Aalborg Universitet

AALBORG UNIVERSITY

DENMARK

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

Link to publication from Aalborg University

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, jrj, 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. As 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 big 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

is introduced in related fields. In [8, 9] a fan-chirp transform is suggested as an alternative to the traditional Fourier transform to analyse harmonic signals. The frequency is here allowed to vary linearly over time, leading to more sharp peaks in the spectrum when applied to a speech signal. Also in the field of speech recognition, non-stationary speech is taken into consideration by using gammachirp filters instead of traditional gammatone filters [10, 11], making the methods more robust to noise. In parameter estimation, a harmonic model extended to take non-stationarity into account has been considered in [12,13]. In [12], the basis is a very flexible model including both a chirp parameter to take changes over time into account and a detuning parameter which can account for individual variations away from the harmonic frequencies. The model is then approximated with a Taylor polynomial which leads to bigger and bigger deviations from the original model as the harmonic number increases, as is also mentioned in the paper. In [13], a harmonic chirp model is used to describe the voiced speech signal. This model has a harmonic structure, but the instantaneous fundamental frequency is allowed to change linearly within each segment, making the model capable of coping with non-stationary speech. The focus in these papers is, however, not on speech enhancement.

In this paper, we investigate the harmonic chirp model used in [13] in relation to speech enhancement. The model is compared to the traditional harmonic model [14], a common model used to describe voiced speech (see, e.g., [15–17]) which is the major component of speech signals. Voiced/unvoiced detectors [18] make it possible to discriminate voiced and unvoiced parts and only use the model on the relevant parts. The unvoiced parts can then be filtered by, e.g., a Wiener filter. The harmonic model assumes that the desired signal is composed of a set of sinusoids having frequencies given by an integer multiple of a fundamental frequency. In the traditional harmonic model, the fundamental frequency is assumed constant in segments of 20–30 ms, whereas the harmonic chirp model allows the fundamental frequency to vary linearly within each segment by introducing a chirp parameter in the model. In the harmonic framework, signals are often filtered by use of the Linearly Constrained Minimum Variance (LCMV) filter or the Amplitude and Phase EStimation (APES) based filter [14, 17, 19]. The principle in these filters is to pass the desired signal undistorted while the noise is reduced as much as possible. We derive the LCMV and APES based filters using the harmonic chirp model and compare their performance on synthetic and real speech signals to similar filters based on the traditional harmonic model. As a part of the derivation of the APES based filter, a noise covariance matrix estimate is obtained which takes the non-stationarity of speech into account.

In Section 2, the harmonic chirp model is introduced, in Section 3, the LCMV and APES based filters are derived according to the harmonic chirp model, and, in Section 4, their performance is compared to similar filters based on the har-

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, ω_l, can then be expressed as:

$$\omega_l(n) = l(\omega_0 + kn), \tag{1}$$

for time indices $n = 0, \cdots, N - 1$ where ω_0 is the normalised fundamental frequency and k is the fundamental chirp rate. The instantaneous phase, θ_l, of the harmonic components of the speech signal is given by the integral of the instantaneous frequency:

$$\theta_l(n) = l\left(\omega_0 n + \frac{1}{2}kn^2\right) + \phi_l \tag{2}$$

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

$$s(n) = \sum_{l=1}^{L} \alpha_l e^{jl(\omega_0 n + k/2n^2)}, \tag{3}$$

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, $\alpha_l = A_l e^{j\phi_l}$, 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, $x(n)$,

$$x(n) = s(n) + v(n). \tag{4}$$

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, $\sigma_x^2 = \sigma_s^2 + \sigma_v^2$.

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

$$\hat{s}(n) = \mathbf{h}^H \mathbf{x}(n) = \mathbf{h}^H \mathbf{s}(n) + \mathbf{h}^H \mathbf{v}(n), \tag{5}$$

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

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:

$$\min_{\mathbf{h}} \mathbf{h}^H \mathbf{R}_x \mathbf{h}, \quad \text{s.t.} \quad \mathbf{h}^H \mathbf{Z} = \mathbf{1}^T, \tag{6}$$

where $\mathbf{1} = [1 \cdots 1]^T$, $\mathbf{R}_x$ is the covariance matrix of the observed signal defined as:

$$\mathbf{R}_x = \mathbf{E}\{\mathbf{x}(n)\mathbf{x}^H(n)\}, \tag{7}$$

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

$$\mathbf{Z} = [\mathbf{z}(\omega_0, k)\, \mathbf{z}(2\omega_0, 2k)\, \cdots\, \mathbf{z}(L\omega_0, Lk)], \tag{8}$$

with

$$\mathbf{z}(l\omega_0, lk) = \begin{bmatrix} 1 \\ e^{jl(\omega_0 + k/2)} \\ \vdots \\ e^{jl(\omega_0(M-1) + k/2(M-1)^2)} \end{bmatrix}. \tag{9}$$

The solution to the minimisation problem is:

$$\mathbf{h} = \mathbf{R}_x^{-1} \mathbf{Z}(\mathbf{Z}^H \mathbf{R}_x^{-1} \mathbf{Z})^{-1} \mathbf{1}. \tag{10}$$

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

$$\widehat{\mathbf{R}}_x = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} \mathbf{x}(n)\mathbf{x}^H(n). \tag{11}$$

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:

$$\text{MSE} = \frac{1}{N - M + 1} \sum_{n=0}^{N-M} |\mathbf{h}^H \mathbf{x}(n) - \mathbf{a}^H \mathbf{w}(n)|^2, \tag{12}$$

where $\mathbf{a} = [\alpha_1\, \alpha_2\, \cdots\, \alpha_L]^H$ and

$$\mathbf{w}(n) = \begin{bmatrix} e^{j(\omega_0 n + k/2n^2)} \\ e^{j2(\omega_0 n + k/2n^2)} \\ \vdots \\ e^{jL(\omega_0 n + k/2n^2)} \end{bmatrix}. \tag{13}$$

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

$$\mathbf{h} = \mathbf{Q}^{-1} \mathbf{Z}(\mathbf{Z}^H \mathbf{Q}^{-1} \mathbf{Z})^{-1} \mathbf{1} \tag{14}$$

with

$$\mathbf{Q} = \widehat{\mathbf{R}}_x - \mathbf{G}^H \mathbf{W}^{-1} \mathbf{G}, \qquad (15)$$

$$\mathbf{G} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{x}^H(n), \qquad (16)$$

and

$$\mathbf{W} = \frac{1}{N-M+1} \sum_{n=0}^{N-M} \mathbf{w}(n)\mathbf{w}^H(n). \qquad (17)$$

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, $\mathbf{R}_x$, whereas the covariance matrix used in the APES based filter, $\mathbf{Q}$, 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),

$$\text{oSNR}(\mathbf{h}) = \frac{\sigma_{s,\text{nr}}^2}{\sigma_{v,\text{nr}}^2} = \frac{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}{\mathbf{h}^H \mathbf{R}_v \mathbf{h}}, \qquad (18)$$

where $\mathbf{R}_s$ and $\mathbf{R}_v$ are the covariance matrices of desired signal and noise, and the signal reduction factor,

$$\xi_{\text{sr}}(\mathbf{h}) = \frac{\sigma_s^2}{\sigma_{s,\text{nr}}^2} = \frac{\sigma_s^2}{\mathbf{h}^H \mathbf{R}_s \mathbf{h}}. \qquad (19)$$

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 $L = 10$, $A_l = 1 \forall l$, random phases, fundamental frequency and fundamental chirp rate in the intervals: $\phi_l \in [0, 2\pi]$, $f_0 \in [150, 250]$ Hz, $k \in [0, 200]$ Hz^2. The signals were added white Gaussian noise with a variance calculated to fit the desired input SNR,

$$\text{iSNR} = \frac{\sigma_s^2}{\sigma_v^2}. \qquad (20)$$

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, 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 chirp LCMV (LCMV$_\text{c}$) and the harmonic LCMV (LCMV$_\text{h}$) and likewise with the two APES based filters, APES$_\text{c}$ and APES$_\text{h}$. The two APES based filters perform better than the corresponding LCMV filters, and the two chirp based filters perform better
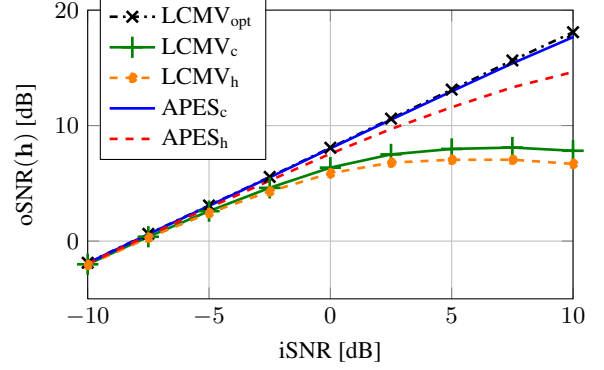

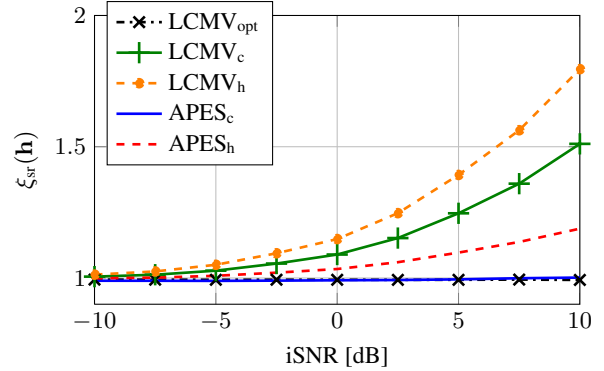Figure 1: Output SNR as a function of the input SNR for a synthetic chirp signal.

Figure 2: Signal reduction factor, $\xi_{\text{sr}}(\mathbf{h})$, as a function of the input SNR for a synthetic chirp signal.

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 chirp 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 chirp 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 chirp 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 chirp versions performing better than the harmonic ones. The difference in output SNR for the two APES based filters, oSNR$_\Delta$ = oSNR(APES$_\text{c}$) − oSNR(APES$_\text{h}$) 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 chirp APES
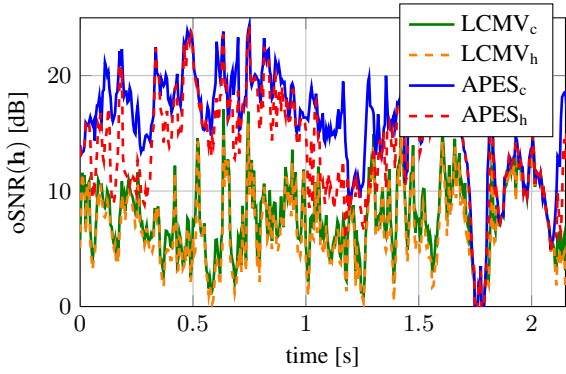
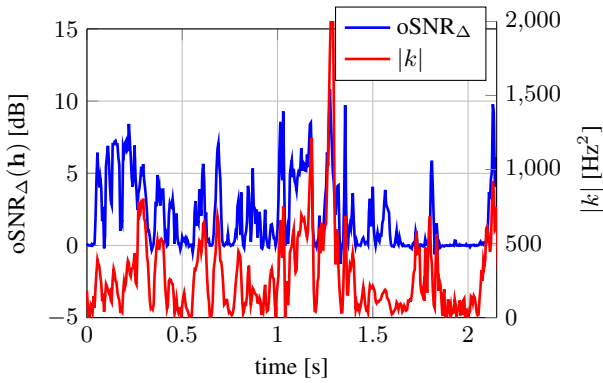Figure 3: Output SNR over time for a speech signal with input SNR = 10 dB.

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

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

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

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 de- 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.

In Figs. 5-7 the output SNR, signal distortion and Perceptual Evaluation of Speech Quality (PESQ) score [21] are shown as a function of the input SNR. The results are averaged over 50 Monte Carlo simulations. Here it is seen that the speech signal follows the same tendencies as the synthetic signal. The output SNRs of the filters are very similar to the output SNRs in the synthetic case, however, the signal distortion is increased for all filters, but the chirp APES based filter still has the lowest distortion. Also in terms of PESQ score the same conclusions can be drawn. The chirp filters perform better than their harmonic counterparts.
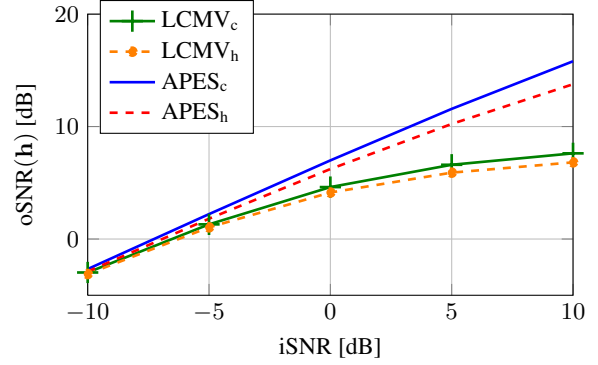
5. Conclusions

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.
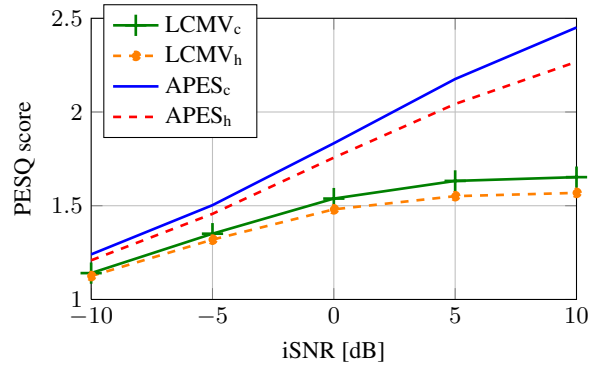

Figure 7: PESQ score as a function of the input SNR for a speech signal.

6. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[3] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, Jan. 1999.

[5] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[6] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech and Language Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[7] F. R. Drepper, "A two-level drive-response model of non-stationary speech signals," *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 1, pp. 125–138, Apr. 2005.

[8] M. Képesi and L. Weruaga, "Adaptive chirp-based time–frequency analysis of speech signals," *Speech communication*, vol. 48, no. 5, pp. 474–492, May 2006.

[9] L. Weruaga and M. Képesi, "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, vol. 87, no. 6, pp. 1504–1522, Jun. 2007.

[10] Z. Tüske, P. Golik, R. Schlüter, and F. R. Drepper, "Non-stationary feature extraction for automatic speech recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 5204–5207, May 2011.

[11] Z. Tüske, F. R. Drepper, and R. Schlüter, "Non-stationary signal processing and its application in speech recognition," *Workshop on statistical and perceptual audition*, Sep. 2012.

[12] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 2009, pp. 3985–3988.

[13] M. G. Christensen and J. R. Jensen, "Pitch estimation for non-stationary speech," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2014, accepted for publication.

[14] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[15] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, Sep. 2001, pp. 437–440.

[16] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech and Language Process. (TASLP)*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[17] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. on Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[18] K. Molla, K. Hirose, N. Minematsu, and K. Hasan, "Voiced/unvoiced detection of speech signals using empirical mode decomposition model," in *Int. Conf. Information and Communication Technology*, Mar. 2007, pp. 311–314.

[19] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Pearson Education, Inc., 2005.

[20] N. Metropolis, "The beginning of the monte carlo method," *Los Alamos Science*, no. 15, pp. 125–130, 1987.

[21] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.