Enhancement of Non-Stationary Speech using Harmonic Chirp Filters

Nørholm, Sidsel Marie; Jensen, Jesper Rindom; Christensen, Mads Græsbøll

Published in:
INTERSPEECH 2015

Publication date:
2015

Citation for published version (APA):
Enhancement of Non-Stationary Speech using Harmonic Chirp Filters

Sidsel Marie Nørholm, Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University
{smn, jrrj, mgc}@create.aau.dk

Abstract

In this paper, the issue of single channel speech enhancement of non-stationary voiced speech is addressed. The non-stationarity of speech is well known, but state of the art speech enhancement methods assume stationarity within frames of 20–30 ms. We derive optimal distortionless filters that take the non-stationarity nature of voiced speech into account via linear constraints. This is facilitated by imposing a harmonic chirp model on the speech signal. An implicit part of the filter design, the noise statistics are also estimated based on the observed signal and parameters of the harmonic chirp model. Simulations on real speech show that the chirp based filters perform better than their harmonic counterparts. Further, it is seen that the gain of using the chirp model increases when the estimated chirp parameter is large corresponding to periods in the signal where the instantaneous fundamental frequency changes fast.

Index Terms: speech enhancement, single-channel, non-stationary signals, harmonic chirp model

1. Introduction

Speech enhancement is important in many systems such as mobile phones, hearing aids and teleconferencing systems where the desired signal is corrupted by noise. Speech enhancement can be approached in different ways, common ones being spectral subtraction [1, 2] performed in the frequency domain or Wiener filtering performed in the frequency or time domain [2, 3]. These, and most other speech enhancement methods, assume that the signal is stationary within an analysis window, for speech this window is often assumed to be 20–30 ms.

Often, a noise driven approach is taken to speech enhancement where the power spectral density is estimated after transformation to the frequency domain. This can be done in speech free periods using a voice activity detector (VAD) [4] and extrapolating to periods with speech. In [5], this is expanded to also include new calculations in short speech pauses and brief breaks in between words, and in [6] the VAD is substituted with a speech probability, but, still, the noise estimation relies primarily on speech pauses. Therefore, the noise has to be stationary for longer periods than 20–30 ms in order for these methods to work properly. Alternatively, a signal driven approach can be taken where a model for the desired signal is assumed. An often used model is the harmonic model. Here, the signals, speech and noise, are assumed stationary within the window of 20–30 ms. However, this assumption is not fulfilled [7] since the speech signal is non-stationary and varies continuously over time.

Speech enhancement of non-stationary speech is not well covered in the literature, but the issue of non-stationary speech
monic model. The paper is concluded in Section 5.

2. Harmonic Chirp Model

Often it is assumed that the desired signal is stationary within blocks of 20-30 ms. In such a framework a normally used model for voiced speech is the harmonic signal model. However, the assumption of stationarity does not hold since the frequencies of the harmonics are changing continuously over time. Therefore, we here suggest to use a model which does not assume stationarity but instead assumes that the harmonic frequencies change linearly within one of these short segments. This can be done by using a linear chirp model and the instantaneous frequency of the \( l \)'th harmonic, \( \omega_l \), can then be expressed as:

\[
\omega_l(n) = l(\omega_0 + kn),
\]

for time indices \( n = 0, \ldots, N - 1 \) where \( \omega_0 \) is the normalised fundamental frequency and \( k \) is the fundamental chirp rate. The instantaneous phase, \( \theta_l \), of the harmonic components of the speech signal is given by the integral of the instantaneous frequency:

\[
\theta_l(n) = l \left( \omega_0 n + \frac{1}{2} k n^2 \right) + \phi_l
\]

where \( \phi_l \) is the initial phase of the \( l \)'th harmonic. Thereby, the harmonic chirp model can be expressed by:

\[
s(n) = \sum_{l=1}^{L} a_l e^{j(l(\omega_0 + kn)/2 + \phi_l)},
\]

where \( L \) is the number of harmonics, and the initial phase is included in the amplitude term to give the complex amplitude of the \( l \)'th harmonic, \( a_l = A_l e^{j\phi_l} \), with \( A_l > 0 \) being the real amplitude. We choose to work in the complex domain since this leads to simpler expressions. A real signal can be transformed to a complex signal by use of the Hilbert transform, and back again by only considering the real part of the complex signal.

We are looking at the case where the desired signal, \( s(n) \), is corrupted by noise, \( v(n) \), to give the observed signal, \( x(n) \),

\[
x(n) = s(n) + v(n).
\]

The signal and noise are assumed uncorrelated and, therefore, we have that the variance of the observed signal is the sum of the variances of desired signal and noise, \( \sigma^2 = \sigma^2_s + \sigma^2_v \).

The enhancement problem considered in this paper is then to get a good estimate of the desired signal, \( \hat{s}(n) \), based on filtering of the observed signal

\[
\hat{s}(n) = h^H x(n) = h^H s(n) + h^H v(n),
\]

where \( h = [h(0) \ h(1) \ \cdots \ h(M-1)]^H \) is the filter with length \( M \), \( x(n) = [x(n) \ x(n+1) \ \cdots \ x(n+M-1)]^T \), \( v(n) \) and \( s(n) \) are defined in a similar way to \( x(n) \) and \( \{\}^H \) denotes the (Hermitian) transpose. Again, under the assumption of uncorrelated signals, we have that \( \sigma^2 = \sigma^2_{s,ss} + \sigma^2_{s,ss} \), where \( \sigma^2_{s,ss} = h^H R_s h \) is the variance of the observed signal after noise reduction, and similar for \( \sigma^2_{s,ss} \) and \( \sigma^2_{v,ss} \).

3. Filters

One filter that can be used for extracting harmonic signals is the LCMV filter [14] which is minimising the output power of the filter while passing the desired signal according to the signal model undistorted. This filter can be modified to fit harmonic chirp signals instead and is then the solution to the optimisation problem:

\[
\min_h h^H R_s h, \quad \text{s.t.} \quad h^H Z = 1^T,
\]

where \( 1 = [1 \ \cdots \ 1]^T \), \( R_s \) is the covariance matrix of the observed signal defined as:

\[
R_s = E\{x(n)x^H(n)\},
\]

with \( E\{\} \) denoting statistical expectation, and \( Z \) is constructed from a set of modified Fourier vectors:

\[
Z = [z(\omega_0, k) \ z(2\omega_0, 2k) \ \cdots \ z(L\omega_0, Lk)],
\]

with

\[
z(\omega_0, k) = \begin{bmatrix}
1 \\
\vdots \\
e^{j\omega_0 k/2}
\end{bmatrix}.
\]

The solution to the minimisation problem is:

\[
h = R_s^{-1} Z(Z^H R_s^{-1} Z)^{-1} 1.
\]

The harmonic LCMV filter is a special case of this filter for \( k = 0 \), and in this case the problem reduces to the one in [14].

In practice the covariance matrix is not known but has to be estimated. This is often done by use of the sample covariance estimate

\[
\hat{R}_s = \frac{1}{N-M+1} \sum_{n=0}^{N-M} x(n)x^H(n).
\]

However, in this estimate it is assumed that the signal is stationary over the set of \( N \) samples. This is not the case when non-stationary speech is considered. Therefore, we also suggest a modification of the APES based filter [17]. As a part of the design of this filter, an estimate of the noise covariance matrix is generated. This is done by subtracting the part coming from the desired signal from the covariance matrix of the observed signal. By modifying this filter it will be possible to obtain a noise covariance matrix which is independent of the part of the desired signal aligning with the chirp signal model.

The APES based filter is the solution to the mean squared error (MSE) between the filtered signal and the signal model:

\[
\text{MSE} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} |h^H x(n) - a^H w(n)|^2,
\]

where \( a = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_L]^H \) and

\[
w(n) = \begin{bmatrix}
e^{j(\omega_0 k + kn)/2n^2)} \\
\vdots \\
e^{jL(\omega_0 k + kn)/2n^2)}
\end{bmatrix}.
\]

The solution to this minimisation, under the same constraint as in (6), is given by:

\[
h = Q^{-1} Z(Z^H Q^{-1} Z)^{-1} 1
\]
with
\[ Q = \hat{R}_s - G^H W^{-1} G, \] (15)
\[ G = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} w(n) x^H(n), \] (16)
and
\[ W = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} w(n) w^H(n). \] (17)

The LCMV filter in (10) and the APES based filter in (14) look very similar. The difference between the two filters is that the LCMV filter uses the covariance matrix of the observed signal, \( R_s \), whereas the covariance matrix used in the APES based filter, \( Q \), can be seen as an estimate of the noise covariance matrix.

### 4. Simulations

The two new harmonic chirp filters are compared to the harmonic LCMV [14] and APES based [17] filters. These filters are special cases of the harmonic chirp filters and are obtained by setting \( k = 0 \). The performance is measured by means of the output signal-to-noise ratio (oSNR),
\[ \text{oSNR}(h) = \frac{\sigma^2_{\tilde{v},w}}{\sigma^2_{\tilde{v},w}} = \frac{h^H R_s h}{h^H R_n h}, \] (18)
where \( R_s \) and \( R_n \) are the covariance matrices of desired signal and noise, and the signal reduction factor,
\[ \xi_{\text{sr}}(h) = \frac{\sigma^2_{\tilde{v},w}}{\sigma^2_{\tilde{v},w}} = \frac{\sigma^2_{\tilde{v},w}}{h^H R_n h}. \] (19)

The output SNR should be as high as possible whereas the signal reduction factor should be as close to one as possible to avoid signal distortion.

The filters were first tested on synthetic harmonic chirp signals made according to (3) through Monte Carlo simulations (MCS) [20]. The signals were generated with \( L = 10 \), \( A_i = 1 \) \( \forall i \), random phases, fundamental frequency and fundamental chirp rate in the intervals: \( \omega_i \in [0, 2\pi] \), \( f_0 \in [150, 250] \) Hz, \( k \in [0, 200] \) Hz\(^2\). The signals were added white Gaussian noise with a variance calculated to fit the desired input SNR,
\[ \text{iSNR} = \frac{\sigma^2_{\tilde{v},w}}{\sigma^2_{\tilde{v},w}} \] (20)

The signal and segment length were set to \( N = 200 \) and the filter length \( M = 50 \). The output SNR and signal reduction factor of the filter were calculated for each realisation of the chirp signal and averaged over 500 MCSs.

In Figs. 1 and 2 the output SNR and signal reduction factor are shown as a function of the input SNR. Five filters are compared in the figures. LCMV\(_{\text{opt}} \) is a chirp LCMV filter with the covariance matrix estimated directly from the noise signal, \( R_n \), and, therefore, it sets an upper limit for the performance of the filters but cannot be used in practice where there is no access to the clean noise signal. The other two LCMV filters are the chip LCMV (LCMV\(_c \)) and the harmonic LCMV (LCMV\(_h \)) and likewise with the two APES based filters, APES\(_c \) and APES\(_h \). The two APES based filters perform better than the corresponding LCMV filters, and the two chip based filters perform better than their harmonic counterparts. At low SNRs all filters perform almost equally, but when the input SNR is increased, the output SNR of the optimal LCMV filter and the chip APES based filter increases almost linearly whereas the output SNR of the other three filters falls off. The signal reduction factor for the optimal LCMV and the chip APES based filter is very close to one for all input SNRs whereas it increases with input SNR for the other filters.

The filters are next evaluated on a speech signal. The signal is a female speaker uttering the sentence "Why were you away a year, Roy?" sampled at \( f_s = 8000 \) Hz. To evaluate the potential of the methods, and since the focus is here on enhancement and not parameter estimation, the fundamental frequency, fundamental chirp rate and number of harmonics are estimated on the clean speech signal using nonlinear least squares (NLS) estimators [13, 14]. Again the noise is white Gaussian and added to give the desired input SNR.

The output SNR over time is shown in Fig. 3 for an input SNR of 10 dB. Except for very few points in time, the chip APES based filter is seen to set an upper limit to the performance of the four filters. The same tendency as for the synthetic signal is seen, with the APES based filters giving a higher output SNR than the LCMV filters and the chip versions performing better than the harmonic ones. The difference in output SNR for the two APES based filters, oSNR\(_{\text{APES}} = \text{oSNR}(\text{APES}_c) - \text{oSNR}(\text{APES}_h) \), is compared to the absolute value of the fundamental chirp rate in Fig. 4. Here, it is again seen that, except for a few places with small negative differences, the difference is positive, meaning that the chip APES...
Figure 3: Output SNR over time for a speech signal with input SNR = 10 dB.

Figure 4: Difference in output SNR between APES\(_c\) and APES\(_h\) from Fig. 3, oSNR\(_\Delta\), and the estimated chirp parameter, |\(k|\).

Based filter gives a higher output SNR than the harmonic APES based filter. In the figure it is also seen that the gain obtained by using the chirp APES based filter instead of the harmonic APES based filter is closely related to the estimated chirp parameter. When the absolute value of the chirp parameter is big, a gain in the oSNR is obtained whereas the gain is close to zero when the chirp parameter is close to zero. This makes sense if the harmonic chirp model describes the speech signal better than the harmonic model. If the fundamental frequency decreases or increases a lot in one segment of the signal, the chirp parameter will have a large absolute value, and the difference between the harmonic and harmonic chirp model will be large, and, thereby, there will be an advantage in using the harmonic chirp model. If the fundamental frequency is almost constant in a segment, the chirp parameter will be close to zero and the chirp harmonic model reduces to the harmonic model, leading to similar output SNRs for the two models.

5. Conclusions

In this paper, the non-stationarity of speech is taken into account to increase the performance of enhancement filters. The voiced speech was described with a harmonic chirp model and two filters based on the Linearly Constrained Minimum Variance (LCMV) filter and Amplitude and Phase Estimation (APES) based filter were presented and compared to their harmonic counterparts. It was shown that the chirp based filters perform better in terms of output SNR, signal distortion and PESQ score. As part of the derivation of the chirp APES based filter, a noise covariance matrix estimate is generated which can be used in other filters as, e.g., the Wiener filter.

Figure 5: Output SNR as a function of the input SNR for a speech signal.

Figure 6: Signal reduction factor, \(\xi_{sr}(h)\), as a function of the input SNR for a speech signal.

Figure 7: PESQ score as a function of the input SNR for a speech signal.
6. References


