SPECTRAL ESTIMATION FOR SPEECH SIGNALS BASED ON DECIMATION AND EIGENANALYSIS

Pirros Tsiakoulis *Student Member IEEE*, Sotiris Karabetsos *Student Member IEEE*, Stavroula-Evita Fotinea, Ioannis Dologlou

Abstract--This paper details on the application of a Decimative Spectral estimation method to speech signals in order to perform spectral analysis and estimation of Formant/Bandwidth values. The method is based on Eigenanalysis and SVD (Singular Value Decomposition) and performs artificial decimation for increased accuracy while it exploits the full set of data samples. The underlying model decomposes a signal into complex damped sinusoids whose frequencies, amplitudes, phases and damping factors are estimated. Correct estimation of Formant/Bandwidth values depend on the model order, thus the requested number of poles. Additionally, some selection criteria are applied regarding finer tracking and estimation of speech formants and their relevant bandwidths.

Index Terms-- Spectral Estimation, Formants, SVD, Decimation, Speech Processing.

I. INTRODUCTION

Various applications in the field of digital signal processing, including speech processing [1] as well as spectroscopy, i.e. quantification of NMR signals, are employing complex damped sinusoidal models in order to represent a signal segment as a sum of exponentially damped complex-valued sinusoids [2][3]. The generalized model we use is given by,

$$s(n) = \sum_{i=1}^{p} (a_i e^{j\varphi_i}) e^{(-d_i + j2\pi f_i)n} = \sum_{i=1}^{p} g_i z_i^n, n = 0,..., N-1 \quad (1)$$

where p is the number of complex damped sinusoids that comprise the measured signal, g_i the complex amplitude and z_i the signal poles. The objective is to estimate the frequencies, damping factors, amplitudes and phases.

In spectrum estimation the use of decimation has played an important role to improve the resolution of the signal under consideration, prior to its quantification. The idea is to artificially move frequency peaks apart -ensuring no aliasing- prior to parameter estimation. Conventional decimation methods used straightforward downsampling of the data, thus, reducing the available data for further processing. More modern methods overcome this inconvenience by using as many data points as possible. However, when employing the methodology we have adopted in this paper, we bring up the issue of data configuration and its importance in the overall performance. The method used here is called DESED (DEcimative Spectral Estimation by factor D) which has already been presented in [4] for decimation factor 2 and in [5] for the general case. The method performs decimation by any factor and it exploits the full data set whereas it is not obliged to reduce the dimensions of the Hankel matrix (no difference in elements of each antidiagonal) as D increases, allowing the use of dimension N/2 approximately. The advantage of DESED relies on the fact that it can benefit from the higher pole resolution obtained by decimation, while at the same time is not bound to use smaller dimensions of Hankel matrices, as other decimative approaches are. Moreover, DESED makes use of Singular Value Decomposition, while it is a generalization of the DESE2 method proposed in [5], which performs decimation by factor 2. This method, along with its TLS counterpart called DESED_TLS, have been successfully used in NMR spectroscopy, compared against methods that lie among the most promising ones for parameter estimation, that solve the same overdetermined system of equations [4].

The idea is to apply this method in the field of speech signal spectral estimation and furthermore to test if it is able to perform Formant/Bandwidth estimation. Consequently, a robust, accurate and representative parameter estimation technique can be used for feature extraction and proved to be valuable for application areas of speech synthesis and recognition.

This paper presents the application of the DESED method in the field of speech signal spectral estimation, where the problem of formants and its respective bandwidth tracking is very important. The rest of this paper is organized as follows. In section II, the main algorithmic steps and notation of the DESED method are given. Section III explains the experimental methodology and provides the obtained results for speech spectral analysis and Formant/Bandwidth estimation on synthetic and real speech signals. Moreover, examples are given concerning formant

Institute for Language and Speech Processing (ILSP), Artemidos 6 & Epidavrou, Maroussi,. GR 151 25, Athens, Greece, Tel: +30-210-6875414, Fax: +30-210-6854270
Email: ptsiak@ilsp.gr

selection criteria. Finally, a brief discussion of conclusions and further work is provided in section IV.

II. Description Of The Desed Algorithm

Let S be the $L \times M$ Hankel signal observation matrix of our N samples signal: $s(n), n = 0,1,..., N-1$ of p exponentials, where, $L-D \le M, p < L-D, L+M-1=N$ and D denotes the decimation factor. The method's algorithmic presentation follows.

STEP 1: Construct the $L \times M$ matrix S from the N data points $s(n)$ of (1).

STEP 2: Construct the matrices $S_{\downarrow D}$ and $S_{\uparrow D}$ as the D order lower shift (top D rows deleted) and the D order upper shift (bottom D rows deleted) equivalents of S. Best results are obtained when we use the $(L-D)xM$ matrices $S_{\downarrow D}$ and $S_{\uparrow D}$ as square as possible.

STEP 3: Compute the enhanced version $S_{\uparrow De}$ of $S_{\uparrow D}$ in the following way: Employ the SVD of $S_{\uparrow D}$, $S_{\uparrow D} = U_{\uparrow D} \Sigma_{\uparrow D} V^{H}{}_{\uparrow D}$ and truncate to order p by retaining only the largest p singular values.

STEP 4: Compute matrix $X = S_{\downarrow D} pinv(S_{\uparrow De})$. The eigenvalues λ_i of X give the decimated signal pole estimates, which in turn give the estimates for the damping factors and frequencies of (1).

STEP 5: Compute the phases and the amplitudes, a least squares (or a total least squares) solution to (1), with z_i replaced by the estimates and $s(n)$ given by the signal data points.

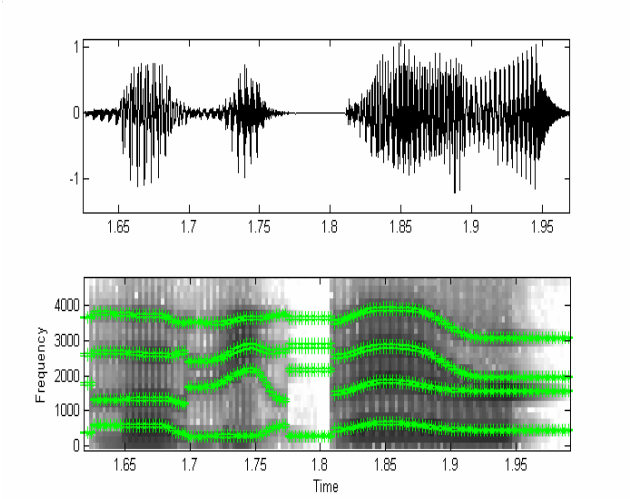

Figure 1: Reference synthetic speech signal and its spectrogram with formant trajectories.

III. Experimentation And Evaluation Of Desed On Speech Signals

The evaluation of the proposed algorithm, with respect to Formant/Bandwidth tracking of a speech signal, was based on experimental results from synthetic speech signals and real speech signals. The use of synthetic speech signals facilitated the use of a straight forward comparison criterion since their Formant/Bandwidth values are a-priori known. Synthetic signals were generated using the Klatt Cascade-Parallel Formant Speech Synthesizer [6]. The experimental methodology is common for both cases of synthetic and real speech signals. Prior to estimation, the signal is passed through a pre-emphasis filter, with a factor of -0.6, for higher frequency bands enhancement. Then, it is divided into overlapping frames with an overlap percent of 50%. The size of every frame acts as a parameter for the whole process and in most cases guarantee the inclusion of at least one pitch period of the signal. No special windowing is applied. Finally, every frame is being processed according to the steps of section II. Additionally, in the absence of noise (case of synthetic signals), preliminary experiments have indicated that using either least squares or total least squares DESED, the obtained results are rather similar. This is to be confirmed in future work with further experimentation. The experimental cases in this paper relate to the application of total least square DESED. In the case of both synthetic and real speech signals, the sampling frequency is 22.050 KHz which facilitates to use a decimation factor of 2, in order to be able to estimate formants in the frequency band of 0-5 KHz.

A. Application on synthetic speech signals

Fig. 1 depicts an example of a segment of a reference synthetic signal (theoretical values being the ones set in the Klatt synthesizer during synthetic production) and its respective spectrogram with Formant/Bandwidth pairs. Fig. 2 (a) to (c) depicts spectral analysis results for the DESED method, applied on the synthetic speech signal. The background image is the spectrogram whereas the points on it specify the apprentice estimated formants. In Fig. 2(a), the order was set to 8, namely, 4 damping sinusoids were sought for. As it can be seen, the estimated resulting sinusoids follow formant trajectories although there is some scattering leading to inconsistency for exact formant values estimation. Additionally, Fig. 2(b) and (c), shows the results when the analysis window is increased to 128 samples with an order of 8 and 12, respectively. It can be seen that the model successfully follows formant tracks but on the other hand, for some frames (for specific model order) it assigns more than one sinusoid to model high energy spectral regions. This fact results in formant values misplacement for some frames.

With the intention of finer tracking and estimation of speech formants, the technique of peak picking on high energy estimated sinusoids was applied in combination with

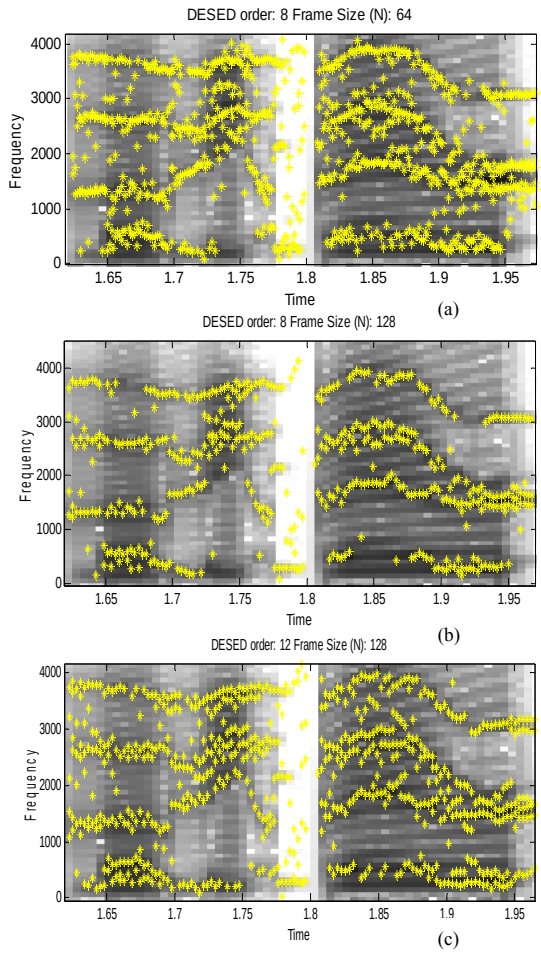

Figure 2: Formant trajectory estimation when using the DESED method on the synthetic speech signal (reference signal). Analysis results are depicted in (a) for window size N=64 and p=8, (b) for window size N=128 and p=8 and in (c) for window size N=128 and p=12.

higher model order, thus requesting more sinusoids than the expected formants. The outcome of this technique is illustrated in Fig. 3(a) and (b) for a frame size of 128 with order 32 and a frame size of 256 and order 64, respectively. From the figures, we observe that for high order model the peak picking estimates coincide with the spectrogram's formant regions, leading to promising results for the estimates of the DESED method.

This is further confirmed from the results of the frequency band chasing technique which are depicted in Fig 4, for a frame size of 256 samples and an order of 64. This technique is based on a-priory knowledge or pre-mark of possible formant locations which, in turn, define initial frequency bands of interest. Then it continuously updates these bands, based on previously estimated formant values and assuming that formant values do not change rapidly in each analysis frame. Additionally, it can be seen that the latter technique is more robust in contrast to peak picking and provides more accurate results. On the other hand, its main drawback is the need for initial frequency band definitions prior to formant estimation (e.g LPC driven).

Finally, as far as Formant/Bandwidth value estimation is concerned, Table I presents some indicative comparison results for the mean deviation from the reference synthesis parameters. The results are presented for the case of frequency bands chasing in Fig. 4. As we see, the proposed method seems to closely estimate Formant/Bandwidth values although some finer estimation has to be done, since mean deviation for the third and fourth formant increases.

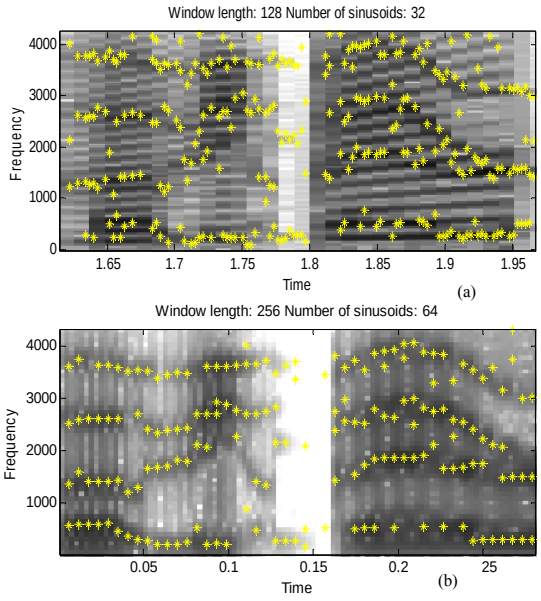

Figure 3: Formant trajectory estimation for the synthetic speech signal using peak picking. Analysis results are depicted in (a) for window size N=128 and p=32, (b) for window size N=256 and p=64.

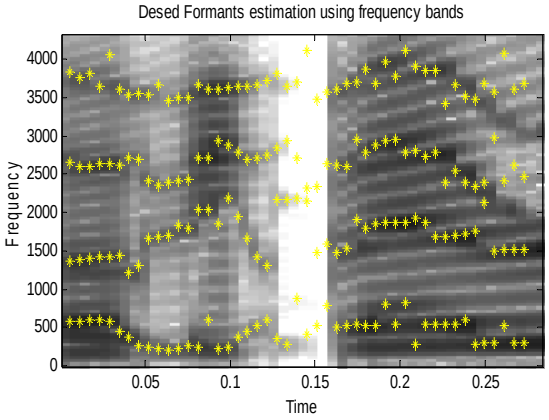

Figure 4: Formant estimation on the synthetic speech signal using the DESED method and applying frequency bands chasing (N=256, p=64).

Contrary to the results of Fig. 3(b), the latter result points out the need for fine tuning of the frequency bands chasing technique so as to decrease formant value misplacements.

Table 1: Formant/Bandwidth mean deviation values in contrast to reference parameters.

Formants	Mean Deviation DESED (Hz)
F1	62
F2	77
F3	88
F4	101
Bandwidths	Mean Deviation DESED (Hz)
B1	68
B2	47
B3	46
B4	34

B. Application on real speech signals

For the case of real speech signals, DESED method was tested on the Greek utterance 'cinoni´a' (meaning 'society') uttered from a male speaker. Fig. 5 shows the time domain speech signal followed by its spectrogram. Possible formant trajectories are signified from high energy regions on the spectrogram.

Fig. 6(a) to (c) depicts spectral analysis results for the DESED method, applied on the real speech signal. The chosen values for the experimentation parameters are identical to the ones presented in subsection A above. In all cases, there are two points of main interest. First, in contrast to Fig 5, it can be seen that formant trajectories are successfully followed although dispersion is apparent especially for high frequency formants. Second, for a given model order, the proposed method tries to model all high energy spectral regions, thus spending more than one sinusoid in some region before moving on to another. Furthermore, since real speech signals concentrate their energy on low frequency bands, the method needs higher order so as to track high frequency formants. Consequently, a selection criterion is needed to locate and extract formant values.

Fig. 7(a) and (b), illustrate the application of peak picking technique on high energy estimated sinusoids in combination with higher model order, thus requesting more sinusoids than the expected formants. The chosen values for the experimentation parameters are identical to the ones presented in subsection A above. From the figures we view that true formant trajectories become apparent but there is an inconsistency in each analysis frame due to spurious peaks of the speech spectrum which are not formants.

Moreover, Fig. 8 presents the estimated formant trajectories when the frequency bands chasing technique is used. From these results, we appreciate that the frequency bands chasing technique is more robust than peak picking but it has the drawback of a-priory definition of the initial frequency formant bands. Moreover, since the DESED method spends more than one sinusoid to model high

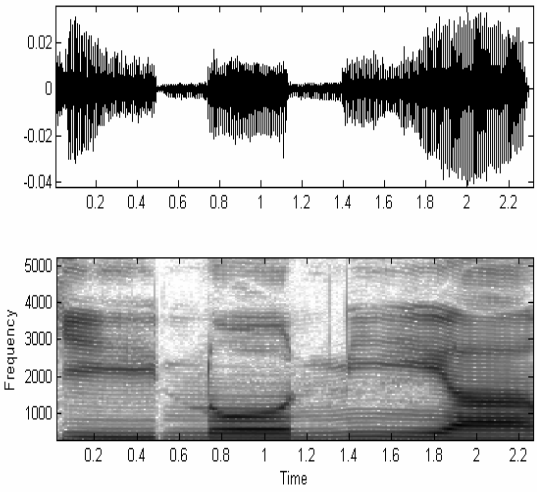

Figure 5: Reference real speech signal and its spectrogram.

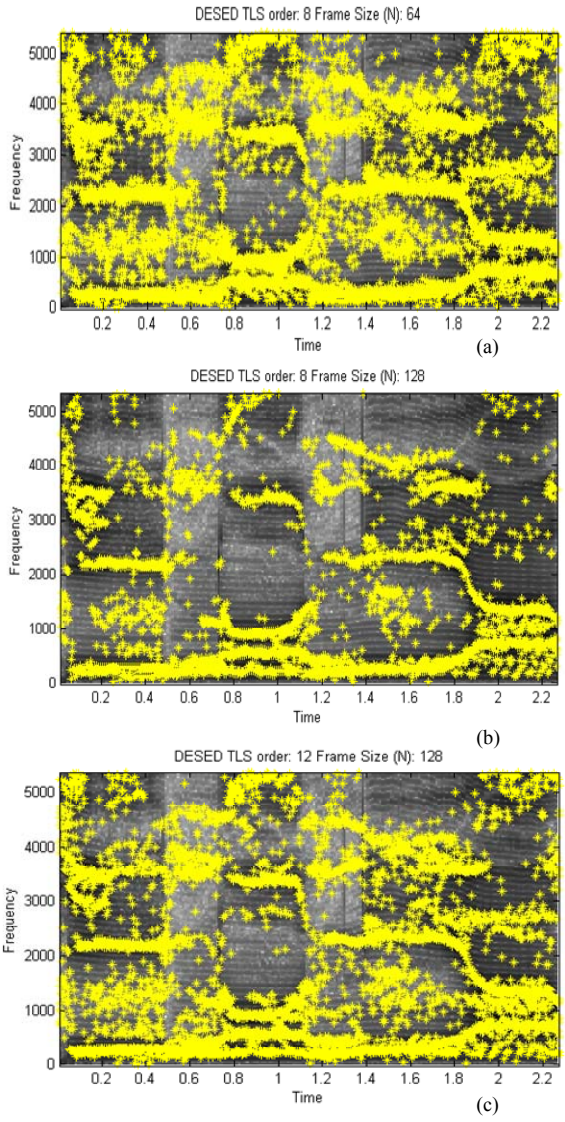

Figure 6: Formant trajectory estimation when using the DESED method on real speech signal. Analysis results are depicted in (a) for N=64 and p=8, (b) for N=128 and p=8 and in (c) for N=128 and p=12.

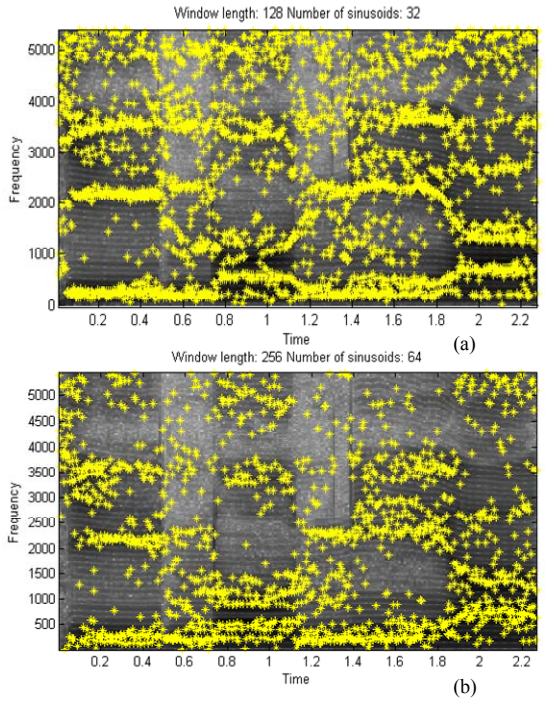

Figure 7: Formant trajectory estimation for a real speech signal using peak picking. Analysis results are depicted in (a) for window size N=128 and p=32, (b) for window size N=256 and p=64.

energy spectral regions, the corresponding formant (or pole) estimation may not necessarily be a true formant.

IV. Conclusions – Further Work

In this paper, we have exploited a high resolution spectral analysis technique that decomposes a signal into complex damped sinusoids by adjusting it to serve the need of spectral analysis and formant/bandwidth parameter tracking of speech signals. The algorithm performs artificial decimation for increased frequency resolution, while it makes use of the full data set available. The experimentation confirms that the proposed methodology successfully estimates formant trajectories and their relative bandwidths. Moreover, some ideas on finer formant/bandwidth estimation have been introduced based on a hybrid exploitation of both pattern matching as well as signal processing techniques. Finally, future work will concentrate on further development and adjustment of the presented techniques, while focusing on an efficient pure signal processing technique that could lead to a speech signal predictor, and an efficient formant/bandwidth estimation scheme.

V. Acknowledgements

This work has been partially supported by the National Technical University grant THALIS/M.I.R.C. 2002.

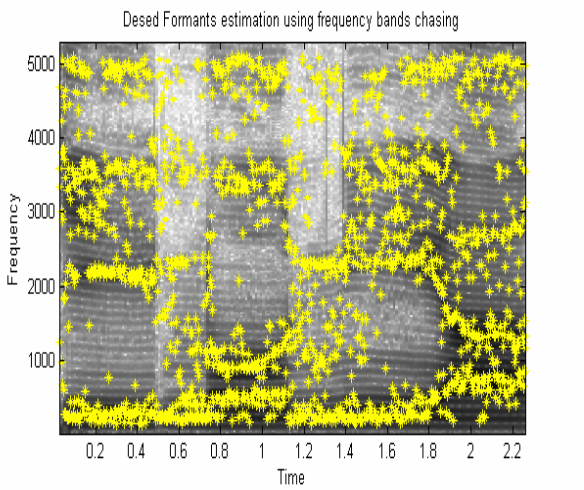

Figure 8: Formant estimation on real speech signal using the DESED method and applying frequency bands chasing (N=256, p=64).

VI. References

[1] Kumaresan, R. & Tufts, D.W. (1982), Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise, *IEEE Trans. Acoust., Speech, Signal Proc.*, 30(6):833-840.

[2] Kung, S.Y., Arun, K.S. & Bhaskar Rao, D.V. (1983), Statespace and singular-value decomposition-based approximation methods for the harmonic retrieval problem, *J.Amer.Opt.Soc*, 73(12): 1799-1811.

[3] Stoica, P. & Moses, R. (1997), *Introduction to spectral analysis*, Prentice Hall, New Jersey.

[4] Fotinea, S-E., Dologlou, I. & Carayannis, G. (2001), Decimation and SVD to estimate exponentially damped sinusoids in the presence of noise", *in Proc. ICASSP2001*, V:3073-3076, Utah, USA.

[5] Fotinea, S-E., Dologlou, I. & Carayannis, G. (2002), A new decimative spectral estimation method with unconstrained model order and decimation factor, *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*, Van Huffel, S., and Lemmerling, P. (Eds), Kluwer Academic Publishers, 321-330.

[6] Klatt, D.H. (1980), Software for a cascade/parallel formant synthesizer, *J. Acoust. Soc. Of Amer., 67:971-995*.