On the use of a Decimative Spectral Estimation Method based on Eigenanalysis and SVD for Formant and Bandwidth Tracking of Speech Signals

Sotiris Karabetsos, Pirros Tsiakoulis, Stavroula-Evita Fotinea, Ioannis Dologlou

Institute for Language and Speech Processing (ILSP)
Artemidos 6 & Epidavrou, Maroussi,. GR 151 25, Athens, Greece
{sotoskar,ptsiak,evita,ydol}@ilsp.gr

Abstract

In this paper, a Decimative Spectral estimation method based on Eigenanalysis and SVD (Singular Value Decomposition) is presented and applied to speech signals in order to estimate Formant/Bandwidth values. The underlying model decomposes a signal into complex damped sinusoids. The algorithm is applied not only on speech samples but on a small amount of the autocorrelation coefficients of a speech frame as well, for finer estimation. Correct estimation of Formant/Bandwidth values depend on the model order thus, the requested number of poles. Overall, experimentation results indicate that the proposed methodology successfully estimates formant trajectories and their respective bandwidths.

1. 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:

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

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 f_i, damping factors d_i, amplitudes a_i and phases ϕ_i, $i=1,...,p$.

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. The method used here makes use of SVD and is called DESED (DEcimative Spectral Estimation by factor D) and has 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 as D increases, allowing the use of dimension $N/2$ approximately, where N is the number of signal samples. The method, along with its TLS (Total Least Square) counterpart 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], [5]).The idea is to test these methods in speech signal spectral estimation, where the problem of formants and its respective bandwidth tracking is of special interest.

The paper is organized as follows. In section 2, we briefly present the algorithmic description of the method. Section 3, presents the experimental procedure along with the obtained results. The performance of the method is evaluated through real and synthetic voice signals. The use of synthetic voice signals facilitated the use of a straight forward comparison criterion since their Formant/Bandwidth values are a-priori known. Finally, in section 4, concluding remarks are discussed.

2. Method description

Let S be the $L\times M$ Hankel signal observation matrix of our signal $s(n), n=0,1,...,N-1$ of p exponentials, where, $L-D\leq M, p<L-D, L+M-1=N$ 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. The 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.

For the purpose of a more robust and efficient Formant/Bandwidth estimation, the above algorithm is modified to process the autocorrelation coefficients (lags) of every analysis speech frame as well. It is interesting to note that the necessary amount of lags needed is bounded by the model order since for an order p we need at least $4 \times p$ lags.

3. Experimentation with speech signals

The key purpose of every spectral estimation method is to overcome, as possible, the problems of frequency resolution and spectral leakage. On the other hand, it would be beneficial if it could be applied as a representative and robust feature extraction technique. In the case of speech signals (non stationary signals), we are mainly interested in the estimation of quantities such as formants and their accompanying bandwidths. A robust estimation of such parameters would be proved to be beneficial for application areas such as speech synthesis and recognition. Although it has been proved that spectral estimation methods based on eigenanalysis techniques, present very good results on resolution and leakage, there is little or no effort on investigating if they could be used as feature extraction methods.

The work on this paper, besides applying the DESED method on speech signals, is furthermore concentrated on techniques that can be used in order to estimate speech formants. We note that DESED is able to estimate *p* complex damping sinusoids, where *p* is the DESED order. In order to track speech formants, we must apply a selection criterion so as to collect those sinusoids of interest. Until now, several selection criteria have been used such as, peak picking, frequency bands chasing, clustering and pattern matching techniques [6]. Furthermore, hybrid methods are also used. Indicative results of these techniques are presented later in this section. Although, the application of the above techniques offer promising results, an automatic, yet highly efficient pure signal processing approach, which could lead to a speech signal predictor, and a formant/bandwidth efficient estimation scheme, would be beneficial.

The evaluation of the proposed algorithm, as far as Formant/Bandwidth tracking of a speech signal is concerned, was based on experimental results from synthetic and real voice signals. The use of synthetic voice signals facilitated the comparison, since their Formant/Bandwidth values are known a-priori. The synthetic signals were generated using the Klatt Cascade-Parallel Formant Speech Synthesizer [7].

In all cases, the speech signal is passed through a pre-emphasis filter and divided into overlapping frames. The pre-emphasis factor was set to 0.6. The overlapping factor is always 50% while the size of the analysis frame varies, ensuring however that at least one pitch period is included in the frame. Every speech frame is being processed without use of windowing. In the case of synthetic 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.

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 is depicted in Figure 1.

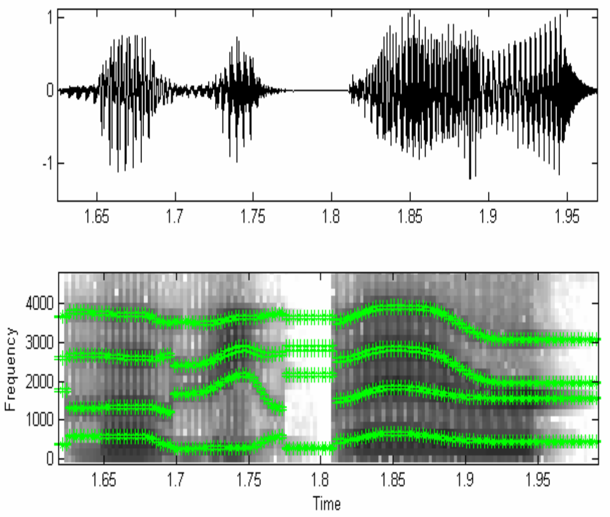

Figure 1: Reference speech frame and its spectrogram (synthetic speech signal).

As we may see from the figure, formant trajectories are easily recognized. Figure 2(a) shows the spectral analysis results for the DESED method, applied on the signal. The frame size is set to 64 and the order is set to 8 in order to estimate four formants. Additionally, Figure 2(b) illustrates the estimated formants when a frame size of 128 is used. The background image is the spectrogram where the superimposed points on it specify the estimated formants. Obviously, the algorithm manages to track formant trajectories but in several cases (frames) fails to correctly estimate some formants while dispersion over frequencies is apparent.

Another, more robust, approach for formant estimation is based on the technique of frequency bands chasing. By making use of an initial and a-priori knowledge of formants locations (e.g., LPC guided), we track formant trajectories on every analysis frame. Some indicative results of this technique are depicted in Figure 3. The main drawback of this technique is the request for pre-processing in order to set the appropriate frequency bands frontiers.

Further experimentation revealed that best formant estimation is achieved when the algorithm process a small amount of speech frame autocorrelation lags, typically 24 to 32. This is illustrated in Figure 4, and should be contrasted to the results depicted in Figures 1 and 2. In Figure 4 we observe that even for small speech frame duration (N=128 or equivalently 5.8 msec), the method successfully estimates formant trajectories. When larger speech segments are used, dispersion of formants is decreased and formant trajectories become smoother.

As far as estimation of Formant/Bandwidths is concerned, Table 1 presents the comparison results (mean deviation from the reference synthesis-parameters values). The table also includes mean deviation results derived when LPC estimation is used. The speech segment used for calculation of the here presented results is the one used in Figures 2 and 4. Deviation values confirm that the proposed method closely estimates actual formants values with a mean deviation close or even better than LPC, especially for the fourth formant. The same behaviour is observed also for bandwidth estimation whereas, except for the first formant bandwidth, the proposed method achieves smaller mean deviation values.

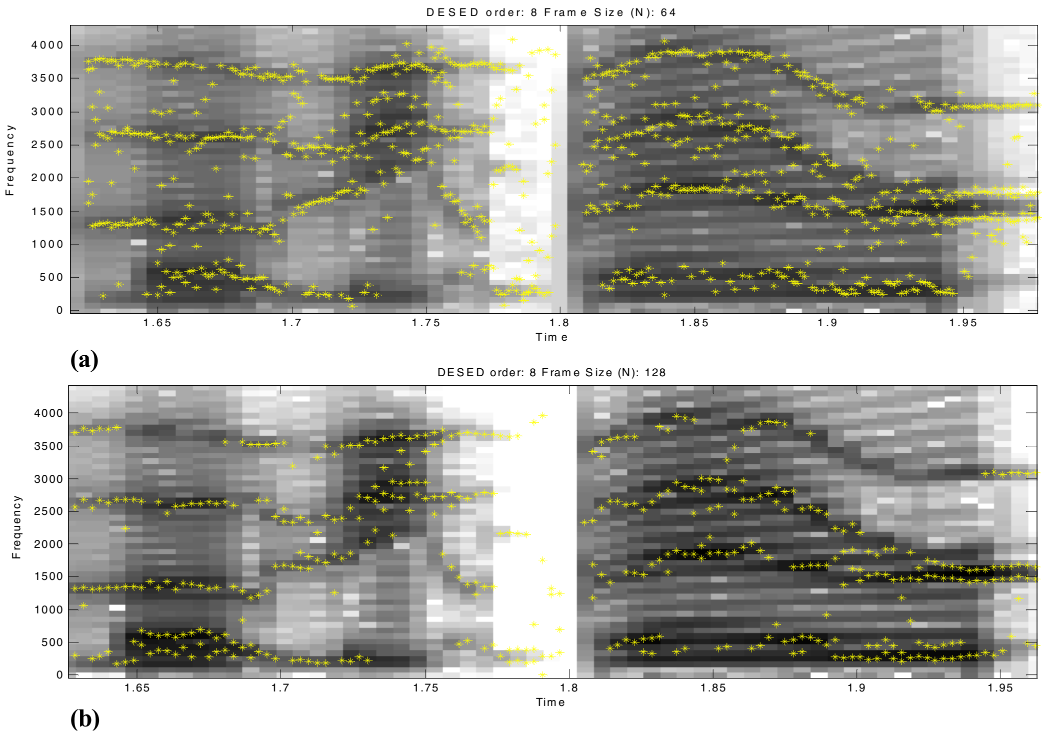

Figure 2: Formant trajectory estimation when using the DESED method on the synthetic speech signal (reference signal). Analysis results are depicted in (a) for window size N=64 and in (b) for window size N=128.

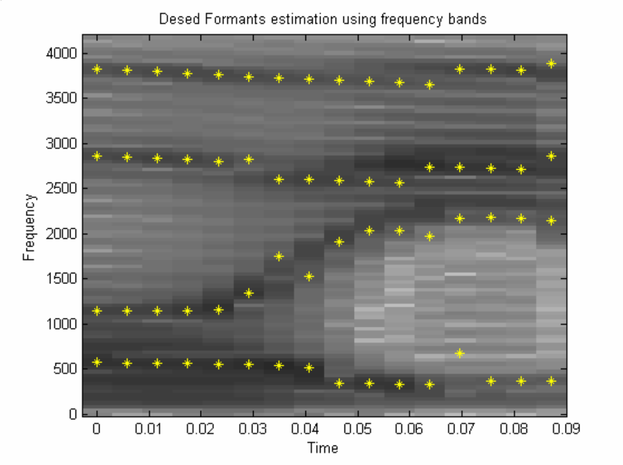

Figure 3: Formant estimation on the synthetic speech signal using the DESED method and applying frequency bands chasing.

In Figure 5 the method’s application in a natural speech signal is illustrated, where the estimated formants are superimposed on the spectrogram. The signal comprises the Greek utterance “ðilaði’ me to o’noma’ mu” (thus, with my name). Finally, a segment of their respective bandwidths, denoted as error bars around the formant value, are plotted in Figure 6. It is observed that the method estimates formant trajectories efficiently; note that high density regions of the spectrogram indicating high energy frequency bands or peaks of the voiced signal spectrum are nicely modeled.

Table 1: Formant/Bandwidth mean deviation

Formants	Mean Deviation DESED (Hz)	Mean Deviation LPC (Hz)
F1	62	88
F2	77	74
F3	88	93
F4	101	152
Bandwidths	Mean Deviation DESED (Hz)	Mean Deviation LPC (Hz)
B1	68	19
B2	47	59
B3	46	89
B4	34	343

4. Conclusions

In this paper, we have presented the use of a decimative spectral estimation method that tries to decompose a signal into complex damped sinusoids, in order to estimate speech formant/bandwidth parameters. The algorithm performs artificial decimation for increased frequency resolution, while it exploits the full data set. The results have shown that the proposed algorithm successfully estimates formant trajectories and their respective bandwidths. Moreover, the proposed method achieves good results even for small window size. Experimentation results on synthetic voice signals with known formant/bandwidth synthesis parameters, confirmed the previous assumption. Finally, we have introduced some ideas for finer formant/bandwidth estimation based on a hybrid of pattern matching and signal processing techniques.

5. Acknowledgements

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

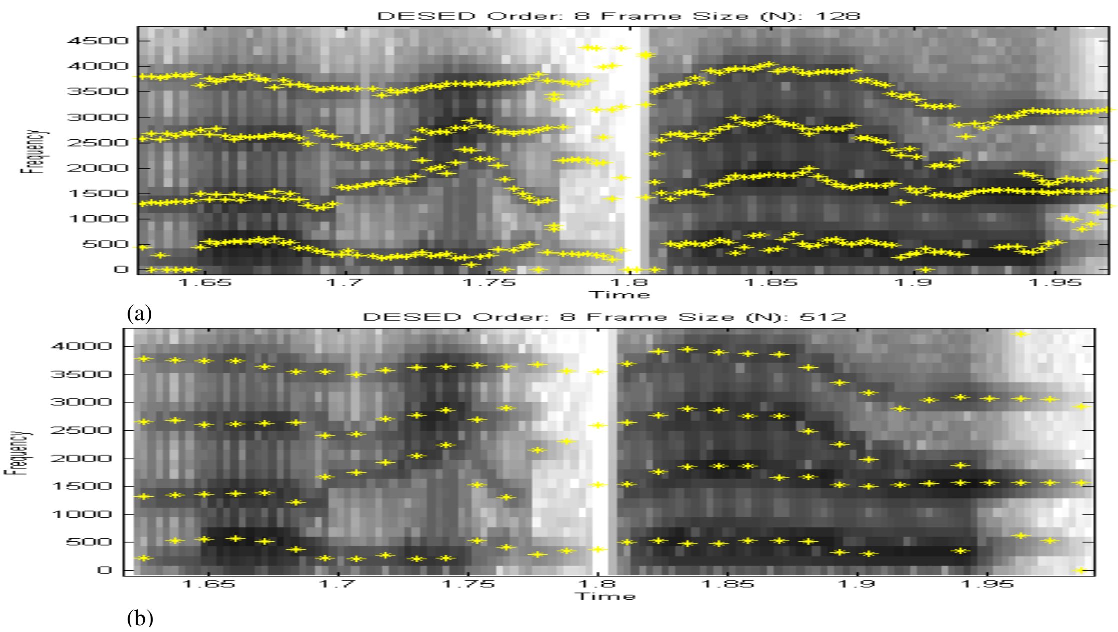

Figure 4: Formant trajectory estimation using the DESED method on the 32 first autocorrelation lags of the synthetic speech signal (reference signal). Analysis results are depicted in (a) for window size N=128 and in (b) for window size N=512.

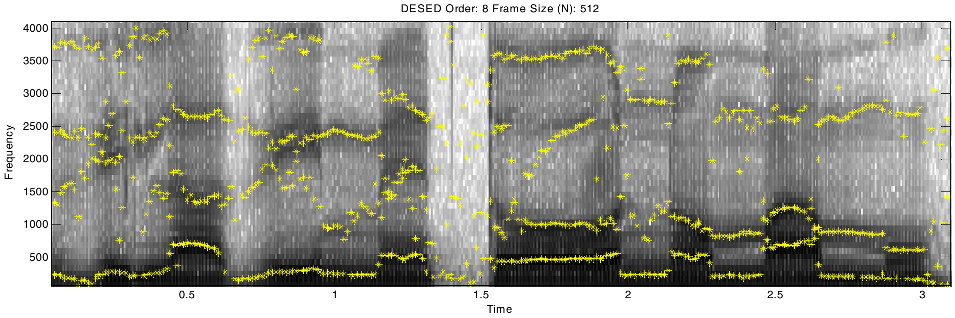

Figure 5: Formant trajectory estimation using the DESED method and respective spectrogram for a natural speech signal.

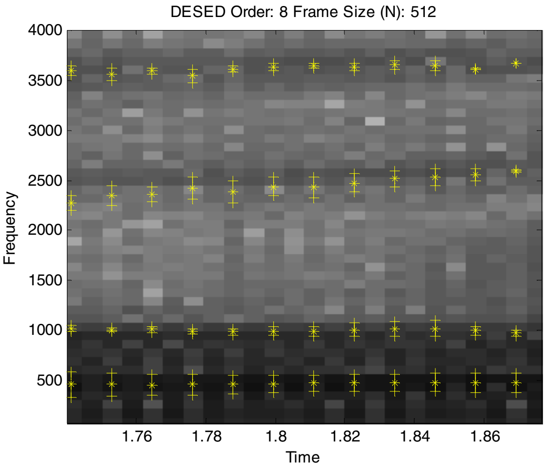

Figure 6: Bandwidth estimation for a natural speech signal.

6. References

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

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

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

[4] Fotinea, S-E., Dologlou, I., and Carayannis, G., "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, 2002.

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

[6] Tsiakoulis, P., Karabetsos, S., Fotinea, S.-E., Dologlou, I., "Spectral Estimation for Speech Signals based on decimation and eigenanalysis", *in Proc. HERCMA-2005 (The 7th Hellenic European Conference on Computer Mathematics & its Applications)*, 2005, to appear.

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