

Robust Energy Demodulation Based on Continuous Models with Application to Speech Recognition

Dimitrios Dimitriadis and Petros Maragos

School of ECE, National Technical University of Athens,
Zografou, 15773 Athens, Greece

[ddim, maragos]@cs.ntua.gr

Abstract

In this paper, we develop improved schemes for simultaneous speech interpolation and demodulation based on continuous-time models. This leads to robust algorithms to estimate the instantaneous amplitudes and frequencies of the speech resonances and extract novel acoustic features for ASR. The continuous-time models retain the excellent time resolution of the ESAs based on discrete energy operators and perform better in the presence of noise. We also introduce a robust algorithm based on the ESAs for amplitude compensation of the filtered signals. Furthermore, we use robust nonlinear modulation features to enhance the classic cepstrum-based features and use the augmented feature set for ASR applications. ASR experiments show promising evidence that the robust modulation features improve recognition.

1. Introduction

There is significant evidence for the existence of amplitude and frequency modulations (AM-FM) in speech resonance signals, which make their amplitudes and frequencies vary instantaneously within every pitch period. Motivated by this evidence, Maragos et al. [1] proposed the modeling of each speech resonance with an AM-FM signal,

$$x(t) = a(t) \cos\left[\int_0^t \omega_i(\tau) d\tau\right] \quad (1)$$

and of the total speech signal as a superposition of a few of such AM-FM signals, one for each formant. The estimation of their instantaneous frequencies $\omega_i(t)$ and amplitude envelopes $|a(t)|$, is referred to as the ‘Demodulation Problem’. Different approaches exist concerning the solution of this problem, such as Hilbert Transform (HT) or the Energy Separation Algorithm (ESA) [1], where the Teager-Kaiser Energy Operator (TEO) $\Psi[x] = \dot{x}^2 - x\ddot{x}$ is used.

One problem of the discrete-time ESA (DESA) is the approximation of the time-derivatives by one-sample differences. This approximation introduces significant modeling noise in the corresponding derivatives, especially for noisy signals. Recently we have used, [4], continuous-time expansions of discrete-time signals, such as sampled speech, to numerically implement the required differentiation with closed formulae. This process provides smooth estimates of the signals’ derivatives and adds robustness to the ESA.

Our motivations for the present research work include the following: (i) Finding continuous-time models for simultaneous interpolation and demodulation of energy-differential of discrete-time signals. (ii) Providing robustness to noisy speech

processing applications. (iii) Adding new information to the ASR feature set, such as instantaneous bandwidths, which can better model various nonlinear and time-varying aspects of speech dynamics with corresponding improvement in ASR.

The paper is structured as follows: In Section 2, novel continuous-time modulation models for speech resonances are introduced. Also, an amplitude compensation algorithm is described to correct artifact modulations introduced by the band-pass filtering. Section 3 discusses the extraction of novel short-time feature vectors from speech signals and the estimation of an augmented set of acoustic features for improving HMM-based phonemic recognition. The improved ASR results for clean and noisy speech data are presented.

2. Continuous-Time Demodulation Models

2.1. Smoothing Splines

The problem of smooth differentiation of signals, especially of the noisy ones, led us to interpolate the signal samples using smoothing splines, whose main advantage (compared to exact splines) is that the interpolating polynomial does not pass ‘precisely’ through the signal samples but ‘close enough’ so as to give smooth derivatives. The smoothing spline interpolating function is defined as the function s_ν that minimizes the mean square error criterion

$$E = \underbrace{\sum_{n=-\infty}^{+\infty} (x[n] - s_\nu(n))^2}_{E_d} + \lambda \underbrace{\int_{-\infty}^{+\infty} \left(\frac{\partial^r s_\nu(t)}{\partial t^r}\right)^2 dt}_{E_s}$$

where E_d is the data fitting error and E_s quantifies the roughness of the interpolant by the mean square value of its derivative.

The application of smoothing splines to interpolating discrete-time signals is thoroughly described in [3, 5]. The interpolating curve of the smoothing splines is given by:

$$s_\nu(t) = \sum_{n=-\infty}^{+\infty} c[n] \beta_\nu(t - n) \quad (2)$$

where $\beta_\nu(t)$ is the B-spline of order ν , and $c[n]$ are the spline coefficients depending only on the data $x[n]$, the parameter λ and the analytic expression of the corresponding B-spline.

The sequence $c[n]$ can be determined uniquely by using the signal sequence $x[n]$ as input of an IIR filter. This IIR filter has a symmetric impulse response, and all its poles are inside the unit circle. Thus, the spline coefficients $c[n]$ can be stably determined via a few recursive equations [3, 4]. The Spline filter

frequency response is

$$C(z) = \frac{1}{B_5(z) + \lambda(-z + 2 - z^{-1})^3} \quad (3)$$

where $B_5(z)$ is the z-transform of the 5^{th} -order B-spline and λ is a parameter that regulates the amount of smoothness of the interpolation curves. The positive design parameter λ controls the trade-off between how smooth the interpolating curves are and how small the close-to-the-data fitting distances will be. The bandwidth of the Spline filter is a function of parameter λ , where larger values of this parameter give narrower bandwidth filters (consequently smoother interpolant curves) and vice-versa.

Smoothing splines are first used to interpolate the discrete-time signals $s[n]$ and then they are differentiated using closed formulae. When these spline-based derivatives of the signal are used to compute first the TEO and then in turn the ESA demodulation estimates of amplitude and frequencies, the resulting algorithm is called '*Spline ESA*'.

2.2. Combination of Gabor Filtering and CT-ESA

ESAs cannot handle wideband signals, such as speech signals, due to inherent limitations of the algorithm. An efficient way to deal with such limitations is the bandpass filtering. For this process, the Gabor filters are chosen for several reasons, well explained in [1].

The continuous TEO Ψ , combined with bandpass filtering and sampled at time instances $t = nT$, is given by:

$$\Psi[x(t)] = \dot{x}^2(t) - x(t)\ddot{x}(t)|_{t=nT} \quad (4)$$

where $x(t) = s(t) * g(t)$, $s(t)$ is the continuous-time signal and $g(t)$ is the Gabor filter impulse response:

$$g(t) = \exp(-b^2 t^2) \cos(\omega_c t) \quad (5)$$

where b and ω_c are the filter parameters.

Since, convolution commutes with time-differentiation, we have,

$$\frac{d^m}{dt^m} (s(t) * g(t)) = s(t) * \frac{d^m}{dt^m} g(t), \quad m = 1, 2, 3, \dots \quad (6)$$

The Gabor time derivatives are given by closed formulae:

$$\frac{dg(t)}{dt} = (-2b^2 t \cos(\omega_c t) - \omega_c \sin(\omega_c t)) \exp(-b^2 t^2)$$

$$\begin{aligned} \frac{d^2 g(t)}{dt^2} &= (4b^2 \omega_c t \sin(\omega_c t) + (4b^2 t^2 - 2b^2 - \omega_c^2) \cos(\omega_c t)) \\ &\quad \times \exp(-b^2 t^2) \end{aligned}$$

Using the above equations in Eq. (4), the output of Ψ acting on the bandpass filtered signal $s(t)$, is given by:

$$\begin{aligned} \Psi[x(t)] &= \Psi[s(t) * g(t)] = \\ &= \left[\frac{d}{dt} (s(t) * g(t)) \right]^2 - (s(t) * g(t)) \left[\frac{d^2}{dt^2} (s(t) * g(t)) \right] \\ &= \left[s(t) * \frac{dg(t)}{dt} \right]^2 - (s(t) * g(t)) \left[s(t) * \frac{d^2 g(t)}{dt^2} \right] \end{aligned}$$

Through this approach, the necessary processes of bandpass filtering and the subsequent differentiations are combined into a single convolution with derivatives of the Gabor response.

Since the output of the continuous-time TEO Ψ is being sampled at time instances $t = nT$, we convolve the discrete-time speech signal $s[n]$ with the discrete-time Gabor derivative filters,

$$g^{(m)}[n] = \frac{d^m}{dt^m} g(t)|_{t=nT}$$

In the sequel, we use the continuous-time ESA formulae for demodulation. This whole algorithm, called '*Gabor ESA*', exhibits some advantages compared to the Spline ESA or to the original discrete demodulation algorithm DESA. At first, bandpass filtering of noisy signals increases the SNR of the filtered signals. Then, the model's parameters are reduced by one compared to the corresponding ones of the Spline ESA where the λ -parameter is introduced. Finally, the differentiation is introduced on the filters and not on the speech signal itself. This fact leads to smooth time-derivatives of the filtered signal. The use of Gabor filter, as well its derivatives, supports this claim, too.

The Gabor ESA is computationally more intensive than the original DESA or the Spline ESA, when applied to bandpass filtered signals. On the other hand as shown in Fig. 1, Gabor ESA provides smoother estimates of the instantaneous frequency compared to the corresponding ones of the Smooth DESA [6], as expected, especially in noisy signals.

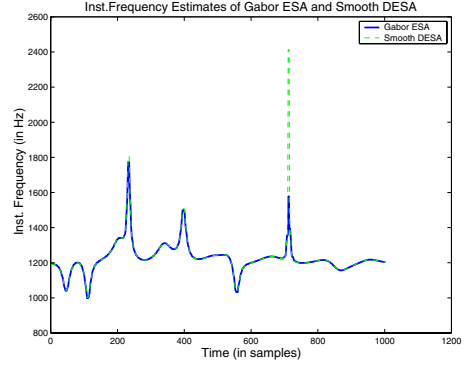


Figure 1: *Instantaneous Frequency Signals of Phoneme /aa/ Using Gabor ESA and Smooth DESA.*

In Fig. 1, a phoneme /aa/, taken from the TIMIT database, is bandpass-filtered with a Gabor filter placed at a center frequency $f_c = 1185$ Hz and with bandwidth parameter $b = 400$ Hz. The filter is manually placed, taking under consideration the spectrum of the phoneme, so that a single resonance is being filtered. Also, white gaussian noise with SNR value equal to 10 dB is added in order to show the algorithms' robustness to noise. Even though, Gabor ESA and Smooth DESA are very robust for noise and give smooth estimates of $f_i[n]$, the Gabor ESA algorithm yields somewhat smoother estimates as expected (smaller spikes).

2.3. Amplitude Compensation

The filtering of an AM-FM signal $s[n] = a[n] \cos(\phi[n])$ using a Gabor filter with frequency response $G(\omega)$ has a great impact on the amplitude of that signal. In [2], Bovik et al. proposed that the filtered signal $x[n] = s[n] * g[n]$ can be approximated by:

$$x[n] \approx a[n] |G(\omega_i[n])| \cos(\phi[n] + \theta[n]) \quad (7)$$

where $\theta[n] = \angle G(\omega_i[n])$ and $\omega_i[n]$ is the instantaneous angular frequency given by $\omega_i[n] = (1/T)d\phi[n]/dn$ (where T is the sampling period).

The instantaneous angular frequency of the filtered signal is not altered when the real-valued Gabor filter is used because $\theta[n] = 0$. Thus, we propose the following filter compensation algorithm. At first, using any one of the demodulation algorithms (DESA, Spline ESA or Gabor ESA) the instantaneous frequency is estimated. Then, if $\tilde{a}[n]$ is the ESA amplitude estimate of the filtered signal, the compensated estimate of the true amplitude $a[n]$ is given by

$$a[n] \approx \frac{\tilde{a}[n]}{|G[\omega_i[n]]|} \quad (8)$$

So, by using the ESA frequency estimate $\omega_i[n]$, we are able to estimate the original instantaneous amplitude $a[n]$ using Eq. (8) and the ESA amplitude estimate $\tilde{a}[n]$.

2.4. Comparison of Demodulation Algorithms

The testing AM-FM signals used for the experiments, are the same as in [1]:

$$x[n] = (1 + 0.05k \cos(\frac{\pi n}{100})) \cos(\frac{\pi n}{5} + m \sin(\frac{\pi n}{100})) + e[n],$$

where $e[n]$ is the added white Gaussian noise of different SNR levels and $m = 1, \dots, 10$, $k = 1, \dots, 10$. The test signals are filtered by a Gabor filter with a center frequency $\omega_c = \pi/5$ and bandwidth parameter $b = 0.1875$. The mean absolute error rates of the instantaneous estimates are calculated over 100 different AM, FM modulation depths and for different SNR values. The demodulation algorithms being tested are the Smooth DESA, Gabor ESA, Spline ESA and Prony ESA. It must be noted that even though the Smooth DESA appeared to have the best performance for the test signals, in total, the Gabor ESA gives the smoother estimates especially when speech signals are used as inputs. As shown in Section 2.2, this can be explained by the fact that time-derivation is introduced on the filters and not on the input speech signals.

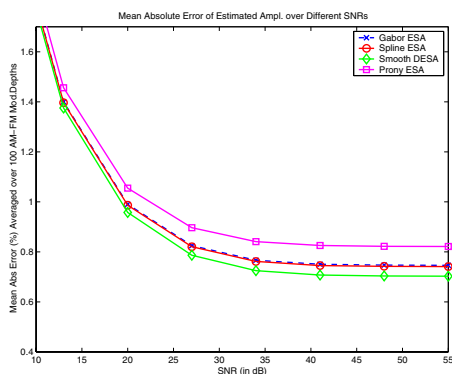


Figure 2: Instantaneous Amplitude-Compensated Signals for Different ESAs.

In Fig. 2 it is shown that the amplitude compensation algorithm is very robust for noisy signals (the error rates are not larger than 2.2% of the original amplitude signal even for small SNR values i.e. (5-15 dB)). Also, the algorithm's performance seems to be very good as the estimation errors of the amplitude signals are significantly small. The estimates of the amplitude-compensated filtered signals are very close to the original input signals.

3. Applications to Speech Recognition

The algorithms described above are applied to speech recognition applications. The proposed feature set consists of the FM-Modulation Depths of the AM-FM signals; for details see [5]. The feature vectors are computed over a 30 ms window and are updated every 10 ms. The novel feature set and its first and second time derivatives are concatenated with the, almost standardized, feature set of MFCCs (Mel Frequency Cepstrum Coefficients), which consists of the first twelve cepstrum coefficients, the mean-square amplitude (i.e. energy) of the signal and their corresponding time-derivatives.

We augment the 'standard' feature vector and thus create a *hybrid feature vector* by incorporating information from the nonlinear structure of speech of the modulation type as additional features. We use feature vectors that contain information both from the smoothed cepstrum of the linear model and from the speech modulations. Analytically, the FM-Modulation depths model the formant track bandwidths between successive pitch pulses. The linear part provides information for the placing of the formant tracks and the corresponding non-linear one for their modulations and oscillations.

We have automated the extraction of modulation features from speech signals in the following way:

First, we use a parallel filterbank of 6 overlapping Gabor band-pass filters, whose center frequencies are spaced in the mel-frequency scale, spanning the whole frequency range $[0..F_s/2]$ Hz, where F_s the sampling frequency. Second, the output signals from each Gabor bandpass filter are demodulated via the Spline ESA or the Gabor ESA into their instantaneous amplitude $a_i(t)$ and frequency $f_i(t)$ component signals. Third, the filtered signals are framed. For each such short-time analysis frame and for each band, the amplitude-weighted mean F_w and standard deviation B_w of the instantaneous frequency signal are estimated as in [7]. Pitch pulses introduce spikes in the instantaneous frequency signals. In order to avoid them, a clipping process is introduced, where a maximum and minimum deviation of the weighted mean frequency estimate is set. Finally, we compute the *frequency modulation depth (FMD)* in each band as the ratio $K = B_w/F_w$, for each analysis frame. The modulation feature vectors consist of the sequence of the FM percentages K_i , $i = 1, \dots, 6$ and their first and second time derivatives, a total of 18 numbers per vector.

We have used the hybrid feature vector with size of 57 feature vector elements (39 samples for the MFCCs and 18 for the FMD) as input to a hidden Markov model (HMM)-based speech recognizer. The HMM back-end recognizer used is the HTK system (version 3.2). For the experiments presented below, context-independent, 3-state, left-right HMMs were used. The input vectors are split into two different data streams, one for the standard features (MFCC) and the other for the modulation features. The two streams are assumed independent. Each one of these streams has independent probability distributions which are modeled by 16 Gaussian mixture probability densities. Finally, the grammar being used is the all-pair unweighted grammar where every pair of phonemes has the same probability to appear. The stream-weights, even though they affect directly the recognition process, they are kept fixed in order to study the recognition results for the novel feature sets. So, stream weights s_1 , s_2 for the two different data-streams, MFCCs and FMD correspondingly, are set equal to $s_1 = 1$ and $s_2 = 0.25$.

¹The percentage number of phonemes correctly recognized is given by the ratio of the number of correct labels minus the insertions to the

Percentage of Phoneme Accuracy ¹	
1 stream (FMD+1st+2nd Time Derivatives)	
Spline ESA with $\lambda = 0.25$	21.59%
Gabor ESA	21.18%

Table 1: Phone Recognition Results of FMD Alone

Percentage of Phoneme Accuracy ¹	
Baseline - MFCCs	53.41%
2 streams (MFCCs+FMD)	
Spline ESA ($\lambda = 0.25$)	53.71%
Gabor ESA	53.70%

Table 2: Phone Recognition Results of FMD Concatenated with the MFCC

As stated in Table 1 the recognition results concerning the non-linear feature set alone are significantly low as the information introduced by them is of minor importance and should be added to the information about the placing of the formants. Formant tracks (linear feature set) is a much more important information set, than the formant variations and oscillations. Also, the results shown in Table 2 indicate that the nonlinear feature set offers additional information, complementary to the linear one, as the recognition results of the two-stream features are better than the baseline results.

It should be noted also that the recognition results using different demodulation algorithms, Gabor or Spline ESA, show small variations, which are not significant. The differences appearing in the estimated instantaneous signals have no significant impact in the recognition tasks.

Another recognition experiment has been held with noisy data. The database used was the Aurora3 Spanish Database. The experiments are well described in [8]. The recognition results for this task are presented in Table 3. The front-end algorithm (using Spline ESA or Gabor ESA) for the estimation of the feature set of FMD is the same as the one described above. The main differences with the TIMIT recognition task are the different backend program being used (for the Aurora task the BLasr program is used, [8]) and the different linear features (Auditory feature set, [8]).

Percentage of Word Accuracy (%) ¹			
Recognition Tasks	WM	MM	HM %
Baseline - Auditory Features	95.4	89.2	84.7
Augmented Feature Set (Auditory+FMD)			
FMD	95.2	88.7	87.2

Table 3: Word Recognition Results of FMD Concatenated with the Auditory Features

In Table 3, it is shown that the FMD feature vectors exhibit robustness in different noise scenarios and this is the main reason for the great improvement of the recognition rates for the HM case, where about a 16.3% relative improvement is obtained. Note that the baseline rates (using solely the auditory feature set for the recognition task) and the recognition rates using the augmented feature set (Auditory+FMD), are obtained

total number of phonemes in the transcription files.

using the same training and testing conditions. This is done in order to compare the contribution of the novel feature set to the recognition tasks.

4. Conclusions-Discussion

In this paper, continuous-time models for speech signals have been proposed. These models exhibit a very good performance in the estimation process of AM-FM testing signals. Especially, the Gabor ESA gives very smooth estimates when applied to speech signals. We have also introduced an algorithm for filter compensation which uses the ESA instantaneous estimates of the frequency and amplitude signals in order to estimate the original amplitude of the input signal. The filter compensation allows us to further investigate the true modulations appearing in the speech signals, which are due to the formants' movements and not to the artifacts created by the filtering process.

Further we have used the above improvements in speech demodulation for feature extraction in ASR applications. The ASR results seem to be promising since the multiband FMD features have been found to improve the recognition rates. The proposed continuous-time models and related algorithms are efficient, especially in noisy signals, because of the bandpass filtering and/or the use of smoothing splines. The novel feature set seems to be robust for noisy signals as shown for the AURORA database recognition task (High-Mismatch case), Table 3, where the noise mismatch of the training and testing sets is efficiently overcome.

5. References

- [1] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy Separation in Signal Modulations with Application to Speech Analysis", *IEEE Trans. Signal Processing*, vol. 41, pp. 3024–3051, Oct. 1993.
- [2] A. C. Bovik, P. Maragos, and T.F. Quatieri, "AM-FM Energy Detection and Separation in Noise Using Multiband Energy Operators", *IEEE Trans. Signal Processing*, vol. 41, Dec. 1993.
- [3] M. Unser, A. Aldroubi and M. Eden, "B-Spline signal processing: Part I—Theory. Part II—Efficient design and applications", *IEEE Trans. Signal Processing*, vol. 41, pp. 821–848, Feb. 1993.
- [4] D. Dimitriadis and P. Maragos, "An Improved Energy Demodulation Algorithm Using Splines", *Proc. ICASSP-01*, Salt Lake, Utah, May 2001.
- [5] D. Dimitriadis, P. Maragos and A. Potamianos, "Modulation Features for Speech Recognition", in *Proc. ICASSP-02*, Orlando, Florida, May 2002.
- [6] A. Potamianos and P. Maragos, "A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation", *Signal Processing*, vol.37, pp.95-120, May 1994.
- [7] A. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking Using Multiband Energy Demodulation", *J. Acoust. Soc. Amer.*, 99 (6), pp.3795–3806, June 1996.
- [8] J. Chen, D. Dimitriadis, H. Jiang, Q. Li, T.A. Myrvoll, O. Siohan, F.K. Soong, "Bell Labs Approach to Aurora Evaluation on Connected Digit Recognition", *Proc. of ICSLP-02*, Denver, CO, Sept. 2002.