# FORMANT ESTIMATION OF SPEECH SIGNALS USING SUBSPACE-BASED SPECTRAL ANALYSIS

*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
phone: + 30 2106875405, fax: + 302106854270, email: sotoskar@ilsp.gr
web: www.ilsp.gr

## ABSTRACT

*The objective of this paper is to propose a signal processing scheme that employs subspace-based spectral analysis for the purpose of formant estimation of speech signals. Specifically, the scheme is based on decimative spectral estimation that uses Eigenanalysis and SVD (Singular Value Decomposition). The underlying model assumes a decomposition of the processed signal into complex damped sinusoids. In the case of formant tracking, the algorithm is applied on a small amount of the autocorrelation coefficients of a speech frame. The proposed scheme is evaluated on both artificial and real speech utterances from the TIMIT database. For the first case, comparative results to standard methods are provided which indicate that the proposed methodology successfully estimates formant trajectories.*

## 1. INTRODUCTION

Formant estimation of speech signals is of special concern since they consist an important feature set that could be used for a wide range of applications, spanning from phoneme classification to speech synthesis and recognition [1].

Many methods for Formant estimation rely on an all-pole assumption of the vocal tract transfer function derived from a Linear Predictive (LP) analysis of speech. More robust estimation is achieved when continuity constraint on formant candidates is applied, either rule-based or through dynamic programming [1][2][3]. Moreover, alternative signal processing approaches have also been proposed. For example, in [4] the AM-FM modulation model and the multiband demodulation analysis scheme are applied to formant tracking. Another approach is presented in [5], in which formants are estimated by pre-filtering the speech signal prior to spectral peak estimation, with a time varying adaptive filter. An improvement on the latter scheme has been proposed in [6]. Other recent approaches include the processing of the differential phase spectrum [7] and the utilization of particle filters [8].

In this paper, an alternative approach is proposed which is based on spectral estimation of speech signals using subspace-based techniques. Early investigations and results were presented in [9] which had provided first indications on the potentials of the method. The current paper reports on the full algorithmic scheme resulting in a more robust methodology that is evaluated using both synthetic and natural speech signals while at the same time is compared against current formant tracking methods. Specifically, the method exploited here is called DESED (DEcimative Spectral Estimation by factor D), which has been successfully used in NMR spectroscopy ([10], [11]) and for the case of formant estimation is modified to process a small amount of the autocorrelation coefficients of a speech frame. For the rest of this paper we call this algorithm DESED-ACOR (DESED on AutoCORellation of speech). Notice that the autocorrelation coefficients seem to convey information for the formants of a speech signal [12]. In addition, the method performs Eigenanalysis and SVD (Singular Value Decomposition) of Hankel matrices structured from signal samples. The underlying model assumes a decomposition of the processed signal into complex damped sinusoids. The number and the accuracy of the estimated sinusoids depend on the requested model order. For the case of formant estimation, since the number of requested frequencies is known within a specified frequency range (e.g. usually four formants in the range 0-4KHz), the required model emerges to be exact.

The rest of the paper is organized as follows. In section 2, we present the DESED-ACOR algorithm and we detail on the proposed scheme for robust formant estimation. In section 3, we present the experimental evaluation of the algorithm through comparative results on synthetic speech signals and illustrative examples of real (male and female) speech signals from the TIMIT database. Finally, in section 4, concluding remarks are discussed.

## 2. THE DESED-ACOR ALGORITHM

The schematic diagram of the proposed methodology is shown in Figure 1. An optional resampling process is available in order to adjust the original sampling frequency to appropriate level when decimation is to be used. Then, the speech signal is passed through a low pass FIR filter that has a twofold purpose. First, it filters the signal so as to retain only the frequency band of interest and second acts as an anti-alias filter when decimation is used. Please note that an FIR filter does not introduce any new poles since its transfer function contains only zeros. Furthermore, the speech signal is passed through a pre-emphasis filter and then divided into overlapping frames. The pre-emphasis factor is tunable. The analysis can be either frame synchronous or pitch-

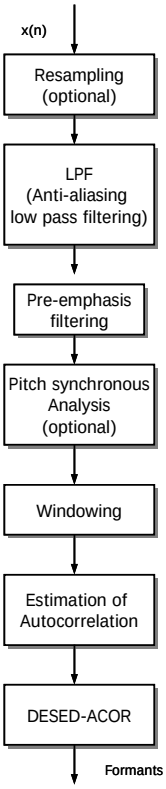


Figure 1: The proposed Formant estimation scheme.

synchronous. Pitch estimation is done using the algorithm presented in [13]. The latter is used when the sampling frequency is high enough to ensure enough samples over a pitch period. Prior to autocorrelation estimation a Hamming window multiplies every analysis frame.

The DESED-ACOR algorithm processes a small amount of the autocorrelation coefficients and yields the estimates of formant frequencies. The DESED-ACOR is outlined in Algorithm 1. Experimental observations have shown that the number of autocorrelation lags depends on the sampling frequency and the requested number of poles. Moreover, for a model order $p$ we need at least $4 \times p$ lags. A typical value for the number of autocorrelation lags range within 24 to 64. The estimation of the autocorrelation function is calculated as,

$$r_{xx}(n) = \sum_{k=0}^{N_s-|n|-1} x(k) \cdot x(k+n) \qquad (1)$$

where, $x(k)$ is the k$^{th}$ sample of a speech frame, $N_s$ is the length of a speech frame and $n$ is bounded by the requested number of lags. We further notice that for the case of formant frequency estimation the model order emerges to be exact. The assumed signal model is of the form:

$$r_{xx}(n) = \sum_{i=1}^{p} g_i z_i^n, \quad n = 0,..., N-1 \qquad (2)$$

where, $p$ is the number of complex damped sinusoids that comprise the measured signal, $g_i$ the complex amplitude, $z_i$ the signal poles and $N = \lambda \cdot p$ where the integer constant $\lambda$ has values $\lambda \geq 4$. As far as formant estimation is concerned, the objective is to estimate the frequencies $f_i$, for $i = 1,..., p$.

---

**Algorithm 1**: *DESED-ACOR*

- Construct the $L \times M$ Hankel matrix $S$ from the $N$ data points $r_{xx}(n)$ of (1), where $L - D \leq M$, $p < L - D$, $L+M-1=N$ and $D$ is the decimation factor.
- 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 $S$.
- Employ the SVD: $S_{\uparrow D} = U_{\uparrow D} \Sigma_{\uparrow D} V^H{}_{\uparrow D}$ and truncate to order $p$ by retaining only the largest $p$ singular values. This results to the enhanced version $S_{\uparrow De}$.
- Compute matrix $X = S_{\downarrow D} pinv(S_{\uparrow De})$. The eigenvalues $\lambda_i$ of $X$ give the decimated signal pole estimates, which in turn give the estimates for the formant frequencies.

---

In this paper we have chosen to evaluate the least squares version of DESED (DESED-LS) instead of its total least squares counterpart (DESED-TLS), since preliminary experiments indicated that the latter did not seem to perform better than the first in formant estimation. This has also been found true for the case of NMR spectroscopy [10]. However, a quantitative comparison between the two solutions would be beneficial and is planned for future research.

## 3. EXPERIMENTAL RESULTS

The formant tracking scheme was tested on several utterances, uttered by both male and female speakers. In order to quantitative evaluate the proposed algorithm, comparison tests were performed on synthetic speech signals using the Klatt parallel synthesizer [14]. The performance of DESED-ACOR is compared to that of two publicly available speech analysis tools namely, Praat (*http://www.praat.org*) and WaveSurfer (*http://www.speech.kth.se/wavesurfer/*). Experimentation also included sentences from the TIMIT database, presented later in this section. Since there is no systematic way or a straightforward criterion to appreciate formant estimates, evaluation is qualitative by visual inspection with the aid of the corresponding spectrograms.

We have evaluated all three methods in voiced regions of three synthetic speech signals (namely *synth_i*, i=1,2,3) having a total duration of 10 sec. In order to provide comparative results we have tried to exclude regions where any of the methods fail to estimate a formant value. For all methods we have manually tuned the parameters for best possible estimation (note that for the case of examples *synth_1* and *synth_2*, the parameters of the Praat software was set up to request for five formants since it was unable to locate the first formant).

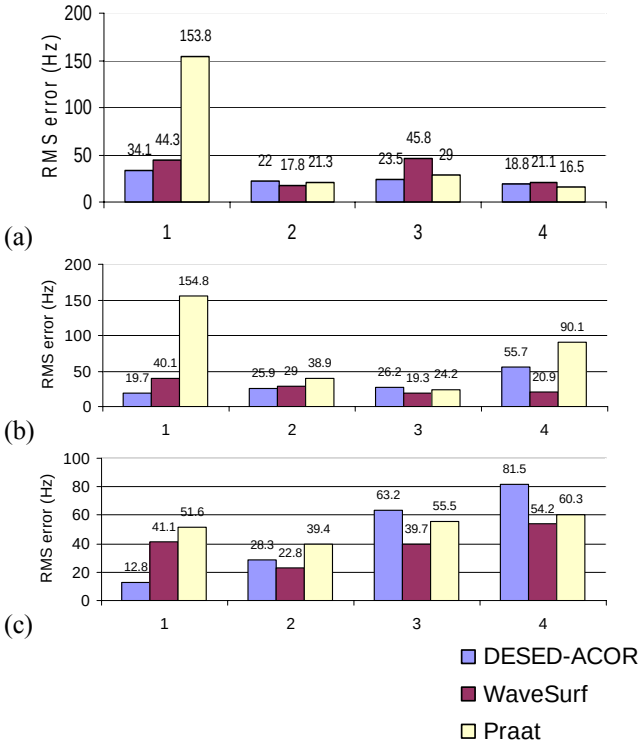


Figure 2: Formant tracking averaged RMS errors for DESED-ACOR, WaveSurfer and Praat: (a) on *Synth_1*, (b) on *Synth_2* and (c) on *Synth_3*. The horizontal axis denotes the formant number.

Figure 2 present averaged RMS (Root Mean Squared) formant frequency estimation errors for all synthetic signals experimented with. It is observed that the DESED-ACOR method achieves reliable formant estimates, especially for the first formant where in all cases it has the smaller RMS error. Furthermore, in most cases DESED-ACOR is among the best two formant estimators and in general provides comparable estimates in relevant to Praat and WaveSurfer. Additionally, it should be noted that, in the case of DESED-ACOR, no stylization or post-processing in formant trajectories and estimates is performed.

An example of formant tracking with DESED-ACOR together with actual formants used in synthesis is depicted in figure 3. It is shown that formant estimates closely follow real trajectories, failing only at unvoiced regions or transitions. Generally, experimentation showed that DESED-ACOR is able to reliably estimate the whole set of formants while for efficiency and best estimation accuracy (minimum RMS error for every individual formant) depends on the proper adjustment of the parameters of the signal processing methodology prior to DESED algorithm application. For example, it has been observed that estimation accuracy for the first formant decreases while for the fourth formant increases by adjusting the pre-emphasis factor. However, a systematic investigation on the parameter values and the interrelation between them is the main goal of planned research. Similarly, as mentioned before, the parameter values of the two other estimators that DESED-ACOR is compared with, are also manually adjusted to guarantee as more efficient estimation results as possible.

An example of real speech concerns a speech sentence from a male speaker from the TIMIT database. The signal comprises the English utterance "*She had your dark suit in greasy wash-water all year*". Figure 4 illustrates the estimated formant trajectories superimposed on the corresponding spectrogram. The sampling frequency is 16 KHz. A frame size of 256 samples with 50% overlap was used. The requested number of formants was set to four (p = 8) while the decimation factor was 2 and the number of autocorrelation lags was 32. The cut-off frequency of the FIR filter was set to 4 KHz. Obviously, the algorithm manages to track smooth formant trajectories with decreased dispersion. This is confirmed from the high density regions of the corresponding spectrogram indicating high energy frequency bands (peaks) of the signal's spectrum.

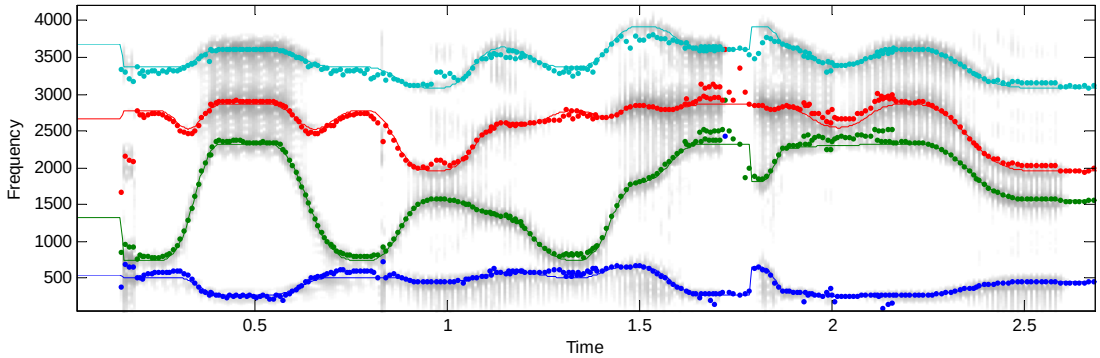


Figure 3: DESED-ACOR formant tracking on synthetic speech. Continuous lines denote actual formant tracks. Dots indicate DESED-ACOR estimates.

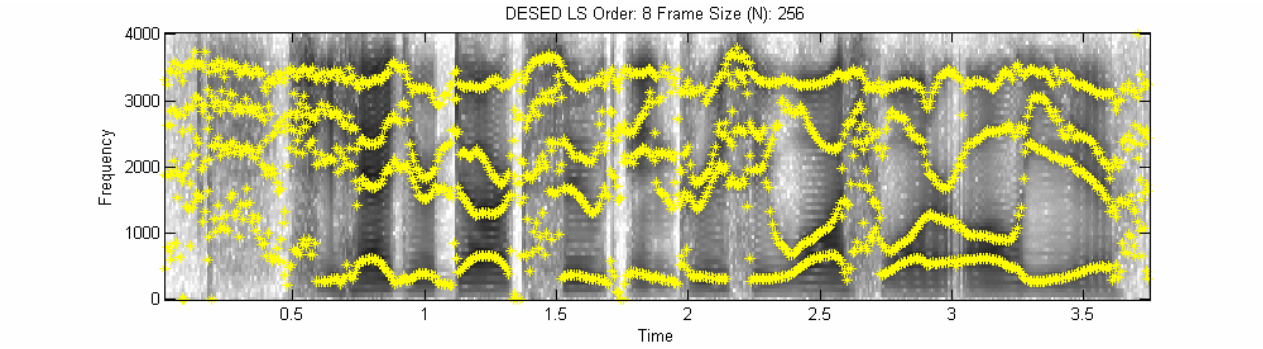


Figure 4: Formant trajectory estimation using the DESED-ACOR (DESED-LS on Autocorrelation coefficients) method on an utterance of a male speaker from the TIMIT database.

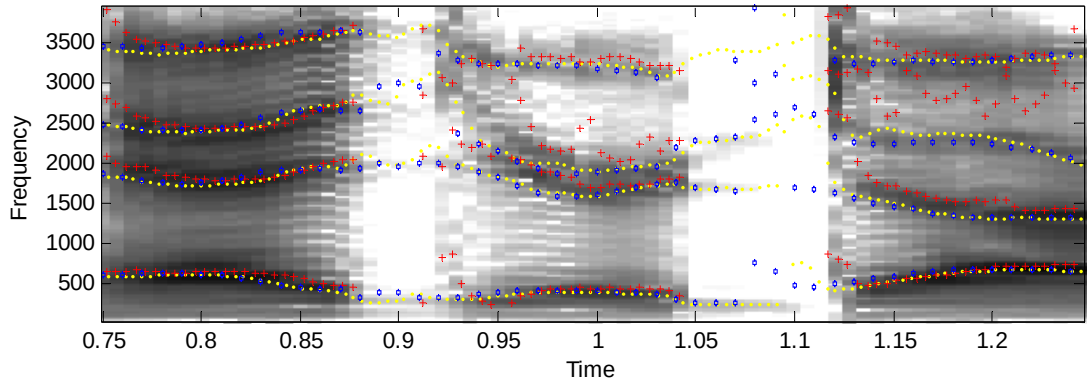


Figure 5: Formant tracking of DESED-ACOR (dots-yellow), WaveSurfer (hexagon-blue) and Praat (cross-red) on a selected region of the natural speech of figure 4.

Figure 5 depicts the estimated formant trajectories derived from DESED-ACOR, WaveSurfer and Praat, on a selected region from the example illustrated in figure 4. It is seen that DESED-ACOR achieves consistent formant tracks which are comparable to that of WaveSurfer and Praat although the latter performs poorly in some cases.

Another paradigm is that of a sentence from the TIMIT database uttered by a female speaker. The utterance is "*She had your dark suit in greasy wash-water all year*". Figure 6, depicts the estimated formant trajectories superimposed on the corresponding spectrogram. The original sampling frequency of 16 KHz is converted to 24 KHz. A frame size of 512 samples with 80% overlap was used. The requested number of formants was set to five (p = 10) while the decimation factor was 1 and the number of autocorrelation lags was 40. The cut-off frequency of the FIR filter was set to 6 KHz since formant values for females are usually higher than males. In accordance with the previous example, the algorithm manages to track the whole set of formants and produces smooth trajectories with decreased dispersion. This is again confirmed from the high density regions of the corresponding spectrogram where high energy frequency bands (peaks) of the signal's spectrum are nicely modelled.

## 4. CONCLUSIONS

In this paper, we have investigated the use of subspace-based techniques on formant frequencies estimation of speech signals and we have proposed a signal processing scheme to enhance their ability for robust tracking. The signal model assumes decomposition into complex damped sinusoids which act as formant candidates. Furthermore, we have seen that the use of autocorrelation lags convey much of the information of speech formants thus requiring an exact model order depending on the number of formants. Consequently, this leads to improved capability for successful formant trajectories estimation. Furthermore, some refinements in the signal analysis process were introduced in order to obtain a robust algorithm for formant extraction.

## 5. ACKNOWLEDGEMENTS

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

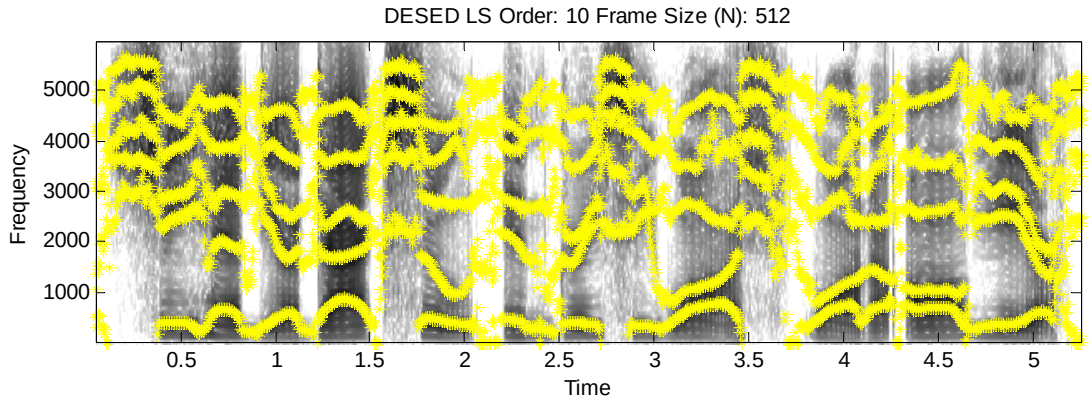


Figure 6: Formant trajectory estimation using the DESED-ACOR method on an utterance of a female speaker from the TIMIT database.

## REFERENCES


[1] M. Lee, J. van Santen, B. Mobius, and J. Olive, "Formant Tracking using Context-Dependent Phonemic Information," *IEEE Trans. Acoust., Speech and Audio Processing*, vol. 13, no. 5, pp. 741-750, Sept. 2005.

[2] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust., Soc. Am.,* Vol. 47, no. 2, pp. 634-648, 1970.

[3] D. Talkin, "Speech Formant Trajectory Estimation Using Dynamic Programming with Modulated Transition Costs," *J. Acoust., Soc. Am.,* S1, pp. S55, 1987.

[4] A. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking using Multiband Energy Demodulation," *J. Acoust., Soc. Am.,* Vol. 99, no. 6, pp. 3795-3806, June 1996.

[5] A. Rao and R. Kumaresan, "On decomposing Speech into Modulated Components," *IEEE Trans. Acoust., Speech and Audio Processing*, vol. 8, no. 3, pp. 240-254, May 2000.

[6] K. Mustafa and I. C. Bruce, "Robust Formant Tracking for Continuous Speech with Speaker Variability," *IEEE Trans. Acoust., Speech and Audio Processing*, vol. PP, no. X, pp. 1-10, accepted for publication, Jan. 2005.

[7] B. Bozkurt, T. Dutoit, B. Doval, and C. D'Alessandro, "Improved differential Phase Spectrum processing for Formant tracking," in *Proc. InterSpeech-ICSLP 2004*, Jehu Island, Korea, 2004, pp. 2421-2424.

[8] Y. Shi and E. Chang, "Spectrogram-Based Formant Tracking via Particle Filters," in *Proc. ICASSP 2003*, Hong Kong, China, Apr. 2003, pp. I-168-I-171.

[9] S. Karabetsos, P. Tsiakoulis, S-E. Fotinea, and I. Dologlou, "On the Use of a Decimative Spectral Estimation Method Based on Eigenanalysis and SVD for Formant and Bandwidth Tracking of Speech Signals", in *Proc. Interspeech-2005*, Lisbon, Portugal, 2005, pp. 709-712.

[10] S-E. Fotinea, I. Dologlou, and G. Carayannis, "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, pp. 321-330, 2002.

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

[12] G. Carayannis and P. Jospa, "On the Analysis of Autocorrelation Function for Speech Spectra Estimation – Application for Nasality detection", in *Proc. ICASSP 1977*, Vol. 2, pp. 754-757, 1977.

[13] I. Dologlou and G. Carayannis, "Pitch Detection based on zero-phase Filtering", *Speech Communication*, vol. 8, No. 4, pp. 309-318, 1989.

[14] D. H. Klatt, "Software for a cascade/parallel formant synthesizer", *J. Acoust., Soc. Am.,* vol. 67, pp. 971-995, 1980.