CHIRP RATE ESTIMATION OF SPEECH BASED ON A TIME-VARYING QUASI-HARMONIC MODEL

Yannis Pantazis[1] , *Olivier Rosec*[2] *and* *Yannis Stylianou*[1]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2]Orange Labs TECH/SSTP/VMI, Lannion, France
email: pantazis@csd.uoc.gr, olivier.rosec@orange-ftgroup.com and yannis@csd.uoc.gr

ABSTRACT

The speech signal is usually considered as stationary during short analysis time intervals. Though this assumption may be sufficient in some applications, it is not valid for high-resolution speech analysis and in applications such as speech transformation and objective voice function assessment for detection of voice disorders. In speech, there are non stationary components, for instance time-varying amplitudes and frequencies, which may change quickly over short time intervals. In this paper, a previously suggested time-varying quasi-harmonic model is extended in order or to estimate the chirp rate for each sinusoidal component, thus successfully tracking fast variations in frequency and amplitude. The parameters of the model are estimated through linear Least Squares and the model accuracy is evaluated on synthetic chirp signals. Experiments on speech signals indicate that the new model is able to efficiently estimate the signal component chirp rates, providing means to develop more accurate speech models for high-quality speech transformations.

Index Terms— Speech modeling, speech analysis, non stationary analysis, chirp rate, f_0 evolution

1. INTRODUCTION

It is well known that speech is, in essence, a non-stationary signal. The origins of these non-stationarities are intimately related to the speech production mechanisms, including vocal tract movements, vocal fold vibration and lip radiation. The classical approach in speech processing is to consider that, under a reasonably small interval, the coarse spectral structure of speech can be considered as stationary. So, it is common practice to separate the speech signal into segments in which the parameters of interest can be estimated in a relatively reliable manner. This assumption enabled the widespread use of techniques based on Fourier analysis, linear prediction and sinusoidal/harmonic models in speech analysis, coding, synthesis and modification. In all of these applications, the underlying technologies must somehow deal with residual information that cannot be captured by the speech model itself. Consequently, we seek to better model the deterministic part of the speech signal, thus reducing the residual information. In speech coding, this improved modeling could lead to more efficient coding schemes, achieving a better compromise between the speech quality and coding rate. In speech synthesis and modification, the residual information is considered to be purely stochastic, though this is not the case. For instance, in the Harmonic plus Noise Model (HNM), the high-frequency region of the spectrum is generated as a modulated noise component, while the lower band is considered deterministic [1]. However, this discrimination between the deterministic and stochastic parts of the signal is rather artificial. In these (and many more) speech applications, there is a clear need for models that better capture the deterministic content of speech signals. A step further in speech modeling would be to take into account the non-stationary nature of speech directly in the modeling. This improved modeling would be particularly helpful in analyzing transient parts of a signal in which rapid movements of the vocal tract, as well as rapid variations of the vocal fold vibrations, are observed. More generally, so-called atypical voices and/or phonations (e.g. pathological voices; voices produced by elderly people or children; emotional speech) may exhibit non-stationarities, even within small observation windows in which one would expect the speech signal to be stable.

The aforementioned non-stationarities manifest themselves in AM-FM modulation of the speech signal. The estimation of such modulation effects has been addressed in various ways. A first class of methods uses non parametric time-frequency representations such as the STFT, Spectrogram or Wigner-Ville distribution [2] to locate the relevant modulation events and estimate their parameters. A second approach is to extract relevant information from the speech signal and then apply operators in order to estimate the AM-FM components. This can be done either with a Hilbert transform [3] or by applying energy operators like the Teager-Kaiser operator [4]. The limitations of these techniques lie in their lack of robustness, especially in the case of multi-component AM-FM signals. Interestingly, a method based on a non-parametric Fan-Chirp analysis has been suggested in [5] [6] to track the frequencies of harmonically related sinusoidal components.

In this paper we propose a method for speech analysis based on a parametric time-varying model that captures the linear evolution of the frequency of sinusoidal components. In previous work [7], important properties of a model initially introduced by Laroche et al. in [8] were revealed: in essence, the complex slope, introduced to capture variations of the harmonic components, can be decomposed into two terms, one for frequency adjustment and the other for the amplitude slope. In this work we extend the model by introducing a second order complex polynomial for each harmonic component. We describe the overall estimation procedure, based on the minimization of a Least Square criterion, and we further propose an iterative scheme to refine the model parameters and, consequently, the estimation of the different sinusoidal components. Typical simulation results carried out on synthetic chirp signals, as well as on real speech signals, illustrate the potential of the proposed method to effectively track the linear evolution of the frequency of each sinusoidal component independently. Note that this is an interesting property since the model is not limited to strictly harmonically related components as in [5]. Moreover it is shown that the model can account for errors in the initial estimates of the sinusoidal component

frequencies.

The paper is organized as follows. Section 2 presents the model and its underlying properties. The estimation procedure is then described in section 3. Section 4 illustrates the behavior of the proposed model on a synthetic chirp signal, as well as on speech signals. Section 5 concludes the paper.

2. MODEL FORMULATION

2.1. Motivation

In this paper we investigate new methods for the analysis of harmonically related sinusoidal components that can be approximated on a certain time interval $[-t_0, t_0]$ by the following equation:

$$s(t) = \sum_{k=-K}^{K} A_k(1+\gamma_{1,k}t+\gamma_{2,k}t^2)e^{j(2\pi k f_0 t+\phi_{2,k}t^2+\phi_{1,k}t+\phi_{0,k})}. \tag{1}$$

where K is the number of harmonics, also known as the order of the model, f_0 is the local fundamental frequency and A_k, $\gamma_{1,k}$ and $\gamma_{2,k}$ are real coefficients that define the amplitude polynomial of the kth component. $\phi_{0,k}, \phi_{1,k}$ and $\phi_{2,k}$ are the coefficients of the phase polynomial of the kth component. Note that for this class of signals the amplitude polynomial of each harmonic is able to model a large variety of amplitude modulations while the phase polynomial is able to capture two phenomena present in speech signals: firstly, the frequency mismatches between kf_0 and the actual kth harmonic frequency which may be different due to an erroneous f_0 estimation or due to the detuning of some harmonics [7], and secondly, the linear evolution of each harmonic frequency through the term $\phi_{2,k}t^2$. The chirp rate of the kth component is given as $2\phi_{2,k}$.

The estimation of the above unknown parameters of the speech signal is a highly nonlinear procedure. In order to obtain a linear estimation problem, a simple yet powerful technique is to approximate the signal in eq. (1) by Taylor series expansion. Thus, for one component, the second order Taylor series approximation gives:

$$\begin{aligned} s_k(t) \approx{}& A_k e^{j\phi_{0,k}}(1+\gamma_{1,k}t+\gamma_{2,k}t^2) \\ & [1+j(\phi_{1,k}t+\phi_{2,k}t^2)-\frac{1}{2}(\phi_{1,k}t+\phi_{2,k}t^2)^2]e^{j2\pi k f_0 t}. \end{aligned} \tag{2}$$

Keeping the order of the polynomial up to 2, $s_k(t)$ is further approximated by:

$$\begin{aligned} s_k(t) \approx{}& A_k e^{j\phi_{0,k}}[1+(\gamma_{1,k}+j\phi_{1,k})t \\ & +(\gamma_{2,k}-\phi_{1,k}^2/2+j[\phi_{2,k}+\gamma_{1,k}\phi_{1,k}])t^2]e^{j2\pi k f_0 t}. \end{aligned} \tag{3}$$

2.2. Model definition

As the approximated signal in eq. (3) has a second order polynomial with complex coefficients, we propose to model the speech signal in eq. (1) by:

$$\hat{s}(t) = \sum_{k=-K}^{K} (a_k + b_k t + c_k t^2)e^{j2\pi k f_0 t}, \quad t = -N, ..., N \tag{4}$$

where, as before, K is the number of harmonics and f_0 is the local fundamental frequency, while $\{a_k, b_k, c_k\}_{k=-K}^{K}$ are complex coefficients which contain both amplitude and phase/frequency information.

2.3. Time-domain Properties

From eq. (4), the instantaneous amplitude for each component is a time-varying function given by:

$$\begin{aligned} m_k(t) &= |a_k + b_k t + c_k t^2| \\ &= \sqrt{(a_k^R + b_k^R t + c_k^R t^2)^2 + (a_k^I + b_k^I t + c_k^I t^2)^2}, \end{aligned} \tag{5}$$

where x^R and x^I are the real and the imaginary parts of x, respectively. The instantaneous phase for each component is given by:

$$\phi_k(t) = 2\pi k f_0 t + atan\left(\frac{a_k^I + b_k^I t + c_k^I t^2}{a_k^R + b_k^R t + c_k^R t^2}\right). \tag{6}$$

Finally, the instantaneous frequency is obtained by differentiating the continuous instantaneous phase:

$$\begin{aligned} f_k(t) &= \frac{1}{2\pi}\phi_k'(t) \\ &= k f_0 + \frac{1}{2\pi}\frac{(a_k^R b_k^I - a_k^I b_k^R) + 2t(a_k^R c_k^I - a_k^I c_k^R) + t^2(b_k^R c_k^I - b_k^I c_k^R)}{m_k^2(t)} \end{aligned} \tag{7}$$

A feature of the model worth noting is that the second term of the instantaneous frequency depends on the instantaneous amplitude.

2.4. Towards the target model

Following the same idea as in a previous work [7], we decompose b_k and c_k into two components one collinear and the other orthogonal to a_k, yielding

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{8}$$

and

$$c_k = \sigma_{1,k} a_k + \sigma_{2,k} j a_k, \tag{9}$$

where $\rho_{1,k}$, $\rho_{2,k}$, $\sigma_{1,k}$, and $\sigma_{2,k}$ are the projections of b_k and c_k onto a_k and ja_k, respectively. Mathematically, the projections are given by:

$$\begin{aligned} \rho_{1,k} &= \tfrac{a_k^R b_k^R + a_k^R b_k^R}{|a_k|^2}, \quad \rho_{2,k} = \tfrac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2}, \\ \sigma_{1,k} &= \tfrac{a_k^R c_k^R + a_k^R c_k^R}{|a_k|^2}, \quad \sigma_{2,k} = \tfrac{a_k^R c_k^I - a_k^I c_k^R}{|a_k|^2}. \end{aligned} \tag{10}$$

With this notation, eq. (4) can be rewritten as:

$$\hat{s}(t) = \sum_{k=-K}^{K} a_k \left[1 + (\rho_{1,k} + j\rho_{2,k})t + (\sigma_{1,k} + j\sigma_{2,k})t^2\right] e^{j2\pi k f_0 t}. \tag{11}$$

Finally, from eq. (3) and (11), an estimate of the kth chirp component parameters can be obtained as follows:

$$\begin{cases} \hat{A}_k = |a_k| \\ \hat{\phi}_{0,k} = \angle a_k \\ \hat{\gamma}_{1,k} = \rho_{1,k} \\ \hat{\phi}_{1,k} = \rho_{2,k} \\ \hat{\gamma}_{2,k} = \sigma_{1,k} + \rho_{2,k}^2/2 \\ \hat{\phi}_{2,k} = \sigma_{2,k} - \rho_{1,k}\rho_{2,k} \end{cases}. \tag{12}$$

3. ESTIMATION

3.1. Least square estimate

Let us consider a speech signal $s(t)$ at time instants $t_{-N}, ..., t_N \in [-t_0, t_0]$ and let us assume that both the model order K and an estimate of the local fundamental frequency f_0 are known. The estimation of the complex quantities $\{a_k, b_k, c_k\}_{k=-K}^{K}$ is then done through Least Squares. In matrix form, the solution is given by:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W \mathbf{s}, \quad (13)$$

where $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{s}$ are the vectors constructed from a_k, b_k, c_k and $s(t)$, respectively. W is a diagonal matrix whose diagonal elements are the weights typically defined by a Hamming window [9]. Finally, $E = [E_0|E_1|E_2]$ where the E_i submatrices for $i = 0, 1, 2$ are given by:

$$E_i = \begin{bmatrix} t_{-N}^i e^{j2\pi(-K)f_0 t_{-N}} & \dots & t_{-N}^i e^{j2\pi K f_0 t_{-N}} \\ t_{-N+1}^i e^{j2\pi(-K)f_0 t_{-N+1}} & \dots & t_{-N+1}^i e^{j2\pi K f_0 t_{-N+1}} \\ \vdots & \vdots & \vdots \\ t_N^i e^{j2\pi(-K)f_0 t_N} & \dots & t_N^i e^{j2\pi K f_0 t_N} \end{bmatrix} \quad (14)$$

It is important to note that the length of the window should be at least 3 pitch periods to avoid matrix ill-conditioning. We chose a length of 4 pitch periods, providing a balance between overfitting and underfitting.

3.2. Iterative Estimation

Once the phase parameters $(\hat{\phi}_{1,k}, \hat{\phi}_{2,k})$ of the signal have been estimated using eq. (12), they can be used to define a new basis for subsequent signal analysis. Hence we suggest an iterative procedure where, at each iteration, the signal is modeled by:

$$\hat{s}(t) = \sum_{k=-K}^{K} (a_k + b_k t + c_k t^2) e^{j(2\pi k f_0 t + \hat{\phi}_{1,k} t + \hat{\phi}_{2,k} t^2)}, \quad (15)$$

where a_k, b_k and c_k are again complex coefficients estimated by the Least Squares method. Technically, the new basis is plugged into the complex exponentials of (14). This procedure is repeated until convergence occurs, i.e. once a criterion (e.g. based on the evolution of the LS error) is satisfied.

4. RESULTS

4.1. Synthetic chirp signals

The proposed time-varying model is first applied to a synthetic mono-component chirp signal whose amplitude varies according to a second order polynomial. The chirp signal is given by

$$x(t) = (1 - 100t + 10^4 t^2) e^{j2\pi(400t + 3000t^2 + 10t + 0.01)}$$

and thus has a chirp rate of $6000Hz/s$. The analysis was performed at $f_0 = 400Hz$ while the instantaneous frequency at the center of the analysis window is $410Hz$. The upper panels of Fig. 1 show the real part of the original chirp signal and the real part of the reconstructed signal as well as their instantaneous frequencies. In the lower panels, the iterative scheme has been applied and after 2 iterations the instantaneous frequency has been correctly estimated.

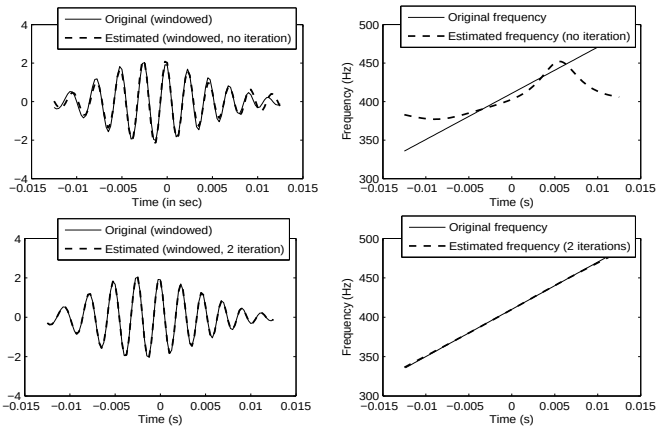

Fig. 1. Mono-component synthetic chirp signal. Left panels: original and reconstructed signals. Right panels: original and estimated instantaneous frequencies.

Another example is given on a synthetic signal consisting of ten harmonically related chirp components. The results presented in Table 4.1 show the potential of the proposed iterative scheme to correctly identify the chirp rates.

harmonic	chirp rate	1 iter	10 iter
1st	200	55	200
2nd	400	194	400
3rd	600	353	599
4th	800	457	800
5th	1000	515	1000
6th	1200	525	1199
7th	1400	519	1400
8th	1600	490	1600
9th	1800	456	1799
10th	2000	441	2000

Table 1. Multi-component synthetic chirp signal: estimated chirp rates in Hz/s after 1 and 10 iterations.

4.2. Real Speech

One female voice is analyzed in Fig. 2. The sampling frequency of the signal was 16kHz and the number of harmonics was set to 5. In this example, after careful manual inspection of the evolution of the glottal cycle, it was observed that within the analysis window, the fundamental frequency approximately decreases from 250Hz to 210Hz. From Fig. 2, it can be seen that the estimated f0 tracks can be recovered, while the other harmonics also exhibit plausible frequency variations.

It must be pointed out however that for speech signal the chirp rate is larger for higher harmonics. Consequently, there may be cases in which the Taylor approximation in eq. (2) is not valid. To handle such cases, it is recommendable to use a single fan-chirp rate α estimated from only the first K_0 components. Then, the analysis is carried using chirp rate $k\alpha$ for the kth harmonic. A weighted average of the chirp rate of each component, $\hat{\alpha} = \frac{1}{K_0} \sum_{k=1}^{K_0} \hat{\phi}_{2,k}/k$,

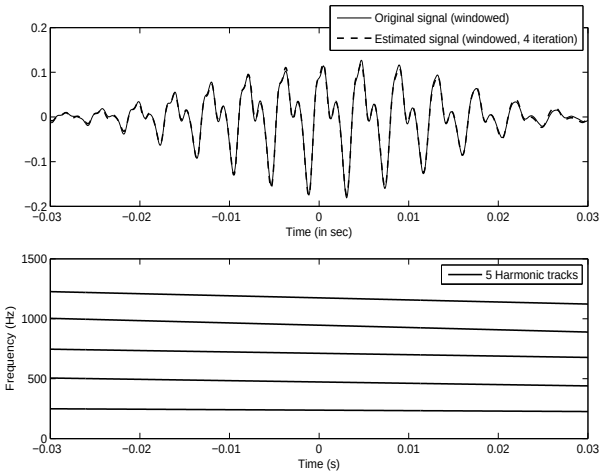

Fig. 2. $60ms$ of female speech. Upper panel: Original signal and the reconstructed signal (SNR = $15.7dB$). Lower panel: The estimated frequency evolution of the 5 first harmonics.

is used as an estimate of the single chirp rate, α. We apply this strategy to the analysis of the /vazivaza/ speech signal depicted on Fig. 3 together with its fundamental frequency contour. The estimation of the fundamental frequency and of the number of harmonics is done using time-domain (autocorrelation-based) and frequency-domain methods respectively as in [9]. Then, the harmonic part of the speech segment is analyzed pitch synchronously with a two pitch period window. Fig. 4 shows that the proposed analysis procedure is able to capture harmonic trajectories that are, in most cases, continuous. Note that K_0 is set to 4 in this example.

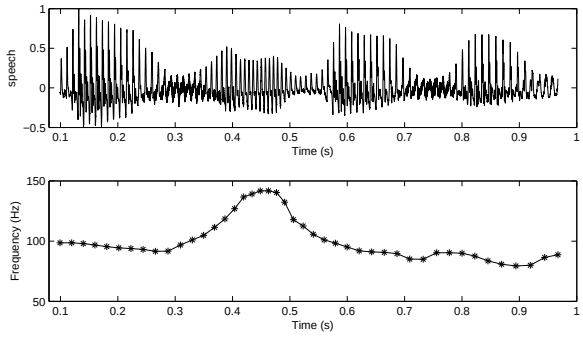

Fig. 3. Speech signal /vazivaza/ and the corresponding pitch contour.

5. CONCLUSION

We have presented a novel model for the analysis of the deterministic part of speech which, for each sinusoidal component, captures modulations in both amplitude and frequency. The ability of the model to gradually estimate and account for the linear evolution of the component frequencies was shown. Forthcoming works will be dedicated to the application of the proposed time-varying speech model in speech synthesis and speech modification.

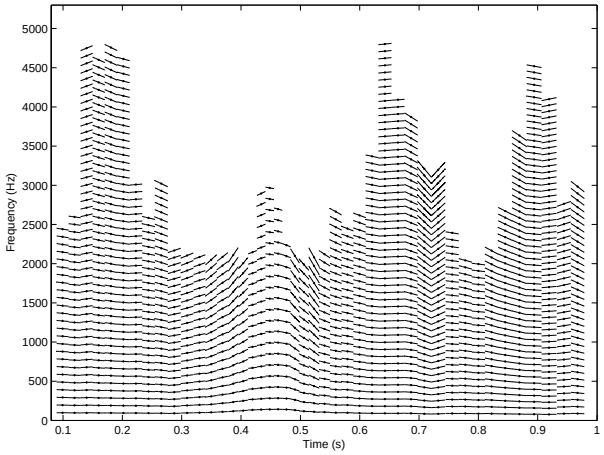

Fig. 4. Analysis of harmonic part of a speech signal /vazivaza/.

6. REFERENCES

[1] J. Laroche, Y. Stylianou, and E. Moulines. HNM: A Simple, Effecient Harmonic plus Noise Model for Speech. In *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 169–172, New Paltz, NY, USA, Oct 1993.

[2] Leon Cohen. *Time-Frequency Analysis*. Prentice Hall, New York, 1995.

[3] David Vakman. On the Analytic Signal, the Teager-Kaiser Energy Algorithm, and other methods for defining Amplitude and Frequency. *IEEE Trans. on Signal Processing*, 44:791–797, 1996.

[4] James F. Kaiser. On a Simple Algorithm to Calculate the 'Energy' of a Signal. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–384, Albuquerque, Apr 1990.

[5] L. Weruaga and M. Kepesi. The Fan-chirp Transform for Non-stationary Harmonic Signals. *IEEE Trans. on Signal Processing*, 87:1504–1522, 2007.

[6] R. Dunn and T. F. Quatieri. Sinewave Analysis/Synthesis based on the Fan-Chirp Transform. In *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 16–19, Oct 2007.

[7] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Interspeech*, pages 1044–1047, Brisbane, Sep 2008.

[8] Jean Laroche. A new analysis/synthesis system of musical signals using prony's method. application to heavily damped percussive sounds. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 2053–2056, Glasgow, UK, May 1989.

[9] Yannis Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, 1996.