TOTAL LEAST SQUARES BASED SUBBAND MODELLING FOR SCALABLE SPEECH REPRESENTATIONS WITH DAMPED SINUSOIDS

Kris Hermus, Werner Verhelst, Patrick Wambacq and Philippe Lemmerling

Katholieke Universiteit Leuven - ESAT/PSI
Kardinaal Mercierlaan 94, B-3001 Leuven, Belgium

e-mail: Kris.Hermus@esat.kuleuven.ac.be

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

ABSTRACT

We describe how Total Least Squares (TLS) algorithms can be applied as a powerful and efficient modelling tool for wideband speech. A detailed description in both time domain and frequency domain illustrates how the modelling functions – damped sinusoids – naturally synthesise non-stationary signals.

Straightforward implementations of TLS applied to fullband speech are known to be computationally hard and they can suffer from numerical sensitivity.

In this paper we introduce a *subband approach*, which leads to a significant reduction of the computational load with an enhanced numerical stability. Moreover, it enables to control the distribution of the TLS components over the spectral range of the input signal such that perceptual criteria can be incorporated in the modelling scheme.

We also address the *scalability* of our design from smallband speech to high quality audio, and provide evidence for the existence of coupled components in TLS modelled segments.

1. INTRODUCTION

In *sinusoidal modelling (SM)* a short segment of speech $s(n)$ is approximated by a limited sum of constant-amplitude, constant-frequency sinusoids:

$$s(n) \approx \sum_{k=1}^{K} a_k \sin(2\pi f_k\, n + \phi_k)\ , n = 1 \ldots N \qquad (1)$$

This representation of speech has gained a lot of attention for speech analysis/synthesis, speech coding and speech modification.

Contrary to LPC analysis that performs a global fit in the spectral domain, SM models the frequency spectrum by spectrally localised contributions from the constituent sinusoids.
The SM approach is very effective to represent the harmonic structure of voiced segments in speech but it is less effective when it comes to model the aperiodic and noise-like segments that are found in transitional parts and unvoiced phonemes. This is a major drawback since often the intelligibility of speech drops with badly modelled consonants.

One way to overcome the limitations of the basic SM lies in the generalisation of the basic modelling functions, i.e. sinusoids, into *damped sinusoids* leading to the *Exponential* Sinusoidal Model (ESM):

$$s(n) \approx \sum_{k=1}^{K} a_k e^{-d_k\, n} \sin(2\pi f_k\, n + \phi_k)\ , n = 1 \ldots N \qquad (2)$$

Support is acknowledged from the Flemish Community - IWT and from the Research Fund K.U. Leuven.

This model, described by Jensen et al. in [2, 3], is indeed very effective to model transients segments, but the extraction of the model parameters (frequency, phase, amplitude, damping) is computationally intensive and numerically sensitive.

In this paper, we propose a solution to these computational problems based on a subband TLS approach. Moreover, this subband ESM model can be used to build scalable speech representations, which could form the basis for a future scalable speech and audio coder.

The outline of this paper is as follows. The next section describes the main algorithmic aspects of ESM. In section 3, a visualisation of ESM modelling in both time domain and frequency domain is given, which leads to the introduction of our subband approach in section 4. The scalability of ESM can be found in section 5. Finally, a summary and conclusions are given in section 6.

2. EXPONENTIAL SINUSOIDAL MODELLING (ESM)

2.1. Theory of TLS

Total Least Squares (TLS) algorithms form a natural extension of the basic LS algorithm, deriving the parameters of an Auto-Regressive (AR) model that *exactly* matches a (slightly) perturbated version of the input signal.

Let $s(n)$, $n = 1 \ldots N$ be one frame (typically 30 msec) from the input signal. The TLS formulation of order L can then be stated as follows.

Find the model parameters $b(l)$, $l = 1 \ldots L$ and $\Delta s(n)$, $n = 1 \ldots N$ that minimise

$$\sum_{n=1}^{N} (\Delta s(n))^2 \qquad (3)$$

subject to

$$\hat{s}(n) = s(n) + \Delta s(n) = \sum_{l=1}^{L} b(l)\,(s(n-l) + \Delta s(n-l)) \qquad (4)$$

for $n = (L+1) \ldots N$

2.2. Algorithmic Implementation

The proper arrangement of the equations in (4) into structured matrices (e.g. Hankel, Toeplitz) leads to an equivalent matrix formulation of TLS, the so-called *structured* TLS (sTLS) problem [4].

To find a solution $\hat{s}(n)$ one has to solve a non-linear optimisation problem with a high computational complexity ($\mathcal{O}(N^3)$).

Furthermore, in the case of speech, a large number of unknowns is involved, leading to poor numerical conditioning. Implementation details will not be presented in this paper; the reader is referred to [2, 4].

For the scope of this text it suffices to mention that – based on matrix characteristics – the solution $\hat{s}(n) = s(n) + \Delta s(n)$ of a sTLS problem can be decomposed in a series of damped sinusoids:

$$\hat{s}(n) = \sum_{k=1}^{K} a_k e^{-d_k\, n} \sin(2\pi f_k\, n + \phi_k) \qquad (5)$$

Throughout this paper, we assume that TLS decomposes the input signal in a *predetermined number of damped sinusoids*, each with its proper phase, frequency, amplitude and damping constants.

3. ILLUSTRATION OF ESM IN TIME AND FREQUENCY DOMAIN

3.1. Time Domain Behaviour

As stated before, TLS algorithms find – in LS sense – a slightly perturbated version of the input signal that *exactly* matches an AR model of order L.

Using the notation in (5) we define the SNR as

$$SNR = 10 \log_{10} \frac{\sum_{n=1}^{N} s^2(n)}{\sum_{n=1}^{N} (s(n) - \hat{s}(n))^2} \qquad (6)$$

which gives us a tool to measure the modelling accuracy of TLS in the time domain.

Figure 1 illustrates the time domain behaviour of TLS, both for a periodic sound (vowel) and a transient segment (plosive); 8 kHz bandwidth signals were used.

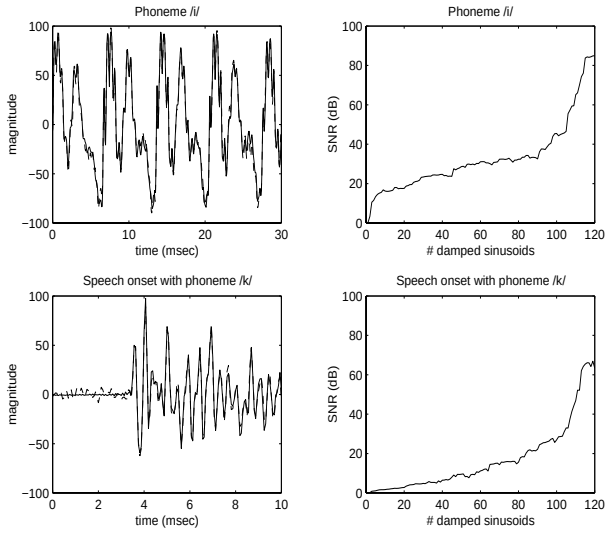

Figure 1: Illustration of TLS in the time domain. Left: original signals (solid) and TLS modelled signals (dashed). Vowel /i/ was modelled with 20 damped sinusoids; 60 damped sinusoids were used for /k/. Right: the SNR as a function of the number of damped sinusoids in the model.

We can state that TLS algorithms are capable of closely fitting speech frames, *even if strong transients are present*. The SNR not always rises with the number of damped sinusoids in the model since TLS not always converges to the global optimum.

3.2. Frequency Domain Behaviour

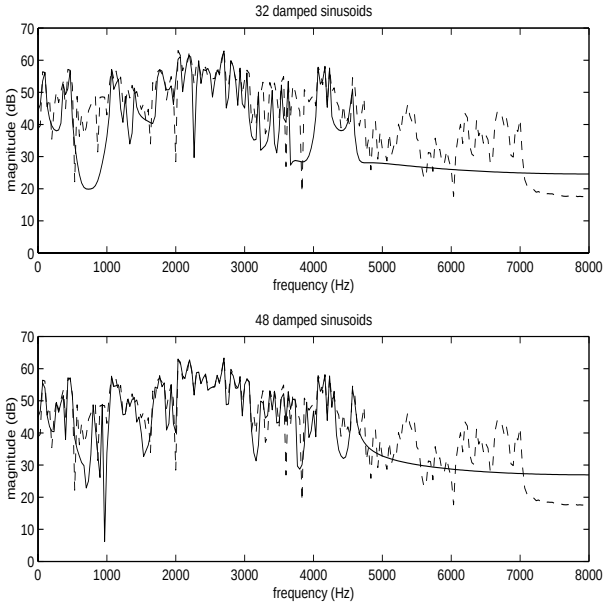

Figure 2: Illustration of TLS in the spectral domain. Original spectra (dashed) and TLS modelled spectra (solid) for phoneme /k/.

Figure 2 illustrates how TLS operates in the spectral domain. From the frame with plosive /k/ in figure 1, we plot the original spectrum (dashed) and the spectra of the TLS modelled versions (solid), for 32 (top) and 48 (bottom) damped sinusoids in the model.

One can clearly notice that TLS models the original spectrum by *spectrally localised* contributions. TLS keeps spending a lot of components to model a local high energetic region, until this local fit is 'good enough', before it goes on to model another high energetic spectral region.

The procedure of *local* fitting is opposite to the *global* fit found in LPC modelling. LPC models the spectral envelope by strongly damped sinusoids; the fine spectral detail is left in the (periodic) residual signal.

3.3. Discussion

TLS accurately models (transient) speech signals in the time domain, but at the expense of a high computational complexity of a numerically sensitive non-linear optimisation problem.

With regard to the spectral domain, TLS does not perform a global fit, as was illustrated in figure 2. This is why, up to a significant number of damped sinusoids, spectral gaps are present in the modelled speech. This is often incompatible with the criterion of optimal *perceptual* quality as required in speech (and music) coding.

4. SUBBAND APPROACH

4.1. Motivation

Drawbacks of straightforward TLS TLS needs to solve a non-linear optimisation problem with a large number of calculations, namely $\mathcal{O}(N^3)$ with N the number of samples in the frame. For fullband speech sampled at 16 kHz and a frame length of typically 30 msec, we obtain $\mathcal{O}(480^3)$ which is far too much to enable real-time operation on common hardware platforms. Recently, faster implementations of TLS have been published, but the complexity remains high.

The number of parameters in the TLS problem stated in eq. 3 and 4 increases with the number of samples N and the order L of the model (L is related to the number of damped sinusoids K in the model). In the case of speech analysis, the number of unknowns ($= L + N$) easily becomes very large ($200 - 600$). This often leads to a poorly conditioned problem that does not converge to a (optimal) solution.

When applied to speech signals, TLS has the undesired effect that a large number of damped sinusoids is needed before the whole spectral range is covered. It would be interesting to have a tool to control the distribution of the components throughout the bandwidth of the incoming signal.

Advantages of a subband approach Limiting the number of samples per frame is an effective way to reduce the number of calculations per TLS problem. This is mainly why we introduce our subband TLS approach : the downsampling at the output of the filter bank reduces the number of samples per TLS problem.

Since the number of parameters in the TLS formulation reduces with the number of samples, we also enhance the numerical stability (faster convergence towards better solutions).

Another motivation for using a subband scheme is that a more uniform distribution of the damped sinusoids over the spectral range of the speech segment can be achieved. The user is free to determine the number of components for each subband individually.
The subband approach also enables to change the modelled bandwidth by selecting/suppressing the appropriate subbands.

4.2. Implementation

Filter bank The subband filtering is implemented by a tree-structured Quadrature Mirror Filter Bank (QMF) [1]. QMF analysis and synthesis filter banks combine a perfect reconstruction with a high degree of aliasing suppression. The QMF subband signals can be fully decimated.

Reduction of Computational Complexity Our subband approach drastically reduces the number of calculations. Indeed, suppose we use m channels in the QMF filter bank (e.g. 16 filters for 16 kHz original sampling rate). By modelling subbands instead of the original fullband signal, the total number of calculations in TLS reduces to :

$$m \cdot \mathcal{O}((N/m)^3) = \mathcal{O}(N^3/m^2) \tag{7}$$

which means for e.g. 16 subbands a reduction by 256.

4.3. Experiments

Modelling performance The impact of our subband approach on the calculation time is illustrated in figure 3. For different numbers of channels in the filter bank, we plot the calculation time '$\circ$', measured in 100 msec units, needed to model a test sentence (4 seconds, 8 kHz bandwidth) with a global SNR of 30 dB.

The SNR is calculated as follows

$$SNR = 10 \log_{10} \frac{\sum_c \left(\sum_n s_{c,n}^2 \right)}{\sum_c \left(\sum_n (s_{c,n} - \hat{s}_{c,n})^2 \right)} \tag{8}$$

with $s_{c,n}$ the n^{th} sample of filter channel c, and $\hat{s}_{c,n}$ the n^{th} sample of the TLS modelled filter channel c.

Since it cannot model across the frequency range of the different channels (each component is restricted to model only its own subband), TLS loses a part of its modelling power. Hence, a subband approach will need more damped sinusoids to obtain the same accuracy as a corresponding fullband model. The number of damped sinusoids needed to obtain the global SNR of 30 dB is indicated with 'x'.

The total number of used damped sinusoids rises with the number of channels in the filter bank, but the total decrease in computation time remains significant.

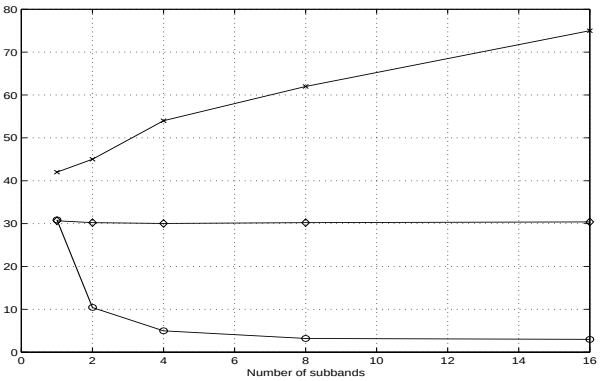

Figure 3: Computational complexity of subband TLS. As a function of the number of channels : calculation time measured in 100 msec ($\circ$), total number of damped sinusoids (x), and global SNR of the sentence ($\diamond$).

Spectral Properties of Subband TLS Using subbands enables us to *control the distribution of the damped sinusoids over the whole frequency range.* In its basic implementation, we assign to each subband an equal number of damped sinusoids. The impact on the frequency spectrum is illustrated in figure 4. We used the same speech frame as in figure 2, but in this case the signal was first sent through a filter bank with 16 output channels. Each subband is modelled with 2 (top) or 3 (bottom) damped sinusoids.

The difference in modelling behaviour compared to the fullband modelling in figure 2 is apparent : the subband approach leads to a much more uniform covering of the spectral range.

5. SCALABILITY OF THE REPRESENTATION

By omitting one or more subbands, one can readily adapt the system to the required bandwidth : Hifi (0 - 20 kHz), Wideband (50 - 7 000 kHz), Telephone (300 - 3 400 Hz).

However, another aspect of scalability lies in the *reduction of the number of modelled components.* Assume that a large number of components is used in the analysis, such that an extremely accurate modelling of speech is achieved. Some of these components will be perceptually less important than others. In coding, this

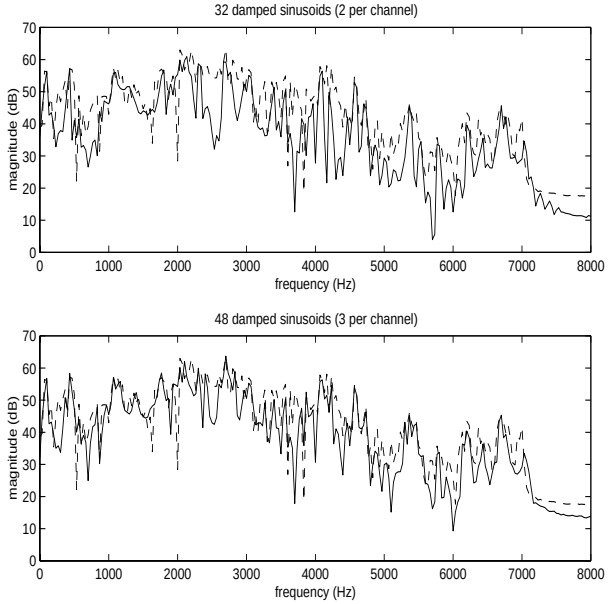

Figure 4: Spectral properties of subband TLS.

fact could be used for bit allocation and to construct a scalable speech coder: a psycho-acoustic model could be used to decide on the relative importance of the different components, and the most important components would be decoded first.

This concept was investigated in a number of preliminary experiments. Two male and two female utterances, sampled at 16 kHz, were used. The signal was split in 32 uniform frequency bands using a tree-structured fully decimated QMF filterbank. The data were analysed with 6 complex exponentials per channel[1], leading to a total of 192 complex exponentials per frame. In informal listening, we verified this to achieve transparent analysis/resynthesis quality.

We then eliminated a number of components from the modelled spectrum before resynthesis.
In a first experiment the components with smallest amplitude a_k, see eq. 5, were eliminated first.[2] This pruning strategy leads to the introduction of audible clicks and pops in the synthesis result.
In a second experiment, the components that were eliminated first were those with $\min_k(\max(a_k, a_k \exp(-d_k N)))$ and this indeed improved the results drastically: we informally judged the sound quality with 160 and 128 remaining components out of the original 192 to be good and comparable to the quality of non-pruned analysis/resynthesis with the same number of components (5 and 4 per channel, respectively).
This points to the fact that the damping factors are indeed useful in modelling speech subbands: some damping factors d_k in eq. 5 must have had a large magnitude.

[1] Here we used an implementation of TLS analysis where the number of complex exponentials needs to be specified, instead of the number of damped sinusoids. (The number of complex exponentials is about twice the number of sinusoids: $\cos(\alpha) = (\exp(\alpha) + \exp(-\alpha))/2)$, but $\cos(0) = \exp(0)$).

[2] Before eliminating components, the spectrum of each frame was equalised in 1 kHz bands to approximate equal perceptual weighting of the bands below 4 kHz and decreasing perceptual weighting above 4 kHz.

Unfortunately, we do not yet fully master the scalability problem as speech quality does not degrade sufficiently gracefully when decreasing the number of remaining model components. Rather, good quality is maintained up to some point after which it drops rapidly with important local distortions appearing. We have some indications that this could be due to the existence of *coupled components* that cannot be pruned separately (e.g., large components with phase-cancellation could model a small signal; when one of them is eliminated without the other this could lead to a severe distortion in the particular frame).

6. CONCLUSION

Total Least Squares algorithms form a powerful and flexible tool to model time signals with damped sinusoids. They are capable of closely fitting the *transient* segments in natural speech.

We introduced the *subband approach* which turns the basic TLS scheme into a feasible optimisation problem, with a significantly lower computational complexity and an improved numerical stability.
We showed how this subband based modelling enables the user to control the distribution of the TLS components over the whole spectral range.

Finally we discussed some aspects of *scalability*, in terms of bandwidth and modelling complexity. Downscaling the number of damped sinusoids revealed the existence of *coupled components*. How to deal with this in quantisation and coding schemes is an interesting aspect for future research.

7. REFERENCES

1. Bobrek M. and Koch D.B. "Music Signal Segmentation Using Tree-Structured Filter Banks", *J. Audio Engineering Society*, 46(5): 412-427, May 1998.

2. Jensen J., Jensen S.H. and Hansen E. "Exponential Sinusoidal Modeling of Transitional Speech Segments", *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. I, pages 473-476, Phoenix, U.S.A., March 1999.

3. Jensen J., Jensen S.H. and Hansen E., "Harmonic Exponential Modeling of Transitional Speech Segments", *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. III, pages 1439-1442, Istanbul, Turkey, June 2000.

4. Van Huffel S., Park H. and Rosen J.B., "Formulation and Solution of Structured Total Least Norm Problems for Parameter Estimation", *IEEE Trans. Signal Processing*, 44(10): 2464-2474, 1996.