QUASI-OPTIMAL ANALYSIS FOR SINUSOIDAL REPRESENTATION OF SPEECH

Jorge S. Marques Luis B. Almeida

INESC, R. Alves Redol, 9-2^0, 1000 Lisboa, PORTUGAL

RESUME

Ce travail présente une estimation presque optimale des amplitudes et phases des répresentations sinusoidales de la parole. En plus, pour éviter la connaissance "a priori" de la phase initiale, une nouvelle base sinusoidale est proposée. Des tests preliminaires ont montré que les phrases sintetisées par cette méthode ont une qualité transparente. Les distortions produites par d'autres méthodes d'estimation ont été complètement eliminées.

SUMMARY

This paper presents a quasi-optimal estimation for the time-varying amplitudes and phases in sinusoidal representations of speech. The estimates of the amplitudes and phases are obtained as the solution of a linear set of equations resulting from the minimization of a weighted squared error functional in a short time horizon. The optimal solution would be obtained with an infinite long time horizon, but experimental results show that time horizons of two or three frames provide very good results.

In addition, a different sinusoidal base function is proposed to avoid the need for a priori knowledge of the initial phase of the sinusoids in each frame. Preliminar experimental results have shown, so far, that synthetic speech obtained with this method is of transparent quality. Distortions produced by previous estimation methods have been completely eliminated.

1. Introduction

Several sinusoidal representations have been proposed in recent years for speech coding and transformation purposes. Some of these methods were developped in a time domain context [4] but other ones were born in a frequency domain framework and are related to other frequency domain techniques. The harmonic model (HM) [1] was developped to improve the performance of adaptive transform coding (ATC) [3] in voiced segments of speech. Since the spectrum of a voiced segment has a harmonic structure, there is a strong correlation among the spectral samples belonging to the same harmonic. It is therefore more efficient to model the spectrum as a sum of lines of a known shape than to code each sample separately. In the original HM [1] the analysis and the synthesis are done in the frequency domain using the Short-Time Fourier Transform (STFT) and the synthetic signal is obtained by the overlap-add method.

The incapacity of this model to reproduce harmonics with fast varying frequencies, when the frame rate was as low as 30 Hz, was soon noticed [2]. To overcome this problem a new time domain synthesis method was proposed which used an interpolation of the parameters of the sinusoids between consecutive frames [2]. Specifically, the speech signal is approximated using a cubic interpolation for the phases and a linear interpolation for the amplitudes of the sinusoids. The analysis is still done using the harmonic model. The quality of the voiced speech obtained by this method is very high.

The sinusoidal representation of speech was extended in [5], [6] to unvoiced segments by dropping the use of a pitch estimator and the restriction on the center frequencies of the sinusoids, which are no longer multiples of a fundamental frequency. Without this restriction it is not obvious how to match the sinusoids of a given frame with the sinusoids of neighbouring frames. An algorithm was proposed in [5] to match the sinusoids of consecutive frames, allowing the sinusoids to be born and to die. The analysis is done in the frequency domain by computing the spectrum and picking its peaks, and the synthesis is done in the time domain using the same interpolation for the phase and the amplitude as in the HM. From our simulations, we concluded that the quality of the speech produced by this analysis/synthesis method is very high in voiced frames but it is much inferior in unvoiced frames, where it presents tonal noises.

Both methods use the same kind of interpolated sinusoids to synthesize speech but neither of the models used in the analysis is optimized taking into account the synthesis model. In this paper, we propose an analysis procedure which is consistent with the synthesis model enabling, in principle, the optimal estimation of the amplitudes and phases of the sinusoids. The optimal solution for the estimation of the parameters is shown to be a least squares problem with restrictions, demanding the knowledge of the whole signal and the simultaneous estimation of the amplitudes of all the sinusoids in all the frames. This solution is impractical, even for small utterances, since it represents an enormous number of parameters to estimate simultaneously, and an unacceptable delay. Fortunately, far away segments have a negligible effect on the estimation at a given frame and the temporal optimization horizon can thus be reduced to a small number of frames, bringing the computational complexity of the method to reasonable values. Another simplification consists of the definition of auxiliary base functions which generate the same space as the sinusoids with interpolated amplitude, but which remove the restrictions among the parameters to be estimated and therefore reduce the dimension of the problem.

2. Sinusoidal Decomposition

Let us define $s^j(t)$ as the j-th frame of the speech signal $s(t)$ shifted to the origin

$$s^j(t) = \begin{cases} s(t-jT+T) & 0 \le t < T \\ 0 & \text{otherwise} \end{cases} \qquad (2.1)$$

where T is the length of the synthesis frame. The purpose of the sinusoidal decomposition is to approximate $s^j(t)$ (j=1,2, ... ,L) by

$$s^j(t) = \sum_{k=1}^{N_j} o^j{}_k(t) \qquad j=1,2,\ldots,L \qquad (2.2)$$

where $o^j{}_k(t)$ are the sinusoidal base functions and N_j is the number of sinusoids used in frame j.

The sinusoidal decompositions used in [2], [5] consider sinusoids with a cubic interpolation for the phase and a linear interpolation for the

amplitude. That is, in the jth segment the kth sinusoid is

$$o^j_k(t) = a^j_k(t)\, \varphi^j_k(t) \qquad (2.3)$$

where

$$a^j_k(t) = A^j_k + (t/T)\, B^j_k \qquad (2.4)$$

$$\varphi^j_k(t) = \cos \psi^j_k(t) \qquad (2.5)$$

$$\psi^j_k(t) = \alpha t^3 + \beta t^2 + \gamma t + \xi \qquad (2.6)$$

The sinusoid $\varphi^j_k(t)$ has unit amplitude and will be called in this paper a segmental base function of frame j (in this context the word "segment" is used in the same sense as "frame"). Each sinusoid $\varphi^j_k(t)$ is matched to a base function $\varphi^{j-1}_p(t)$ of the previous segment (except if it is "born" in the present segment), and it is also matched to a base function of the next frame $\varphi^{j+1}_m(t)$ (except if it "dies" in this segment). These matching concepts can be found in [5] and the continuity conditions to match sinusoids of different frames are the continuity of the amplitude,phase and frequency at the transition between frames

$$\begin{cases} a^j_k(0) = a^{j-1}_p(T) \iff A^j_k = A^{j-1}_p + B^{j-1}_p \\ \psi^j_k(0) = \psi^{j-1}_p(T) \\ \psi'^j_k(0) = \psi'^{j-1}_p(T) \end{cases} \qquad (2.7)$$

$$\begin{cases} a^j_k(T) = a^{j+1}_m(0) \iff A^j_k + B^j_k = A^{j+1}_m \\ \psi^j_k(T) = \psi^{j+1}_m(0) \\ \psi'^j_k(T) = \psi'^{j+1}_m(0) \end{cases} \qquad (2.8)$$

If $\varphi^j_k(t)$ is born in this segment then the envelope must be zero at the beginning of the frame

$$A^j_k = 0 \qquad (2.9)$$

and if $\varphi^j_k(t)$ is a dying base function the envelope must vanish at the end

$$A^j_k = -B^j_k \qquad (2.10)$$

These relationships defined a way of matching a sinusoid in a given frame to a sinusoid of the previous frame and to a sinusoid of the next frame except if the sinusoid is born in the frame or is dying.

We shall define a global base function $\Phi_k(t)$ as a sequence of segmental base functions φ^j_k (with unit amplitude), linked by continuity conditions and such that the first sinusoid is "born" and the last sinusoid "dies". Figure 1 shows three global base functions viewed on a set of 4 frames.

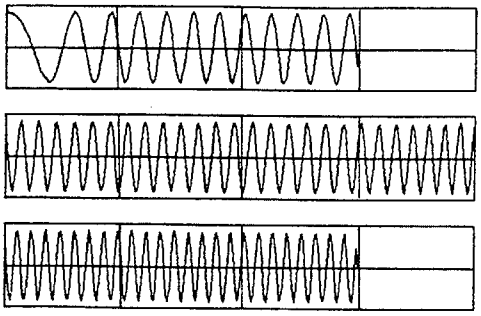

Figure 1. Three global base functions viewed on 4 frames

The sinusoidal decomposition (2.2) can now be expressed in terms of global base functions

$$s(t) = \sum_{k=1}^{Nb} e_k(t)\, \Phi_k(t) \qquad (2.11)$$

where Nb is the number of global base functions defined on the utterance and $e_k(t)$ is a piecewise linear envelope (figure 2).

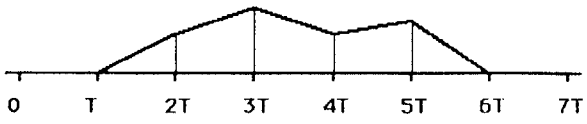

Figure 2. Piecewise linear envelope of a global base function

3. Parameter Estimation

The form of the sinusoidal decomposition was presented in section 2. We now address the problem of parameter estimation. The cost functional used to evaluate the parameters is a weighted squared error criterion

$$E = \int_0^{LT} w(t)\, |s(t) - s(t)|^2\, dt \qquad (3.1)$$

where w(t) is a weighting function, L is the number of segments and T the length of each segment as before. Using (2.2),(2.3),(2.4)

$$E = \sum_{j=1}^{L} \int_0^T w^j(t) \left[s^j(t) - \sum_{k=1}^{N_j} (A^j_k + B^j_k t/T)\, \varphi^j_k(t) \right]^2 dt \qquad (3.2)$$

$$w^j(t) = w(t - jT + T) \qquad (3.3)$$

The segmental base functions $\varphi^j_k(t)$ are assumed to be known. This means that the parameters $\alpha, \beta, \gamma, \xi$ in eqs. (2.5),(2.6) have been evaluated from the estimates of the center frequencies and phases of the sinusoids at the frame boundaries (e.g. by one of the methods presented in [2], [5]).

The minimization of the cost functional (3.1) allows the estimation of the time-varying amplitudes $a^j_k(t)$ or equivalently, the constants A^j_k, B^j_k. The number of unknown parameters is

$$N = 2 \sum_{j=1}^{L} N_j \qquad (3.4)$$

where N_j is the number of sinusoids in frame j. However, these parameters must satisfy the restictions (2.8), (2.9), (2.10), and so, the optimal estimate is the solution of a linear least squares problem with linear restrictions, which involves the simultaneous evaluation of all the parameters of the sinusoids used to synthesize the whole speech utterance. This direct optimal solution has two drawbacks:

- It is not pratical to simultaneously estimate all the parameters which model the whole utterance. This usually means the simultaneous estimation of thousands of parameters, and an unacceptable delay.
- It is a waste to estimate 2 parameters per segmental sinusoid when there is at least one restiction on these parameters

To overcome the first difficulty we note that when we want to estimate the parameters of a segment j_0 far away segments have a negligible effect on this estimation. Therefore, we propose to estimate the parameters of the segment j_0 using a time horizon of H segments to perform the minimization of (3.1) instead of using the whole utterance (figure 3).

Figure 3. Time horizon for the parameter estimation

After solving the least squares problem in the time horizon we store the parameters of frame j_0 and start to estimate the next frame shifting the time horizon forward by one frame. This solution is suboptimal but tends to the optimal solution when we increase H. In section 4 it is shown that H=3 frames is usually enough for T=10 ms and no significant improvement is obtained by considering H=5 frames.

The second drawback of the direct implementation is overcome simply by a reformulation of the problem into a simpler equivalent one. The key point is to note that a piecewise envelope (figure 1) can be obtained as the sum of triangular envelopes (figure 4).

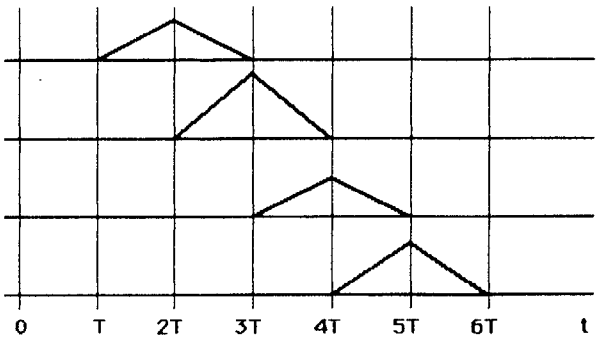

Figure 4. Triangular envelopes

It is equivalent to multiply the global base function $\phi_k(t)$ by $e_k(t)$ in (2.11) or to multiply $\phi_k(t)$ by each of the triangular functions of figure 4, with appropriate amplitudes, and add the results.

The product of a global sinusoid $\phi_k(t)$ by a triangular envelope is called an auxiliary base function $\mu_i(t)$. Assuming that the global base function $\phi_k(t)$ is born in segment b_k and dies in segment d_k, it generates $d_k - b_k$ auxiliary base functions. Figure 5 shows the auxiliary base functions generated by the first global sinusoid presented in figure 2. The amplitude of the first base function is already known from the condition of continuity with the previous segment. Thus, we can subtract its contribution from the speech signal and estimate only the amplitudes of the other auxiliary base functions.

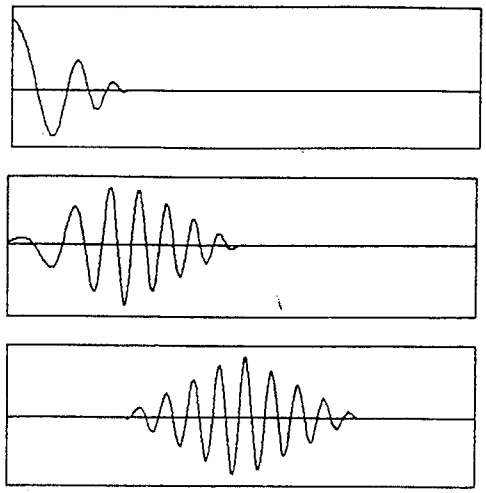

Figure 5. Auxiliary base functions for the sinusoidal decomposition

The use of these auxiliary base functions automatically guarantees that the envelopes of the global sinusoids are piecewise linear and continuous. We have therefore managed to obtain an unconstrained decomposition which generates the same space as the $o_k(t)$ with the restrictions on A_k and B_k.

The estimation will be done according to the following steps:

1- determine the auxiliary base functions from the speech signal
2 - solve a standard unconstrained least squares problem to express the speech signal in terms of the auxiliary base.

Step 2 consists of the solution of a system of linear equations

$$R\,a = r \qquad (3.5)$$

where R is the matrix of correlations between auxiliary base functions; its elements are

$$R_{ij} = \int_0^{HT} w(t)\,\mu_i(t)\,\mu_j(t)\,dt \qquad (3.6)$$

and r is the cross correlation vector with components

$$r_j = \int_0^{HT} w(t)\,s(t)\,\mu_j(t)\,dt \qquad (3.7)$$

The weighting function w(t) which was used, is constant in each frame and inversely proportional to the energy of the speech signal in that frame. The minimization of the cost functional (3.1) corresponds therefore to the minimization of the segmental SNR.

4. A New Sinusoidal Decomposition

Section 3 describes a quasi optimal solution for the piecewise linear amplitude estimation of the global base functions. The method is not restricted to the particular choice of base functions used in the decomposition. It is applicable to any set of global base functions $\phi_k(t)$ with piecewise linear envelopes.

The class of base functions which has been used in sinusoidal representations of speech was described in section 2 (eq. (2.5),(2.6)). Each sinusoid is defined by a set of parameters $\alpha, \beta, \gamma, \xi$, which are evaluated from the phase and frequency of the sinusoid at the boundary of the frame (eq. (2.7),(2.8)). The use of these sinusoids is dependent on the estimation of the frequencies and phases of the sinusoids at the segment boundaries.

Ideally, we would like to have the optimal estimates of the amplitudes,frequencies and phases of the sinusoids according to some optimization criterion. As we have shown, amplitude estimation is a linear problem, but optimal estimation of frequency and phase is much harder, and some pragmatic non-optimal solutions have been used like the estimation of the parameters by peak picking of the spectrum [5]. However, by using a slightly different set of sinusoidal base functions, the phase estimation problem can be overcome; the frequency estimation is the only nonlinear problem that still remain.

To avoid the a priori knowledge of phase we will consider a different choice for the sinusoids by replacing each sinusoid by a cosine and a sine

$$\phi^j{}_k(t) = \cos\psi^j{}_k(t) \qquad (4.1)$$

$$\phi^j{}_k(t) = \sin\psi^j{}_k(t) \qquad (4.2)$$

and adopting a quadratic evolution for the phase

$$\psi^j{}_k(t) = c_1 t^2 + c_2 t + c_3 \qquad (4.3)$$

Since the estimates of the frequency of each sinusoid at the beginning and at the end of the frame are available, we have

$$\psi'^j{}_k(0) = w_1{}^j{}_k$$

$$\psi'^j{}_k(T) = w_2{}^j{}_k \qquad (4.4)$$

and the phase evolution becomes

$$\psi^j_k(t) = (w_2{}^j_k - w_1{}^j_k)\, t^2/(2T) + w_1{}^j_k\, t + c_3 \qquad (4.5)$$

The parameter c_3 is determined by the condition of phase continuity at the transition between frames.

This new decomposition avoids the estimation of the phase of the sinusoids by using an alternative method: the independent estimation of the amplitudes of two sinusoids in quadrature with each other. The duplication of the base functions, and therefore, the duplication of the parameters to be estimated, does not mean that we have increased the total number of parameters of the sinusoidal representation, which remains the same: In the previous representation, we had to find, for each sinusoid, its frequency, phase and amplitude once per frame. These have now been replaced by one frequency and two amplitudes per frame. The two amplitudes can be optimally estimated by the method presented in section 3.

5. Experimental Results

We have implemented the sinusoidal decomposition with the base functions of section 4 with various time horizons. The increase of time horizon length H does not increase the computational effort to evaluate the correlations (3.6) since after the method is running, it is only necessary to compute new correlations for the last frame of the time horizon, the other ones being saved from the previous position of the horizon. The same reasoning applies to the cross correlations (3.7) except that they must also be computed for the first frame since we have subtracted the auxiliary functions dying in frame 1 from the speech signal (see section 3) and therefore changed the cross correlations of frame 1. But the increase of the time horizon length increases the dimension of the system (3.5) (which is solved iteratively by the Gauss Seidel Method).

Figure 6 shows the evolution of the segmental SNR as a function of the time horizon length H.

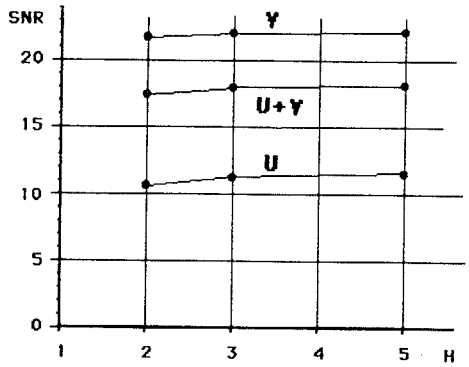

Figure 6. Segmental SNR in voiced segments (V) unvoiced segments (U) and both (U+V)

This figure was constructed using a single utterance "rice is often served in round bowls", spoken by a female speaker, and using T=10 ms and a maximum of 20 sinusoids (20 cosines + 20 sines). It shows that very good results are obtained with small values of time horizon length H.

Table I shows a comparison between our implementation of the sinusoidal decomposition of [5],[6] (MQ), and the quasi-optimal decomposition. It shows the segmental SNR for a set of 4 utterances processed by both methods using 40 sinusoids and T=10 ms. The improvement in unvoiced regions obtained by the quasi-optimal estimation procedure together with the base functions of section 4 is apparent. In voiced regions some improvement is also obtained, though smaller.

MQ	1 (M)	2 (M)	3 (F)	4 (F)	utterance
U+V	12.0	10.5	16.0	10.4	
V	19.1	16.8	21.0	16.3	
U	7.3	6.6	7.9	7.1	

quasi-opt	1 (M)	2 (M)	3 (F)	4 (F)	utterance
U+V	20.2	19.4	21.0	18.0	
V	23.1	21.3	22.6	19.0	
U	18.6	17.5	18.8	17.5	

Table I - Comparison between two sinusoidal decompositions; segmental SNR in voiced segments (V), unvoiced segments (U), and both (U+V); male speaker (M), female speaker (F)

Listening of utterances synthesized by our simulation of the method [5], [6], with 40 sinusoids and T=10 ms, shows a transparent quality in the voiced regions, but the unvoiced segments are not well reproduced and have tonal noises, showing that the estimation method [5], [6] is not adequate for unvoiced frames. With the quasi-optimal estimation method described in this paper, the synthesized speech is undistinguishable from the original, the tonal noises having disappeared completely.

6. Conclusion

We presented a method for quasi optimal estimation of the time-varying amplitudes and phases in sinusoidal decompositions. The use of convenient sinusoidal base functions together with this estimation algorithm compares favourably with other sinusoidal representations, yielding synthetic speech of transparent quality. We think that this quasi-optimal decomposition is an important step towards an efficient representation of speech by means of sinusoids.

References

[1] - L. B. Almeida, J. M. Tribolet, "Nonstationary Spectral Modelling of Voiced Speech", IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-31, pp.664-678, June, 1983

[2] - L. B. Almeida, F. M. Silva "Variable-Frequency Synthesis: an Improved Harmonic Coding Scheme", Proc. International Conf. on Acoustic, Speech, Signal Processing, San Diego, p 27.5.1, 1984

[3] - J. M. Tribolet, R. E. Crochiere "Frequency Domain Coding of Speech", IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-27,pp 512-530, Oct., 1979

[4] - P. Hedelin "A Tone-Oriented Voice-Excited Vocoder", Proc. International Conf. on Acoustic, Speech, Signal Processing, Atlanta, p. 205, 1981

[5] - R. J. McAulay, T. F. Quatieri "Mid-Rate Coding Based on a Sinusoidal Representation of Speech", Proc. International Conf. on Acoustic, Speech, Signal Processing, Tampa, p. 945, 1985

[6] - R. J. McAulay, T. F. Quatieri "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-34, pp.744-754, Aug, 1986