MIXTURE GAUSSIAN ENVELOPE CHIRP MODEL FOR SPEECH AND AUDIO

Bishwarup Mondal and T. V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore, India
bishwa@sasi.com, tvsree@ece.iisc.ernet.in

ABSTRACT

We develop a parametric sinusoidal analysis/synthesis model which can be applied to both speech and audio signals. These signals are characterised by large amplitude variations and small frequency variation within a short analysis frame. The model comprises of a Gaussian mixture representation for the envelope and a sum of linear chirps for the frequency components. A closed form solution is derived for the frequency domain parameters of a chirp with Gaussian-mixture envelope, based on the spectral moments. An iterative algorithm is developed to select and estimate prominent chirps based on the psycho-acoustic masking threshold. The model can adaptively select the number of time-domain and frequency-domain parameters to suit a particular type of signal. Experimental evaluation of the technique has shown that about 2 to 4 parameters/ms is sufficient for near transparent quality reconstruction of a variety of wide-band music and speech signals.

1. INTRODUCTION

Speech and music signals are inherently non-stationary. The important time-varying characteristics of these signals are the time-varying harmonics and the signal time-envelope as determined by the signal analysis and psycho-acoustic experiments. The currently successful models of speech and audio, assume the signal characteristics to be time-invariant over an analysis window and estimate either the spectral envelope (LPC) or the individual harmonics (sinusoids). The time-varying information is then generated either through interpolation between successive windows or as a residual within the window of the time-invariant envelope. There have been attempts to directly estimate time-varying components in the signal [1, 5, 6], which have used quite restrictive models in terms of modelling the time-envelope or the frequency variation. Also, the models in [2, 5, 6, 7] have been mainly for speech signals; audio signal models in [3, 8, 9] use stationary sinusoidal model appended with residual waveform representation. We develop a general model for speech and audio signals which provides for time-varying characteristic of the short-time envelope as well as time-varying spectral components. This results in minimisation of interpolation between successive windows or the need for using residual signals. The improved envelope model has resulted in reduced pre-echo [4] which is a common problem in audio signals.

2. MIXTURE GAUSSIAN CHIRP MODEL (MGC)

The mixture Gaussian model is an extension of the earlier model of speech which used Gaussian linear chirps [5, 6]. In general, time-varying tonal signals, such as notes of musical instruments and voiced speech, can be considered as a sum of partials with a time-varying envelope and frequency, over the duration of the note or the phoneme. Over a short analysis segment, much smaller than a note, the signal is usually assumed to be time-invariant (quasi-stationary), which has lead to the development of either the sinusoidal model or linear system (LPC) model of the signal. It is clear that over many segments of speech or audio, the time-invariance is not valid, leading to poor quality synthesis. To improve this situation, we can consider linear chirps for the frequencies of the partials and a Gaussian mixture function for the envelope of the signal segment of the order of $100ms$. The reconstructed signal in the m^{th} frame can be represented as $\widehat{s}_m(n) =$

$$\left[\sum_{i=1}^{G_m} A_{i_m} e^{-\alpha_{i_m}(n-\mu_{i_m})^2}\right]\left[\sum_{l=1}^{L_m} B_{l_m} e^{j(\omega_{l_m} n+\beta_{l_m} n^2+\phi_{l_m})}\right] \quad (1)$$

which is a sum of $L_m/2$ distinct chirp components (for a real valued signal) where the instantaneous frequency (IF) of the l^{th} component is given by $\Omega_{l_m}(n) = \omega_{l_m} + 2\beta_{l_m} n$. The total signal envelope is represented as a sum of G_m number of Gaussian components with means μ_{i_m}, variances $1/2\alpha_{i_m}$ and scale factors A_{i_m}. For short segments of the signal, the Gaussian mixture model is considered adequate since it can represent sudden attacks or decays or multi-modal shapes.

3. MODEL ESTIMATION

The joint estimation of the parameters in eqn (1) is quite complex and may not yield any useful results. Instead, we resort to a sequential approach of first estimating the envelope parameters and then the spectral components, successively.

3.1. Envelope parameters

The parameters which determine the time-envelope of the signal are G_m, A_{i_m}, μ_{i_m} and α_{i_m}. Let $\sigma(n)$, $n = -N, .., 0, .., N$, be the envelope of the signal in a frame of duration $2N+1$. We estimate $\sigma(n)$ by filtering the rectified signal with a low-pass filter having a transition band in the range of $40 - 80Hz$. Fig.1 illustrates the algorithm for the envelope parameter estimation. Among the various algorithms for mixture Gaussian approximation, this approach is found simple and effective. In this procedure, major peaks in the envelope function $\sigma(n)$ are identified and each is fitted with a Gaussian function, sequentially. The parameters of each Gaussian are obtained from the sample mean and sample variance of the windowed envelope, $\sigma_w(n) = \sigma(n).w(n)$;

$w(n) = e^{-\alpha_w (n-\mu_w)^2}$ where μ_w is at the chosen peak and α_w depends on the adjoining valleys. Assuming $\sigma(n) \approx Ae^{-\alpha(n-\mu)^2}$, around μ_w, A, μ, α can be easily solved from the computed sample mean μ_s, sample variance α_s and the chosen μ_w and α_w.

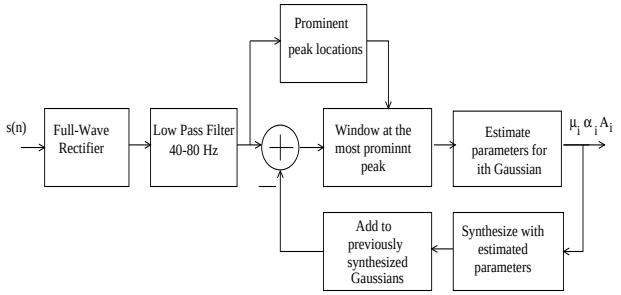

Fig. 1. *Identification and estimation of envelope parameters*

3.2. Spectral parameters

The spectral parameters are ω_{l_m}, β_{l_m}, B_{l_m}, ϕ_{l_m} and L_m. For this estimation, instead of the MMSE (minimum mean square error) approach, we have formulated a method based on spectral moments, which leads to closed-form solution of the spectral parameters for a single chirp. Taking a successive approximation approach, the parameters of all the chirps are estimated, sequentially.

3.2.1. Single chirp

Let s(t) be a linear chirp having an envelope of a sum of G Gaussians; i.e., $s(t) = [\sum_{i=1}^{G} k_i g_i(t)] B e^{j\beta t^2/2 + j\omega_0 t + \phi_0}$ where $g_i(t) = e^{-\alpha_i (n-\mu_i)^2}$, and the parameters of $g_i(t)$ and k_i have been already estimated through the time-domain envelope estimation procedure. Let us consider the spectral moments of the signal $s(t)$,

$$\langle \omega \rangle = \omega_0 + \beta \frac{E_t}{E} \tag{2}$$

$$\langle \omega^2 \rangle - \langle \omega \rangle^2 = \beta^2 \left(\frac{E_{t^2}}{E} - \frac{E_t^2}{E^2}\right) + \frac{E_c}{E} \tag{3}$$

where $E = \int |s(t)|^2 \, dt = \int [\sum_{i=1}^{G} k_i g_i(t)]^2 \, dt$,
$E_t = \int t |s(t)|^2 \, dt = \int t [\sum_{i=1}^{G} k_i g_i(t)]^2 \, dt$,
$E_{t^2} = \int t^2 |s(t)|^2 \, dt = \int t^2 [\sum_{i=1}^{G} k_i g_i(t)]^2 \, dt$,
$E_c = \int [\sum_{i=1}^{G} \sum_{j=1}^{G} k_i k_j \alpha_i \alpha_j (t-\mu_i)(t-\mu_j) g_i(t) g_j(t)] \, dt$.

Since each $g_i(t)$ is a Gaussian function, we can easily avoid the integration to evaluate E, E_t, E_{t^2} and E_c and instead obtain them through closed form expressions in terms of k_i, α_i, and μ_i, which have been earlier estimated. Also, we can compute the spectral moments in eqns (2, 3) from the computed $|S(\omega)|^2$. Thus, using E, E_t, E_{t^2}, E_c and the moments, we can solve for ω_0 and β in a closed form. The sign of β is determined from the curvature of the phase around ω_0. The complex valued scale factor of the chirp is obtained using a least square error formulation [5, 6], which results in $Be^{j\phi_0} = \frac{\int X^*_{est}(\omega) S(\omega) \, d\omega}{\int |X_{est}(\omega)|^2 \, d\omega}$ where $X_{est}(\omega) \iff [\sum_{i=1}^{G} k_i e^{-\alpha_i (n-\mu_i)^2}] e^{j\beta t^2/2 + j\omega_0 t}$

3.2.2. Multiple chirps

The case of multiple chirps is solved using a successive approximation (Fig 2) approach of identifying the dominant component, isolating it using a frequency-domain window and estimating the chirp parameters. After the parameters of one chirp are estimated, the chirp is synthesised and subtracted from the signal and the next chirp is estimated from the spectrum of the residual signal. This process is continued until all the identified chirps are estimated.
For resolving between chirps, it is clear that longer time-domain analysis window would be beneficial whereas time localisation of the beginning and ending of the chirp would be affected by the long window. Hence frame-window has to be chosen optimally and the frequency-domain window (for isolating the chirps) should be a function of the time-window. Let $\Omega_l(\omega)$ be the frequency-domain window for estimating the l^{th} chirp component. The windowed spectrum, $\Omega_l(\omega)|S(\omega)|^2$, approximately represents a single chirp with mixture Gaussian envelope. Let $\widehat{\omega_l}$, $1 \leq l \leq L_m$ be the prominent spectral peaks of the multicomponent signal. Using $\widehat{\omega_l}$ as the centre-frequency, a Gaussian frequency-domain window can be used: $\Omega_l(\omega) = e^{-\gamma_l \sigma_a^2 (\omega - \widehat{\omega_l})^2}$, where σ_a^2 is the variance of the analysis time-window and γ_l is a factor chosen such that the width of the window $\Omega_l(\omega)$ can be scaled depending on the nature of the peaks in the spectrum. The scaling is important because, a broad frequency-window will be affected by adjacent peaks whereas a narrow window causes underestimation of the chirp rate. From the sample mean and variance computed from the windowed spectrum $|S(\omega)|^2 \Omega_l(\omega)$, ω_l and β_l can be solved through the eqns. (2, 3).
It may be noted that all the frequency-domain parameters can be refined through subsequent iterations of re-estimating each chirp from the residual signal obtained after subtracting the contributions of all other identified chirps with the latest estimated parameters. We can also have different envelope functions for each chirp (generalisation to the present model), by re-estimating the time-domain parameters also, after each chirp is estimated and subtracted from the signal. But, initial experiments did not indicate a need for this generalisation.

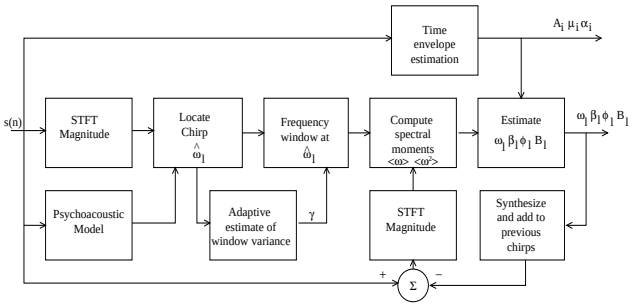

Fig. 2. *Frequency domain parameter estimation through successive approximation*

4. PERFORMANCE COMPARISON

The performance of the new parameter estimation technique is compared with other techniques available in the literature for estimation of constant amplitude multi-component chirp signals. Also, the performance of the model on real signals of speech and audio is determined using a measure of the number of parameters required to synthesise the signal, with near transparent quality.

4.1. Synthetic Signals

The first signal is a two component constant-amplitude chirp signal, reported in [5, 6]. The signal has significant spectral overlap because of high β_l as well as small ω_l separation. The actual parameter values and those estimated, are compared in Table-1. It can be seen that estimation error is $< 1\%$ and this compares favourably with that reported in [5, 6]. The second signal, shown in Table-2, simulates a voiced speech segment. It can be seen that the new model is able to estimate most of the parameters with $< 1\%$ error, only the β_l of the high-frequency components (11 to 15) are found to have more error. This is a considerable improvement over the results for the same signal, reported using an algorithm based on MMSE [5], which showed that it was possible to estimate only the first 5 harmonics.

Table 1. Two component constant amplitude chirp signal parameters and MGC estimates

Cmp	A_l	$\widehat{A_l}$	ϕ_l	$\widehat{\phi_l}$	ω_l	$\widehat{\omega_l}$	β_l	$\widehat{\beta_l}$
1	1	0.99	0	-0.03	1.0	1.0	1e-3	1e-3
2	0.8	0.79	1	1.03	1.2	1.2	1e-3	1e-3

Table 2. Synthetic speech segment parameters and estimated parameters using the MGC model

Cmp	A_l	$\widehat{A_l}$	ϕ_l	$\widehat{\phi_l}$	ω_l	$\widehat{\omega_l}$	$\beta_l \times 10^4$	$\widehat{\beta_l}$
1	0.8	0.80	0	0.0	0.2	0.20	1.5	1.5
2	0.7	0.70	1	1.0	0.4	0.40	3.0	3.0
3	0.9	0.90	0	0.0	0.6	0.60	4.5	4.5
4	1.0	1.00	-1	-1.0	0.8	0.80	6.0	6.0
5	0.9	0.90	0	0.0	1.0	1.00	7.5	7.5
6	0.8	0.80	1	1.0	1.2	1.20	9.0	9.0
7	0.6	0.60	0	0.0	1.4	1.40	10.5	10.5
8	0.4	0.40	-1	-1.0	1.6	1.60	12.0	12.0
9	0.3	0.30	0	-0.0	1.8	1.80	13.5	13.5
10	0.4	0.40	1	1.0	2.0	2.00	15.0	15.0
11	0.5	0.49	0	0.0	2.2	2.19	16.5	16.4
12	0.4	0.39	-1	-0.9	2.4	2.39	17.5	16.9
13	0.4	0.39	0	-0.0	2.6	2.60	19.0	20.4
14	0.3	0.29	1	1.1	2.8	2.79	20.5	17.5
15	0.2	0.17	0	0.0	3.0	2.97	22.0	16.4

4.2. Speech/Audio signals

The concern in modelling speech and audio is to obtain a minimum number of components ($G_m + L_m$) i.e., the number of Gaussians for the envelope and the number of chirps, which can result in transparent quality reconstruction. It is seen that different types of signals (transient, tonal) require different G_m and L_m. Hence, the MGC parameter estimation algorithm determines these quantities adaptively for each frame, to meet the psychoacoustic threshold [10] for the frame. Thus, we can quantify the average modelling cost as $(\frac{1}{M}\sum_{m=1}^{M} 3G_m + 4L_m)/framesize(ms)$, where each Gaussian in the envelope is 3 parameters and each chirp is 4 parameters. It may be noted that the measure of parameters/ms is similar to the measure of perceptual entropy, indicating the minimum number of parameters to achieve near transparent quality reconstruction.

In the first experiment, a male-female conversational speech (of GSM evaluation tests) sampled at 8kHz is used for evaluation. A overlapped frame analysis of 25ms frame size and 64ms Kaiser window (beta =8) is chosen to minimise spectral domain side-lobe leakage. An overlap-add synthesis is performed. It is found that an average of 3.7852 parameters/ms results in near transparent quality speech. In the second experimental condition (case-2), the signal envelope is restricted to be one Gaussian which resulted in an overall parameter reduction of 15.27%. But, the perceptual quality showed several artifacts. In contrast, constraining the spectral components and leaving the envelope components unconstrained, resulted in a 18.66% reduction, but no artifacts; the quality is quite close to case-1. This shows the importance of mixture-Gaussian envelope modelling. Fig (3) compares voiced speech modelling using chirps and constant sinusoids. It is clear that the chirps are more effective in reducing the modelling error especially in the mid-high frequency regions where the spectral peaks are broader and cannot be accurately modelled by sinusoids.

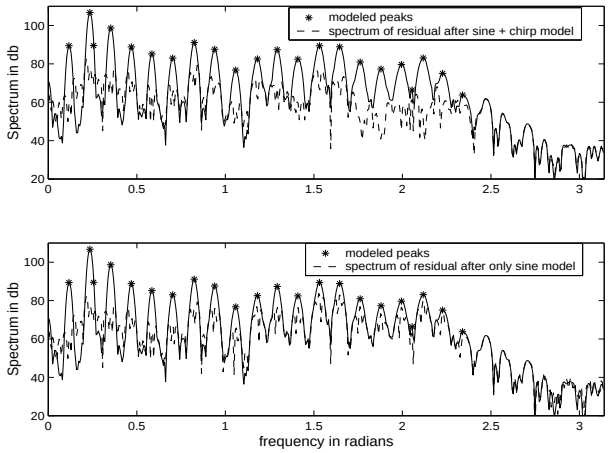

Fig. 3. *Spectral fit for one frame of voiced speech (a) residual spectrum after sine + chirp model is $\sim 20dB$ below the original at peak locations throughout the frequency scale (b) residual with sine only model shows that broader peaks at frequencies > 1 rad (1.3 kHz) are poorly modelled*

Table 3. MGC modelling of speech; NT = near transparent, A = artifacts, NA = no artifacts

Case	Gauss/ms	Sines+Chirps/ms	param/ms	quality
1	0.1603	0.9197	4.1597	NT
2	0.0800	0.8000	3.4400	A
3	0.1603	0.7197	3.3597	NA

Next, several pieces of instrumental music are selected from the 'SQAM' test data (16kHz samples and monophonic). The MGC analysis is done with 50ms frame size and 128ms Kaiser window (beta=8) with overlap-add synthesis. The parametric representation of signals 1 to 5 (Table-4) resulted in near transparent quality reconstruction of the original sound. Trumpet has the maximum richness among the five signals (which is also perceptually justified), whereas flute, piccolo and oboe have lower parametric entropy. As expected, the percussion type of signal has the maximum rate of parameters for the envelope function and trumpet has the maximum frequency domain load. It has been observed that reducing the number of frequency-domain parameters by 30% without restricting the time-domain parameters reduces the perceptual quality but do not introduce artifacts. It shows that modelling the

time-envelope of speech/audio signals is crucial in preserving their natural quality and restricting the frequency-domain parameters offers a more graceful degradation in perceptual quality. On the other hand, reducing the time-domain parameters results in pre-echo or other artifacts.

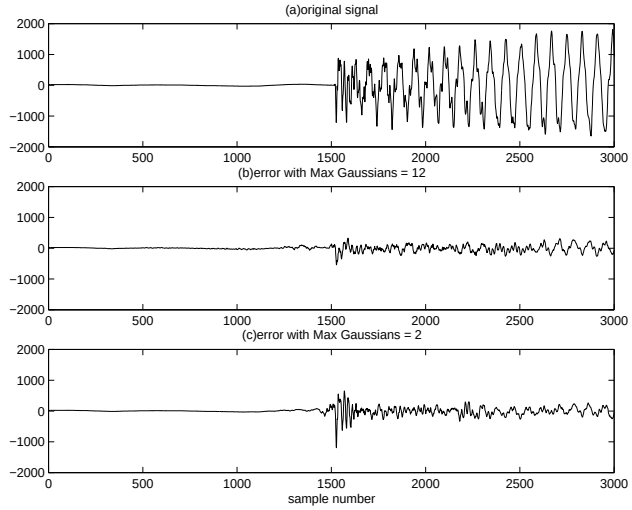

Fig. 4. *Reduction of pre-echo using mixture Gaussian envelope modelling (a)time-signal of mari showing sharp attack (b)residual after modelling with $Max(G_m) = 12$ (b)residual after modelling with $Max(G_m) = 2$*

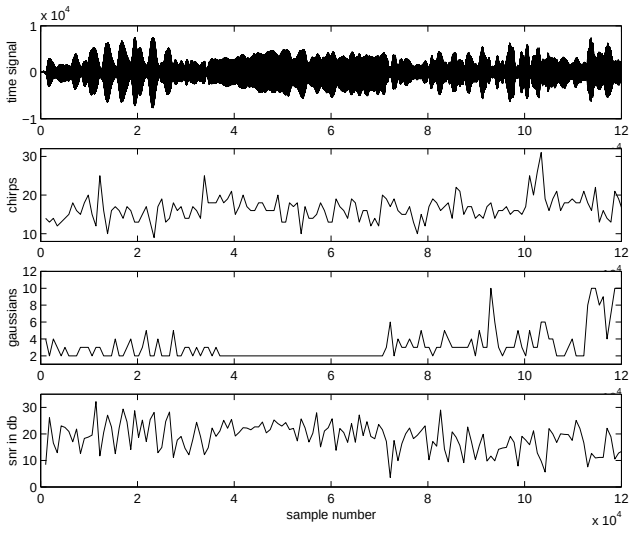

Fig. 5. *Adaptive choice of parameters based on signal characteristics (a)time-signal of piccolo (b)the number of chirps (L_m) varies depending on the psychoacoustic threshold. (c)the number of Gaussians (G_m) required to model the time-envelope (d)segmental SNR with average SNR = 18.36db*

5. CONCLUSION

A mixture Gaussian envelope chirp model (MGC) is proposed for representing time-varying signals as a sum of linear chirp components. This is a generalisation of the classical stationary sinusoidal model developed for speech and later extended to sinusoids with time-varying amplitude and linear chirps with a single-Gaussian envelope. While the existing models have been found to be insufficient for high quality reconstruction of audio signals, the new MGC model is shown to be effective for a variety of signals that are harmonic, tonal, transient and also noise-like speech signals. This has been achieved because of effective modelling of the signal envelope as well as a built-in tradeoff of the number of parameters used for time-domain envelope and frequency domain chirps in the overall model. The parameter estimation of the new model is developed around a closed-form solution for a mixed-Gaussian envelope chirp parameters using spectral moments. Another novelty of the proposed scheme for real signals is the use of psychoacoustic criteria in the selection of the chirps.

Table 4. MGC modelling of audio for near transparent quality

Case	Sig	Gauss/ms	Sines+Chirps/ms	param/ms
1	trumpet	0.1025	0.4824	2.2371
2	mari	0.2239	0.3391	2.0281
3	oboe	0.0731	0.3613	1.6645
4	picc	0.0700	0.3288	1.5252
5	flute	0.0918	0.3281	1.5858

6. REFERENCES

[1] L. B. Almeida and J. M. Tribolet, Non-Stationary spectral modeling of Voiced Speech, *IEEE Trans. ASSP, vol. ASSP-31, No. 3, 1983*

[2] E. B. George and M. J. T. Smith, Speech Analysis/Synthesis and Modification Using an Analysis-by -Synthesis/Overlap-Add Sinusoidal Model, *IEEE Trans. SAP, vol-5, No. 5, 1997*

[3] S. N. Levine and J. O. Smith III, A Switched Parametric and Transform Audio Coder, *from http://www-ccrma.stanford.edu/~scottl*

[4] S. N. Levine, T. S. Verma, J. O. Smith III, Multiresolution Sinusoidal Modeling for wideband audio with modifications, *ICASSP, 1998*

[5] J. S. Marques and L. B. Almeida, A Background for Sinusoid Based Representation of Voiced Speech, *ICASSP, 1986*

[6] J. S. Marques and L. B. Almeida, Frequency-Varying Sinusoidal Modelling of Speech, *IEEE Trans. ASSP, vol-37, No. 5, 1989*

[7] R. J. McAulay and T. F. Quatiery, Speech Analysis/Synthesis Based on a Sinusoidal Representation, *IEEE Trans. ASSP, Vol ASSP-34, No. 4, 1986*

[8] X. Serra, A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition, PhD thesis, Stanford University, 1990

[9] K. N. Hamdy, M. Ali, A. H. Tewfik, Low Bitrate high quality audio coding with combined harmonic and wavelet representations, *ICASSP, 1996*

[10] B. Edler, H. Purnhagen, C. Ferekidis, ASAC - Analysis/Synthesis audio codec for vry low bitrates *100th AES Convention, 1996*