SINUSOIDAL MODELING OF VOICED AND UNVOICED SPEECH

Jorge S. Marques and Luis B. Almeida

INESC, R. Alves Redol n° 9-2, P-1000 Lisboa, Portugal

Abstract
Speech modeling is an active research area which has received considerable attention in recent years. Application areas include speech coding, recognition and synthesis. Several models have been proposed both in the time and frequency domain contexts. Early frequency domain models approximated the voiced segments by locally periodic signals. The generalized harmonic model was able to deal with departures from periodicity resulting from vocal tract variation, but had difficulty in handling fast pitch variations. Sinusoidal models were developed to overcome these limitations. In this paper, we briefly review the basic concepts related to sinusoidal modelling of speech and present the results of an extensive experimental comparison between several sinusoidal modeling variants.

1. INTRODUCTION
Speech modeling is an active research area which has received considerable attention in recent years. Application areas include speech coding, recognition and synthesis. Several models have been proposed in recent years, both in the time and frequency domain contexts. Early frequency domain models [12] approximated the voiced segments by locally periodic signals. The generalized harmonic model [1] was able to deal with departures from periodicity resulting from vocal tract variation, but had difficulty in handling fast pitch variations [2]. Sinusoidal models [2,4,5,6,9] were developed to overcome these limitations. They are based on the idea of representing the speech signal, within each frame, by a superposition of sinusoids with continuously varying amplitudes and frequencies. They are known to be able to reproduce voiced speech with high quality, and have also been extended to the representation of unvoiced sounds. In this paper, we briefly review the basic concepts related to sinusoidal modelling of speech. We then present the results of an extensive experimental comparison between several sinusoidal modeling variants.

This paper is organized as follows: section 2 briefly reviews sinusoid based models; section 3 describes several parameter estimation algorithms; section 4 presents an experimental evaluation of several sinusoidal model variants, and section 5 concludes.

2. SINUSOIDAL MODELS OF SPEECH
Sinusoid based models generally represent the speech signal, within each frame, by a superposition of sinusoids with time-varying amplitudes and frequencies. The sinusoids which are used to synthesize the signal are speech dependent and vary from frame to frame. However, sinusoids in consecutive frames are not independent since it is important to keep smooth transitions of the amplitudes, frequencies and phases of the sinusoids at the frame boundaries, at least in voiced speech segments. Two kinds of sinusoidal model will be considered in this paper. A widely used model which will be called basic sinusoidal model (BSM), and an alternative model where each sinusoid is split into two quadrature components: the so-called quadrature sinusoidal model (QSM).

A - Basic Sinusoidal Model [9]
In the basic sinusoidal model, the speech signal is divided into frames of length T and approximated in each frame by a sum of sinusoids

\[ s(t) = \sum_{k=1}^{n} a_k(t) \cos(\theta_k(t)) \quad 0 < t < T \]  

(1)

where n is the number of sinusoids, \( a_k(t) \) is the amplitude of the k-th sinusoid and \( \theta_k(t) \) is its phase, the time-varying frequency of the k-th sinusoid being given by the derivative of the phase

\[ \phi_k(t) = \dot{\theta}_k(t) \]  

(2)

A popular approach to describe the evolution of the amplitude and phase is the use of a first order polynomial for the amplitude and a third order polynomial for the phase [2, 9]

\[ a_k(t) = A_k + B_k \]  

(3)

\[ \phi_k(t) = c_3 k^3 + c_2 k^2 + c_1 k + c_0 \]  

(4)

This choice provides an interpolation rule for the instantaneous values of amplitude, frequency and phase between frame boundaries. If we denote by \( a_0, a_1, a_0, a_1, \theta_0, \theta_1 \) respectively, the parameters of the interpolation laws are easily obtained after some algebra

\[ \begin{bmatrix} a_0 \\ \theta_0 \\ a_1 \\ \theta_1 \end{bmatrix} = \frac{1}{T^3} \begin{bmatrix} 3 & 1 \\ T & 2 \end{bmatrix} \begin{bmatrix} 0 \\ \phi_0 \end{bmatrix} \]  

(6)

where

\[ \phi = \phi_0 - \left( \theta_0 + \frac{\phi_0 + \phi_1}{2} \right) T \]  

(7)

\[ \phi_0 = \phi_1 + \phi_0 \]  

(8)

\[ \phi_0 \] we have dropped the order of the sinusoid for simplicity
can be interpreted as a phase and a frequency prediction error. When the signal is periodic both the phase and frequency prediction errors are zero and the phase interpolation law becomes a straight line.

There is one difficulty which may not be apparent, in the evaluation of the prediction error in equation (7). Only the principal value of the phase is usually obtained from estimation algorithms. If $\Phi_0, \Phi_1$ are the principal values of the phase at the frame boundaries then

$$\epsilon_\Phi = \Phi_1 - \Phi_0 + 2\pi m - \frac{\phi_0 + \phi_1}{T}$$

and to evaluate the prediction error we must unwrap the phase, determining the integer $m$. The value of $m$ is usually obtained by forcing the prediction error to be in the interval $[-\pi, \pi]$. Therefore, $m$ is the integer closest to

$$m^* = \frac{1}{2\pi} \left[ \Phi_1 - \Phi_0 - \frac{\phi_0 + \phi_1}{2\pi} + \frac{T}{2} \right]$$

B - Quadrature Sinusoidal Model [7]

In the quadrature sinusoidal model each sinusoid is split into two quadrature components

$$s(t) = \sum_{k=1}^{n} a_k(t) \cos \psi_k(t) + a_k(t) \sin \psi_k(t), \quad 0 < t < T$$

The two components of each pair have independent amplitudes $a_{k1}(t)$, and $a_{k2}(t)$ but share a common phase $\psi_k(t)$. The amplitudes are usually described by first order polynomials on $t$ and the phase by a second order polynomial [6,7]

$$a_k(t) = A_k + B_k t$$

This allows to specify the sinusoid parameters from the values of the amplitudes and the frequency at the frame boundaries. The phase parameters, for instance, are given by

$$d_1 = \alpha_0$$

$$d_2 = \frac{\phi_0 - \phi_1}{2T}$$

parameter $d_0$ being determined from a phase continuity condition at the beginning of the frame. In this model there is no explicit evaluation of the phase of the sinusoids from the speech signal. Instead, phase matching between model and signal is achieved by the relative amplitudes of the two quadrature components of each pair.

A - Frequency Estimation

Frequency estimation is a delicate aspect of sinusoidal modeling since it has a strong influence on the perceptual quality of the speech produced by these models. In voiced speech segments, the spectrum of the speech signal usually consists of spectral lines approximately located at multiples of the fundamental frequency and it is important to guarantee a close matching between the location of the spectral lines and the frequencies of the sinusoids. Although some frequency deviations can be tolerated on the synthesis stage without significant distortion, the estimation of amplitudes and phases is strongly affected by frequency errors. In unvoiced speech segments, the role of sinusoid frequencies appears to be different. Experiments performed with unvoiced fricatives and synthetic noise signals suggest that the density of frequencies along the frequency axis is more important than the exact location of each frequency. Below, the two main frequency estimation algorithms are briefly described.

A.1 Peak Picking Algorithm [9]

The peak picking (PP) algorithm evaluates the frequencies of the sinusoids by detecting the local maxima of the short-time spectrum of the speech signal, computed at the frame boundaries [9]. If the number of peaks exceeds the maximum number of frequencies allowed in the model, the smallest peaks are discarded. In the experiments presented in this paper, the speech signal is windowed with an adaptive Hanning window and the short-time spectrum is computed using a 512 point FFT. Figure 1 shows the location of the analysis windows used to estimate the frequencies at the beginning and the end of the synthesis frame.

A.2 Harmonic Frequency Estimation [7]

In voiced speech segments, the short-time spectrum of the speech signal exhibits an harmonic structure, i.e. a line structure with spectral lines approximately located at multiples of the fundamental frequency. This structure has traditionally been exploited, in the so-called Harmonic Models [1,2,4], by assuming that the frequencies of the sinusoids are multiples of the fundamental frequency (we shall call this harmonic restriction). This restriction leads to very good frequency estimates provided that the fundamental frequency is accurately evaluated. In addition, the description of all frequencies by a single parameter is very useful in coding applications, where the total number of parameters to be transmitted should be as low as possible.

Since the exact locations of the sinusoids, in unvoiced speech segments, appears to be secondary, the main concern being to guarantee a high enough "density" of sinusoids across the whole spectrum, the use of harmonically related frequencies has been extended to unvoiced frames using the maximum number of sinusoids allowed in the model, distributed from zero to half the sampling frequency [7]. The frequencies of the sinusoids are therefore given by

voiced frames: $\phi_k = k \phi_0$
unvoiced frames: $\phi_k = k \frac{\phi_0}{2n_{\text{max}}}$

where $\phi_0$ is the fundamental frequency, $\phi_k$ is the sampling frequency and $n_{\text{max}}$ is the maximum number of sinusoids allowed in the model.

this means that, in coding applications, amplitudes and phases should be evaluated using unquantized frequency estimates.

usually, a separate pitch detector is employed for this purpose [3]
B - Matching Algorithm [7]

Since the estimation of instantaneous frequencies at the beginning of the frame is independent from the estimation at the end, one must specify the correspondence of frequencies between frame boundaries. In fact, the problem is slightly more complex, since one must allow frequencies to be created or to disappear within each frame. A matching algorithm to perform this task is described in [9]. It basically tries to pair frequencies at both boundaries according to a nearest neighbor criterion, i.e. minimizing

$$d_{kn} = |\theta_{0k} - \theta_{1m}|$$

where $\theta_{0k}$ is a frequency estimated at the beginning of the frame and $\theta_{1m}$ a frequency estimated at the end. The matching of the frequencies is not allowed when their difference exceeds a given threshold. Frequencies which are not matched by this algorithm are said to correspond to the birth or death of sinusoids within the frame.

The use of the matching algorithm is essential when peak picking is used, but it can also enhance the robustness when the frequencies are harmonically related, making the model less sensitive to errors in the estimation of the fundamental frequency (e.g. pitch doubling and halving). However, wrong frequency matches are often observed in voiced frames, since the algorithm as given above does not take into account the variation of the fundamental frequency (figure 2a). This difficulty can be easily overcome by the use of an expansion factor $\alpha$ in the distance expression to account for pitch variation [7,10].

$$d_{kn} = |\alpha \theta_{0k} - \alpha \theta_{1m}|$$

with $\alpha = \theta_{1f}/\theta_{0f}$

where $\theta_{0f}, \theta_{1f}$ are the fundamental frequency estimates at the frame boundaries. When using the expansion factor, care must be taken to avoid the effect of pitch errors. This can be accomplished by limiting this factor to a specified range of values (e.g. $[1/\sqrt{2}, \sqrt{2}]$).

```
Figure 2 - Matching of Frequencies in a Voiced Frame
  a) without expansion factor; b) with expansion factor
```

C - Amplitude and Phase Estimation

Two algorithms have been widely used in the estimation of the amplitudes and phases of the sinusoids. One of them uses the short-time spectrum of the speech signal computed with windows centered at the frame boundaries (figure 1) [4,9]. The amplitudes and phases of the sinusoids are obtained by sampling the magnitude and phase of the spectrum at the sinusoid frequencies. This algorithm will be called spectral sampling, (SS) algorithm.

Another estimation technique tries to approximate the speech signal by a sum of sinusoids with constant amplitudes and frequencies in the vicinity of the frame boundaries. The frequencies of the sinusoids are assumed to be known and the complex amplitudes are evaluated by a squared error criterion leading to a set of linear equations (table I) [1]. Since this algorithm performs the estimation using a stationary sinusoidal model around each boundary, it will be denoted by stationary least squares (SLS).

To understand the relationship between both algorithms, one can note that spectral sampling can also be interpreted as the solution of a least squares problem, if we model the speech signal by a single complex exponential at a time and evaluate its amplitude by a weighted squared error criterion (table I). Apart from the difference in the weighting, which can be compensated by an appropriate choice of the window, the spectral sampling estimation is identical to the stationary least squares algorithm except that it discards the interaction among sinusoids, i.e., it neglects the non-orthogonality of the set of sinusoids. The spectral sampling algorithm is therefore less optimal, the difference being small when the spectral lines are well separated in frequency. In terms of computational effort, both algorithms have comparable complexity since the linear set of equations can be efficiently solved [7].

One last aspect which is worth mentioning is the choice of the analysis window $w(t)$ (see table I). We have used Hamming window, centered at the frame boundaries, and with adaptive length, in both algorithms; the length of the window was 2.5 times the average pitch in voiced frames, as suggested in [9]. The choice of the length of the analysis window in unvoiced frames will be discussed later together with the experimental evaluation of sinusoidal models.

<table>
<thead>
<tr>
<th>Spectral Sampling</th>
<th>Stationary Least Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\theta_{0k} - \theta_{1m}</td>
</tr>
<tr>
<td>$\sum w(t) h(t) \theta_0(t)^2 dt$</td>
<td>$\sum w(t) h(t) \theta_0(t)^2 dt$</td>
</tr>
<tr>
<td>$a_k = \frac{\sum w(t) \theta_0(t) e^{-j\phi_k t} dt}{\sum w(t) dt}$</td>
<td>$a_k = \frac{\sum w(t) e^{j\phi_k t} dt}{\sum w(t) dt}$</td>
</tr>
</tbody>
</table>

Table I - Amplitude and Phase Estimation

The SS and SLS algorithms share many common features and specifically, they both decouple the estimation of the parameters at each end of the frame from the estimation at other end, ignoring the way amplitude and phase are actually interpolated within the frame. An exact least squares estimation, taking into account the actual amplitude and phase variation within the frame, as well as the influence of the present estimates on future frames, leads to hard optimization problems. A quasi-optimal (QO) solution [6] involves considering only a "horizon" of a limited number of frames forward, for the estimation of the parameters of the present frame. The estimation of amplitudes then becomes a manageable linear problem (though it can still be computationally heavy). In the case of the basic sinusoidal model, the estimation of phases must be done by an alternate technique (SS or SLS). The quadrature sinusoidal model replaces each amplitude-phase pair with a pair of amplitudes, and therefore needs no separate phase estimation.

4 EXPERIMENTAL EVALUATION

A systematic evaluation of sinusoidal models was performed by means of objective and informal subjective tests. With this experimental evaluation we have tried to compare the performance of all analysis techniques described above namely the peak picking and harmonic estimation of frequency and the three amplitude and phase estimation algorithms. All combinations of these techniques were tested with the
basic sinusoidal model and, in the case of the Quasi-Optimal amplitude estimation, tests were extended to the quadrature sinusoidal model as well. We have also systematically measured the effect of two design options: the use of an expansion factor in the matching algorithm, and two different strategies for the choice of the analysis window in unvoiced frames. These tests were performed using a set of 24 speech utterances (12 uttered by male speakers and 12 uttered by female speakers), corresponding to a total of 60 s of speech, sampled at 8 kHz.

In the processing of voiced frames, two test conditions were considered: use of an expansion factor in the interval $[1/2, \sqrt{2}]$, or no expansion factor. In the processing of unvoiced frames, two choices of window length were tested: an adaptive length, equal to the length used in the last voiced frame (as suggested in [9]), or a fixed, "long" window of 32 ms. In Table II, the combinations of the options above are indicated as follows: na - no expansion, adaptive window; nl - no expansion, long window; ea - expansion, adaptive window; el - expansion, long window. In all tests, a maximum of 40 sinusoids were allowed.

Table II shows the segmental SNR results for different frame lengths in the range 10-30 ms. The main overall conclusion is a severe degradation of SNR when the length is increased from 10 to 30 ms. This behaviour is partly explained by the restriction on the number of sinusoids, which remained the same for all experiments. Important differences also exist between different analysis techniques. Frequency estimation is best performed by harmonic evaluation of frequencies in voiced frames, and by peak picking in unvoiced ones. In 10 ms frames, amplitude and phase estimation is best performed by the quasi-optimal technique, the SLS ranking second and the SS algorithm ranking third. However, when the frame length increases, SLS loses its superiority over SS. Concerning the expansion factor, its use has improved the performance of the models when the frame length is greater than 10 ms. The choice of long analysis windows in unvoiced frames significantly improved the results of the peak picking frequency estimation. The effect of the frame length on the performance of the peak picking algorithm is easily understood since long analysis frames provide higher frequency resolution and therefore increase the number of sinusoids used by the model in unvoiced frames.

<table>
<thead>
<tr>
<th>SS</th>
<th>SS</th>
<th>SLS</th>
<th>SS</th>
<th>SS</th>
<th>SLS</th>
<th>SS</th>
<th>SS</th>
<th>QM</th>
<th>QQ</th>
<th>QQ</th>
<th>QQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSM BSM BSM BSM BSM BSM BSM BSM BSM BSM BSM BSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14.4 14.2 14.9 13.6 14.3 15.4 15.8 19.3 19.3 19.3 19.3 19.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.8 12.3 12.1 12.3 12.3 12.3 12.3 12.3 12.3 12.3 12.3 12.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.6 5.6 5.6 5.6 5.6 5.6 5.6 5.6 5.6 5.6 5.6 5.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

We have briefly presented the main concepts involved in the sinusoidal modeling of speech, and the main variants that exist in parameter estimation techniques. The results of an extensive experimental evaluation of these variants were given. They showed that very high synthetic speech quality can be obtained, for certain combinations of frame length and estimation techniques. They also showed that, in applications like speech coding, where the number of parameters must be reduced to a minimum, a single parameter can be used to represent all sinusoidal frequencies, both in voiced and unvoiced frames, with no degradation of the perceptual quality. The use of these models in coding applications has been discussed elsewhere [8]. Interested readers can contact the authors for obtaining software related to these models.

REFERENCES