_{a_m}^{b_m}\left|\hat{\Xi}_m(e^{j\omega'},\delta_0,A_m=1)\right|^2 d\omega'} \qquad (9)$$

where * denotes a complex conjugate. The integration is performed only within the main lobe of the window, so the limits a_m and b_m are symmetric around the harmonic frequency $m\overline{\omega}_0$ with displacement of ±Δω':

$$a_m = m\overline{\omega}_0 - \Delta\omega', \quad b_m = m\overline{\omega}_0 + \Delta\omega' \qquad (10)$$

$$\Delta\omega' = \min(\overline{\omega}_0/2,\ 2\pi(Q+1)/N) \qquad (11)$$

The estimation of the remaining parameters of the model, δ_0 and $\overline{\omega}_0$, depends on the actual application of the proposed modeling technique. Usually these parameters can be coarsely estimated using some simplified algorithms and then refined based on the Analysis By Synthesis (ABS) approach by direct minimization of (7). The initial slope can also be estimated from the conventional spectrum shape of the low frequency harmonics using algorithm given in [5].

3.2. Approximation of the GFT of the model

The parameter estimation can be simplified if the model spectrum $\hat{\Xi}(e^{j\omega'},S)$ is approximated by a closed form equation. First, the GFT of an adaptive window $w(n,\delta_0)$ will be calculated, by substituting $x(n)=1$ in the equation (5). The resulting GFT spectrum, $W(e^{j\omega'},\delta_0)$, can be approximated as a product of the conventional spectrum of the conventional window $w(n,0)$, $W_0(e^{j\omega})$, multiplied by a linear phase term that depends on the slope δ_0:

$$W(e^{j\omega'},\delta_0) \cong e^{-j\omega'\delta_0 N^2/8}\cdot W_0(e^{j\omega'}) \qquad (12)$$

Conventional window spectrum $W_0(e^{j\omega})$ can be determined from the coefficients α_0 to α_Q according to the well-known equation:

$$W_0(e^{j\omega}) = \alpha_0\frac{\sin(\omega N/2)}{\sin(\omega/2)} + \sum_{q=1}^{Q}\frac{\alpha_q}{2}\frac{\sin((\omega+\frac{2q\pi}{N})N/2)}{\sin((\omega+\frac{2q\pi}{N})/2)} + \sum_{q=1}^{Q}\frac{\alpha_q}{2}\frac{\sin((\omega-\frac{2q\pi}{N})N/2)}{\sin((\omega-\frac{2q\pi}{N})/2)} \qquad (13)$$

The accuracy of the approximation in (12) is very good even for very high slopes, (e.g. for linear variation of 20% in $N/2$ samples). Once the window spectrum is defined, the GTF of the m^th harmonic with unit amplitude, can be approximated using the familiar expression:

$$\hat{\Xi}_m(e^{j\omega'},\delta_0,A_m=1) \cong W(e^{j(\omega'-m\overline{\omega}_0)},\delta_0)/2 + W(e^{j(\omega'+m\overline{\omega}_0)},\delta_0)/2 \qquad (14)$$

Again, the accuracy of this approximation is very good as long as the harmonic frequency is low. However, for harmonics that are close to the Nyquist frequency, the modeling accuracy strongly depends on δ_0. The problem is that for high slope δ_0, the basis functions that are close to π may exceed the Nyquist frequency as a result of the linear slope. These functions wrap-around causing aliasing that affects the orthogonality. A half-width of the main spectral lobe of the window is equal to $Bw=2\pi\ (Q+1)/N$. If the harmonic frequency is within $[Bw,\ \pi\text{-}Bw]$ and if the window has sufficient side-lobe attenuation, then the second term of (14) corresponding to the mirror image can be ignored to simplify the modeling. The corresponding maximum approximation error for the Hamming window is shown in the Figure 1., as a function of the harmonic frequency and the normalized slope. The solid curve corresponds to zero-slope case, while the dashed curves are for δ_0 from 2% to 20 % within $N/2$ samples. It can be observed that for moderate slopes (<5%), only the last 10% of the spectrum are affected by the aliasing problem, while the modeling accuracy of the remaining part is identical to the zero-slope case.

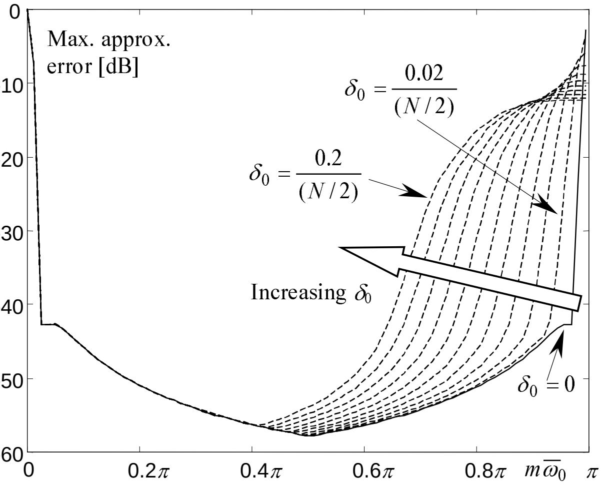

Figure 1: Accuracy of the spectral modeling (*N*=161).

4. Experimental results

To illustrate the potential of the improved analysis procedure based on GFT, an experiment was performed on the real speech data. Sixteen short utterances spoken by 8 male and 8 female speakers were used. To eliminate the problem of the spectral envelope variations, the proposed modeling was performed on the normalized residual signal of the LPC analysis. Speech was sampled at 8kHz, high-pass filtered above 80Hz, pre-emphasized using a fixed factor of 0.9375, and analyzed using the 10th order LPC. The LPC analysis window width was 25 ms with an update rate of 100 frames/s. Fixed bandwidth expansion of 10 Hz was applied to all the formants. The LPC residual signal was formed by inverse filtering and was normalized by a piecewise-linear amplitude envelope derived from the LPC gains. The voiced/unvoiced segmentation and an initial pitch contour were derived manually.

Harmonic analysis and modeling was performed for voiced frames with a 32ms Hamming window. The proposed GFT estimation was compared to the baseline fixed-frequency approach with $\delta_0=0$.

For the baseline case, the estimation procedure becomes identical to the MBE technique [3], but with the assumption of completely voiced excitation (all bands classified as voiced). The modeling error, $\varepsilon(S)$, was minimized using the technique known as pitch refinement [3], by selecting $\overline{\omega}_0$

from a candidate list and determining the optimal amplitudes for each candidate according to (9). One hundred candidates were used in the interval of ±10% around the initial estimate of the fundamental frequency for each of the voiced frames. Due to the exhaustive nature of this ABS technique, the final results represent the best possible stationary harmonic match to the input signal. Estimation was performed in the DFT domain with the spectral resolution of M=512 samples. A modeling gain, G, was calculated for each frame as a quotient of the energy of the windowed signal and the energy of the modeling residual $\varepsilon(S)$. Scatter plot of the MBE modeling gain G expressed in dB is shown in Figure 2. as a function of the normalized slope δ_0 of each of the analysis frames. The slope on the x-axis is converted to the continuous time domain by multiplication with the sampling rate f_s, such that it can be interpreted as a percentual frequency increase on the 10 ms intervals.

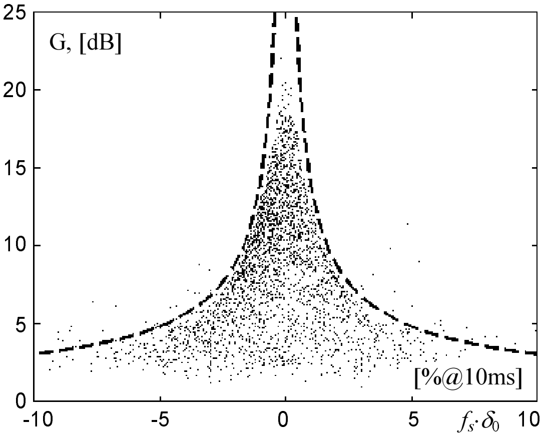

Figure 2: MBE modeling gain vs. normalized slope

It can be observed, that the modeling gain of the convetional analysis drops rapidly as δ_0 is increased. It is roughly bounded by the dashed curve given by:

$$G \le 14 \cdot \left| f_s \delta_0 \right|^{-2/3} \quad \text{[dB]} \tag{15}$$

The results of the first procedure were used as initial estimates for the GFT based analysis. The initial slope $\delta_0(i)$ for the analysis frame i was calculated from the estimated pitch values of the frames i-1, i and i+1 according to:

$$\delta_0(i) = \frac{1}{2L} \frac{\overline{\varpi}_0(i+1) - \overline{\varpi}_0(i-1)}{\overline{\varpi}_0(i)} \tag{16}$$

To evaluate the maximum possible benefit of the proposed GFT based analysis technique, the optimal model parameters were estimated using the exhaustive search over a two dimensional grid with 21x21 $\overline{\varpi}_0 / \delta_0$ candidate pairs. The search grid for each frame was centered on the initial MBE estimates. The slope was varied by ±3%@10ms around the initial δ_0, while $\overline{\varpi}_0$ was varied by ±2%. The final solution was derived from 2-D interpolation of the sampled $\varepsilon(S)$ surface determined in the search procedure.

The increase of the modeling gain, ΔG was calculated for each frame and it is shown in Figure 3 as a function of the signal slope δ_0. As expected, for stationary frames with δ_0 close to 0, both techniques resulted in the same gain, but for frames that contain pitch variations the modeling accuracy can be significantly improved. Average modeling gain improvement for all voiced frames in the database was: 1.64 dB for the interpolated solution; 1.54 dB for the best solution on the grid; and 0.92 dB for the initial $\overline{\varpi}_0 / \delta_0$ estimate. On a large number of frames (cca. 8%) the improvement was greater then 5dB.

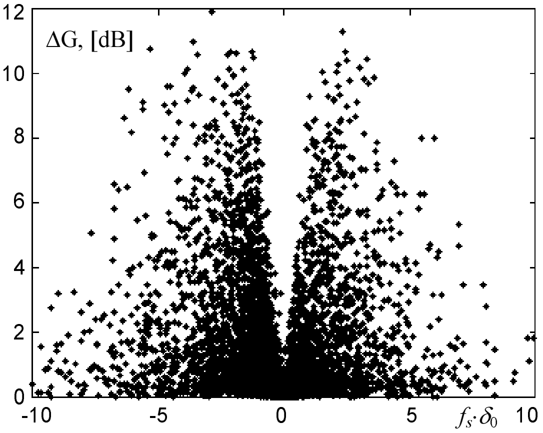

Figure 3: Increase of the modeling gain

5. Conclusions

A new speech analysis technique was described, that can improve the accuracy of the sinusoidal voiced speech modeling especially for speech with significant time variability, e.g. conversational or emotional speech. The proposed method addresses the problem of signal frequency variability encountered in the analysis systems with fixed and relatively wide analysis windows. By adapting the signal transformation to the local frequency variations of the signal, the overall modeling gain can be improved by 1.6 dB.

6. References

[1] Almeida, L.B. and Tribolet, J.M., "Nonstationary spectral modeling of voiced speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 31, no. 3, pp. 664-678, June 1983.

[2] McAulay, R.J. and Quatieri, T.F., "Speech analysis/synthesis based on sinusoidal representation", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 744-754, August 1986

[3] Griffin, D.W. and Lim, J.S., "Multiband excitation vocoder", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 8, pp. 1223-1235, August 1988.

[4] Almeida, L.B. and Silva, F.M., "Variable-frequency sinthesys: an improved harmonic coding scheme", *Proceedings of IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1984, pp. 27.5.1-27.5.4

[5] Marques J.S. and Almeida, L.B., "A background for sinusoid based representation of voiced speech", *Proceedings of IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1986, pp. 24.3.1-24.3.4

[6] Marques J.S. and Almeida, L.B., "Frequency-varying sinusoidal modeling of speech", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 5, pp. 763-765, May 1989