5th Biennial Workshop on DSP for In-Vehicle Systems, Kiel, Germany, 2011

Generalized Spectral Refinement and its Applications

Mohamed Krini[1], Bernd Iser[1], Gerhard Schmidt[2]

[1]: SVOX Deutschland GmbH, Acoustic Signal Enhancement, Germany
[2]:Christian-Albrechts-Universität zu Kiel, Digital Signal Processing and System Theory, Germany
E-mail: mohamed.krini/bernd.iser@svox.com, gus@tf.uni-kiel.de

Abstract

In this contribution a generalized method for spectral refinement (GSR) is presented which is applied as a post-processing stage after a conventional frequency analysis of speech signals. The principle idea of GSR is to refine each subband signal after a frequency decomposition individually and to compute additional frequency supporting points in between using a linear combination of a the current as well preceding and successive short-term spectra. For its efficient implementation a simplification of the GSR method is derived – it can shown that the refinement can easily be implemented using short FIR-filters in each subband. This results in a very low computational complexity. The proposed method has been applied as a pre-processing stage for fundamental frequency estimation as well as for echo cancellation. Evaluations have shown that pitch frequency estimation methods can significantly be improved for all SNR considered when employing GSR method. Echo cancellation experiments confirm that GSR can enhance the performance by means of improved steady-state convergence.

Keywords Pitch frequency, echo cancellation, filterbank, spectral refinement, higher resolution, speech enhancement

1. INTRODUCTION

In different applications such as hands-free telephony or speech dialogue within a car, the desired speech signal is disturbed by the background noise (engine, wind noise, etc.). In order to reduce the disturbing components (while keeping the speech signal as natural as possible) speech enhancement algorithms are utilized. Most often enhancement algorithms like noise reduction or echo cancellation are applied in the subband domain to reduce computational complexity and to achieve fast convergence for adaptive filters [8, 9]. The microphone signal is usually first segmented into overlapping blocks of appropriate size ($20 - 30$ ms) and subsequently weighted with a window function. Afterwards, the windowed signal frames are transformed into the frequency domain using a DFT. The obtained short-term spectrum (STS) can be employed for estimating the power spectral density of the background noise which is required, e.g., for noise reduction. After several signal processing stages the enhanced STS is converted back to the time domain using an inverse DFT. The resulting overlapped signal blocks are added to obtain the broadband output signal. This type of overlap-add based scheme is also known as a DFT-modulated not-critically-sampled filterbank.

For the window often a Hann function is applied which on the one hand allows for an appropriately chosen subsampling factor (also known as frameshift) a perfect reconstruction at the output. On the other hand it shows good aliasing properties which is important for adaptive subband filters such as echo cancellation. However, windowing of successive signal blocks has most often the negative effect, that a significant frequency overlap of adjacent DFT subbands arises. Thus, adjacent fundamental frequency trajectories are sometimes hard to separate which is important for speech enhancement schemes that involve fundamental frequency estimation. In addition, aliasing components that appear due to large subsampling factors can remarkably degrade the convergence behavior of the echo cancellation schemes[5].

Increasing the DFT order to reduce the spectral overlap and the aliasing effects one should consider that for hands-free telephone systems several restrictions have to be fulfilled: the tolerable front-end delay of a hands-free system connected to a GSM network should not exceed 39 ms [6]. However, increasing e.g. the DFT order from $N = 256$ to $N = 512$ at a sampling frequency of $f_s = 11025$ Hz results in a delay of approx. 46 ms in the signal path, which doesn't fulfill ITU and ETSI recommendation. To overcome this, the herein proposed method for SR can be utilized. It is applied as a linear combination of a few weighted subband signal vectors at the output of a DFT.

The contribution is organized as follows: First, a brief overview about conventional methods will be given. Afterwards, the novel generalized method for SR as well as computational efficient approximations will be presented. In the next section applications of SR by means of improving fundamental frequency estimation schemes and echo cancellation systems are shown. The paper concludes with some simulation results and a conclusion.

2. CONVENTIONAL METHODS

To enhance the frequency selectivity of DFT-modulated filterbanks, they can be extended to a so-called non-critically

subsampled polyphase filterbanks [3]. In this case the length of the analysis and synthesis window functions are allowed to be larger than the number of used subbands (determined by the DFT order N). A polyphase filterbank introduces much lower aliasing components and the computational complexity is only increased marginally. As a consequence a frameshift close to the DFT order can be selected (depending on the used length of the prototype filters). In the literature (e.g. [16]) design procedures are described that achieve a frameshift of about $3/4\ N$ using filter orders of about $6-8\ N$. While a polyphase filterbank is able to reduces the computational complexity for large frameshifts one the one hand it also increases significantly the delay. Unfortunately, a high delay is very undesirable for applications such as hands-free telephony.

In [5] critically subsampled systems have been investigated. It was suggested to use adaptive cross filters in order to explicitly cancel the aliasing components. The consequences of using such cross filters lies in increasing the computational complexity significantly and it has been found to have problematic convergence speed.

In [12] a delayless structure has been proposed where adaptive filter weights are computed in the subband domain and then transformed to an equivalent time-domain filter. With this structure the filtering operation is performed in the time domain. A similar technique was developed in [2] and [15] for an acoustic echo canceller whereas the adaptive processing part takes place in the frequency domain. However, all mentioned time-domain based filtering approaches have the consequences of higher computation complexity. Also, mixed schemes (the first part of the impulse response is convolved in the time domain, the remaining part in the subband domain) have been published by various authors.

The [13] the authors addressed the issues of computational complexity and delay of subband adaptive filtering for applications of acoustic echo control. It has been suggested to use filterbanks based on allpass polyphase IIR structure as an alternative to the FIR based filterbanks. The use of allpass polyphase IIR filter banks achieve very high sidelobe attenuation and it has been shown to be computational efficient while keeping the aliasing components low. It has to be noticed that with this approach non-linear phase distortions and appearance of narrowband high energy aliasing terms at the filter boundaries arise.

In [7] an efficient prototype filter design method for an oversampled DFT filterbank has been proposed where the aliasing components are minimized while the total filterbank group delay is pre-specified. It could be shown that the estimation accuracy for non-critical decimated filterbanks is close to the fullband solution and significantly better than for the critically decimated case.

In contrast to the state-of-the-art approaches, the herein proposed method for GSR is employed as a post-processor for analysis filterbanks. The enhanced frequency selectivity of the analysis is achieved either by reducing the spectral overlap of adjacent subbands or by computing additional subbands. The refinement procedure can easily be implemented using short FIR filters in each subband channel – this results in a very low computational complexity and an insignificant additional delay in the signal path.

3. GENERALIZED SPECTRAL REFINEMENT

In contrast to the contribution in [11] a generalized method for spectral refinement method will be derived in the following which is applied afterwards to enhance pitch frequency estimation and to improve the performance of echo cancellation. For the derivation of the generalized spectral refinement (GSR) method a basic DFT of order N and an increased DFT of order

$$\widetilde{N} = N + k_0\, r \qquad \text{with} \qquad k_0 \in \Big\{1, 2, 3, \ldots\Big\} \tag{1}$$

are assumed, whereas r corresponds to the used subsampling factor. All quantities in this contribution that characterize a high order will be designated with a tilde symbol. As input we define an overlapped and windowed input signal frame of a low order N as well as of a high order $\widetilde{N}$. Applying DFTs to the weighted input signal vectors result in N and $\widetilde{N}$ frequency supporting points of the short-term spectra:

$$Y\left(e^{j\Omega_\mu}, n\right) = \sum_{k=0}^{N-1} y(nr-k)\, h_k\, e^{-j\Omega_\mu k}\,, \tag{2}$$

$$\widetilde{Y}\left(e^{j\widetilde{\Omega}_{\tilde{\mu}}}, n\right) = \sum_{k=0}^{\widetilde{N}-1} \widetilde{y}(nr-k)\, \widetilde{h}_k^{(\tilde{\mu})}\, e^{-j\widetilde{\Omega}_{\tilde{\mu}} k}\,, \tag{3}$$

whereas the parameters n and h_k characterize the frame index and the chosen window function respectively (Eqn. (2) and (3) can also be interpreted as subband signals of an analysis filterbank). Note that for calculating the refined short-term spectrum $\widetilde{Y}(e^{j\widetilde{\Omega}_{\tilde{\mu}}}, n)$ individual window functions $\widetilde{h}_k^{(\tilde{\mu})}$ are utilized. This is specified in order to allow different subband filters within a filterbank. The used frequency supporting points Ω_μ and $\widetilde{\Omega}_{\tilde{\mu}}$ are equidistantly distributed over the normalized frequency range:

$$\Omega_\mu = \frac{2\pi}{N}\mu \quad \text{with } \mu \in \Big\{0, \ldots, N-1\Big\} \quad \text{and} \tag{4}$$

$$\widetilde{\Omega}_{\tilde{\mu}} = \frac{2\pi}{\widetilde{N}}\tilde{\mu} \quad \text{with } \tilde{\mu} \in \Big\{0, \ldots, \widetilde{N}-1\Big\}. \tag{5}$$

For the sake of simplicity the short-term spectra $Y(e^{j\Omega_\mu}, n)$ and $\widetilde{Y}(e^{j\widetilde{\Omega}_{\tilde{\mu}}}, n)$ will be rewritten in matrix-vector notation:

$$\boldsymbol{Y}\left(e^{j\Omega}, n\right) = \boldsymbol{D}\,\boldsymbol{H}\,\boldsymbol{y}(n)\,, \tag{6}$$

$$\widetilde{\boldsymbol{Y}}\left(e^{j\widetilde{\Omega}}, n\right) = \boldsymbol{P}^{(\widetilde{N})}\,\widetilde{\boldsymbol{D}}_{\text{Block}}\,\widetilde{\boldsymbol{H}}_{\text{Block}}\,\boldsymbol{I}^{(\widetilde{N})}\,\widetilde{\boldsymbol{y}}(n)\,. \tag{7}$$

The quantity $\boldsymbol{D}$ characterize a DFT matrix of order N and $\boldsymbol{H}$ denotes a diagonal matrix which consists of window function coefficients according to:

$$\begin{aligned}\boldsymbol{H} &= \text{diag}\{\, h_0,\, h_1, \ldots,\, h_{N-1}\} \qquad (8)\\ &= \begin{bmatrix} h_0 & 0 & 0 & 0 \\ 0 & h_1 & 0 & 0 \\ & & \ddots & \\ 0 & 0 & 0 & h_{N-1} \end{bmatrix}.\end{aligned}$$

To formulate the higher resolution STS from Eq. (3) in matrix-vector notation the input signal is duplicated $\widetilde{N}$-times. This is done by multiplying the input vector $\widetilde{\boldsymbol{y}}(n)$ with a so-called *block unit matrix* defined as:

$$\boldsymbol{I}^{(\widetilde{N})} = \left[\; \boldsymbol{I}_0^{(\widetilde{N})},\;\; \boldsymbol{I}_1^{(\widetilde{N})},\;\; \ldots,\;\; \boldsymbol{I}_{\widetilde{N}-1}^{(\widetilde{N})} \;\right]^{\mathrm{T}}, \qquad (9)$$

whereas each element characterizes an identity matrix: $\boldsymbol{I}_\mu^{(\widetilde{N})} = \text{diag}\{\mathbf{1}\}$ of size $\widetilde{N}$. The resulting vectors are subsequently multiplied by a block-diagonal window matrix:

$$\widetilde{\boldsymbol{H}}_{\text{Block}} = \text{diag}\left\{\; \widetilde{\boldsymbol{H}}_0,\;\; \widetilde{\boldsymbol{H}}_1,\;\; \ldots,\;\; \widetilde{\boldsymbol{H}}_{\widetilde{N}-1} \;\right\}, \qquad (10)$$

with

$$\widetilde{\boldsymbol{H}}_\mu = \text{diag}\left\{\; \widetilde{h}_0^{(\mu)},\;\; \widetilde{h}_1^{(\mu)},\;\; \ldots,\;\; \widetilde{h}_{\widetilde{N}-1}^{(\mu)} \;\right\}. \qquad (11)$$

Afterwards, a DFT is performed for each weighted input signal vector. This is accomplished by using a block-diagonal DFT matrix $\widetilde{\boldsymbol{D}}_{\text{Block}}$ according to

$$\widetilde{\boldsymbol{D}}_{\text{Block}} = \text{diag}\left\{\; \widetilde{\boldsymbol{D}}_0,\;\; \widetilde{\boldsymbol{D}}_1,\;\; \ldots,\;\; \widetilde{\boldsymbol{D}}_{\widetilde{N}-1} \;\right\}, \qquad (12)$$

whereas the diagonal elements of Eq. (12) specify DFT matrices of order $\widetilde{N}$. Once the DFT matrices are applied to individual weighted input signal segments a vector (of lenght $\widetilde{N}\widetilde{N}$) results. The vector contains $\widetilde{N}$ short-term spectra. Thus, the outcome represents the short-term spectra of the current signal frame weighted with $\widetilde{N}$ individual window functions.

Finally, a so-called *selection matrix* $\boldsymbol{P}^{(\widetilde{N})}$ (of dimension $\widetilde{N}\text{x}\widetilde{N}\widetilde{N}$) is multiplied with the short-term spectra in order to select the desired output subband signals $\widetilde{Y}(e^{j\widetilde{\Omega}_{\tilde{\mu}}}, n)$. The *selection matrix* is defined as

$$\boldsymbol{P}^{(\widetilde{N})} = \left[\; \boldsymbol{P}_0^{(\widetilde{N})},\;\; \boldsymbol{P}_1^{(\widetilde{N})},\;\; \cdots,\;\; \boldsymbol{P}_{\widetilde{N}-1}^{(\widetilde{N})} \;\right]. \qquad (13)$$

Each element of Eq. (13) characterizes a diagonal matrix of order $\widetilde{N}$ defined as:

$$\boldsymbol{P}_\mu^{(\widetilde{N})} = \text{diag}\left\{\; p_{\mu,0},\;\; p_{\mu,1},\;\; \ldots,\;\; p_{\mu,\widetilde{N}-1} \;\right\}. \qquad (14)$$

The elements $p_{\mu,i}$ from Eq. 14 have to fulfill the following condition:

$$p_{\mu,i} = \begin{cases} 1\,, & \text{if} \quad i = \mu\,, \\ 0\,, & \text{else}\,. \end{cases} \qquad (15)$$

Once we have formulated the vector-matrix notation of the short-term spectra using a basic and increased DFT of order N and $\widetilde{N}$, respectively, the next step will be the derivation of a general solution for spectral refinement. The principle idea of the GSR method is to determine a refined STS, $\widetilde{\boldsymbol{Y}}\left(e^{j\Omega}, n\right)$, by using the current spectrum $\boldsymbol{Y}\left(e^{j\Omega}, n\right)$ and a number of time-delayed spectra $\boldsymbol{Y}\left(e^{j\Omega}, n-k\right)$ of lower order N without the need for an additional DFT of higher order $\widetilde{N}$:

$$\boldsymbol{S}\begin{bmatrix} \boldsymbol{Y}\left(e^{j\Omega}, n\right) \\ \vdots \\ \boldsymbol{Y}\left(e^{j\Omega}, n-(M-1)\right) \end{bmatrix} = \widetilde{\boldsymbol{Y}}\left(e^{j\tilde{\Omega}}, n\right). \qquad (16)$$

The matrix $\boldsymbol{S}$ refers to the SR matrix with a dimension of $\widetilde{N}\,\text{x}\,NM$, whereas M is the number of the input spectra each shifted by a frameshift of r samples. It is assumed that the lower order STSs $\boldsymbol{Y}\left(e^{j\Omega}, n\right)$ are already available. The might be used, e.g., to estimate the noise power for speech enhancement within a hands-free system. However, in some situations it is desired to determine a higher resolution STS in order to enhance feature extractions schemes such as pitch frequency estimation. For that purpose we suggest to apply a linear combination of the lower order short-term spectra as stated in Eq. (16). The derivation of the generalized SR matrix $\boldsymbol{S}$ as well as its simplified version will be explained in the following.

3.1. Determining the Spectral Refinement Matrix

Before calculating the SR matrix $\boldsymbol{S}$ a constraint for the higher order window functions $\widetilde{\boldsymbol{h}}_\mu$ is introduced:

$$\boldsymbol{A}^{(\tilde{\mu})}\;[\boldsymbol{h},\, \boldsymbol{h},\, ...,\, \boldsymbol{h}]^{\mathrm{T}} = \widetilde{\boldsymbol{h}}^{(\tilde{\mu})}. \qquad (17)$$

This allows for an efficient implementation – as we will see later on. The matrix $\boldsymbol{A}^{(\tilde{\mu})}$ of size $\widetilde{N}\,\text{x}\,MN$ consists of appropriate weights $a_k^{(\tilde{\mu},m)}$ with $k \in [0, N-1]$ and $m \in [0, M-1]$. The structure of the matrix is shown in Eq. (18).

The main task of $\boldsymbol{A}^{(\tilde{\mu})}$ is to weight M window functions of lower order N and shift subsequently adjacent window functions by the chosen subsampling factor r. The so-obtained modified window functions were summed up to obtain a desired higher order window function. Consequently, the window functions $\widetilde{\boldsymbol{h}}^{(\tilde{\mu})}$ consists of a weighted sum of shifted window functions $\boldsymbol{h}$. The coefficients $a_{k,m}^{(\tilde{\mu})}$ can be designed in such a way that a set of low-order window functions are transformed into a desired window function of higher order. The resulting order of the window function $\widetilde{\boldsymbol{h}}^{(\tilde{\mu})}$ from Eq. 17 is given by:

$$\widetilde{N} = N + r\,(M-1)\,. \qquad (19)$$

In the upper part of Fig. 1 an example of weighted and shifted Hann-windows each of lower order (dashed lines with

$$\boldsymbol{A}^{(\tilde{\mu})} = \begin{bmatrix} a_0^{(\tilde{\mu},0)} & \cdots & 0 & & \cdots & 0 & & \cdots & 0 \\ \ddots & \ddots & \vdots & \ddots & \ddots & \vdots & \cdots & \ddots & \vdots \\ a_{r-1}^{(\tilde{\mu},0)} & \ddots & a_0^{(\tilde{\mu},1)} & \ddots & \ddots & & \ddots & & \\ \ddots & & \ddots & a_{r-1}^{(\tilde{\mu},1)} & \ddots & & a_0^{(\tilde{\mu},M-1)} & & \\ a_{N-r-1}^{(\tilde{\mu},0)} & \ddots & & \ddots & a_{N-r-1}^{(\tilde{\mu},1)} & \ddots & \ddots & a_{r-1}^{(\tilde{\mu},M-1)} & \\ \ddots & & & & \ddots & & & \ddots & \\ a_{N-1}^{(\tilde{\mu},0)} & & & & a_{N-1}^{(\tilde{\mu},1)} & & & a_{N-1-r}^{(\tilde{\mu},M-1)} & \\ \vdots & \ddots & \vdots & \ddots & & \vdots & \ddots & & \ddots \\ 0 & \cdots & 0 & \cdots & & 0 & \cdots & & a_{N-1}^{(\tilde{\mu},M-1)} \end{bmatrix} \quad (18)$$

$N = 256$, $M = 5$, $r = 64$) as well as the resulting window function of higher order (solid line with $\widetilde{N} = 512$) is shown. The coefficients used for weighting the window func-

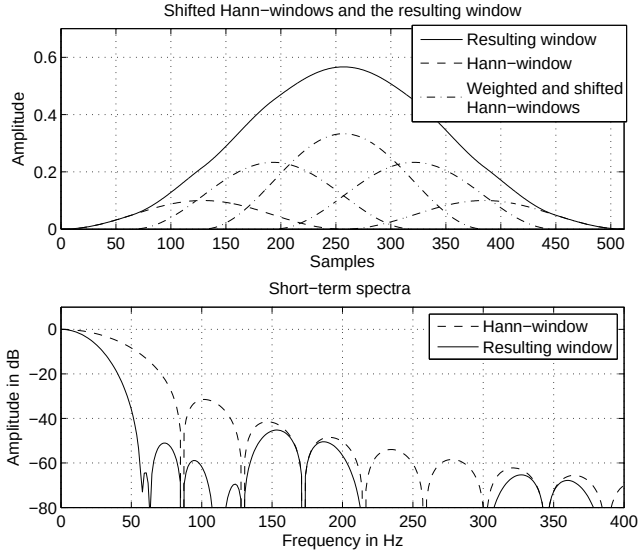

Fig. 1. Upper part shows the weighted and shifted Hann-windows and the resulting window function, lower part depicts the corresponding spectra.

tions have been chosen as follows: $a_k^{(0)} = a_k^{(M-1)} = 0.3\,K_0$, $a_k^{(1)} = a_k^{(M-2)} = 0.7\,K_0$ and $a_k^{((M-1)/2)} = K_0$.[1] As normalization constant $K_0 = 3$ has been applied. In the lower part of Fig. 1 the corresponding analyses of the short-term spectra are depicted. By comparing the results one can see that the main lobe width as well as the side-lobe amplitudes are reduced when using the weighted sum of shifted window functions $\boldsymbol{h}$.

Once the constraint for the window functions is defined, the next step will be the solve the equation system for the SR matrix $\boldsymbol{S}$. By doing so first Eq. (16) is rewritten as follows:

$$\boldsymbol{S}\,\boldsymbol{D}_{\text{Block}}\,\boldsymbol{H}_{\text{Block}}\,\boldsymbol{Y}(n) \;=\; \widetilde{\boldsymbol{Y}}\big(e^{j\widetilde{\Omega}}, n\big)\,. \quad (20)$$

The vector $\widetilde{\boldsymbol{Y}}(n)$ consists of the current input signal frame as well as of the delayed ones each of lower order N:

$$\boldsymbol{Y}(n) \;=\; \big[\boldsymbol{y}(nr), \ldots, \boldsymbol{y}\big(nr-(M-1)\big)\big]^{\mathrm{T}}. \quad (21)$$

The quantities $\boldsymbol{H}_{\text{Block}}$ and $\boldsymbol{D}_{\text{Block}}$ characterize *block-diagonal* matrices of the lower order window function and the DFT. Using the above mentioned constraint (Eq. 17), the higher resolution STS from Eq. 7 can be expressed as follows:

$$\widetilde{\boldsymbol{Y}}\big(e^{j\widetilde{\Omega}}, n\big) \;=\; \boldsymbol{P}^{(\widetilde{N})}\,\widetilde{\boldsymbol{D}}_{\text{Block}}\,\boldsymbol{A}^{(\tilde{\mu})}\,\boldsymbol{I}^{(\widetilde{N}M)}\,\boldsymbol{H}_{\text{Block}}\,\boldsymbol{Y}(n)\,. \quad (22)$$

The *block-unit matrix* $\boldsymbol{I}^{(\widetilde{N}M)}$ comprises N identity matrices each of size NM (analogue to Eq. 9).

Inserting the expression from Eq. 22 in Eq. 20 results in several solutions for the matrix $\boldsymbol{S}$, that depend in general on the input signal vectors $\boldsymbol{y}(n-m)$. A Solution that is independent of the input signal can be obtained by:

$$\boldsymbol{S} \;=\; \boldsymbol{P}^{(\widetilde{N})}\,\widetilde{\boldsymbol{D}}_{\text{Block}}\,\boldsymbol{A}^{(\tilde{\mu})}\,\boldsymbol{I}^{(\widetilde{N}M)}\,\boldsymbol{D}_{\text{Block}}^{-1}\,. \quad (23)$$

After inserting the definitions of the matrices in Eq. 23 the SR matrix $\boldsymbol{S}$ can finally be rewritten in the following way:

$$S_{i,mN+l} \quad (24)$$

$$= \;\frac{1}{N}\sum_{\mu=0}^{\widetilde{N}-1}\sum_{z=0}^{\widetilde{N}-1} P_{i,z+\mu\widetilde{N}}\sum_{k=0}^{N-1} e^{-j\frac{2\pi}{\widetilde{N}}z(k+mr)}\,a_k^{(\tilde{\mu},m)}\,e^{j\frac{2\pi}{N}kl}\,.$$

The parameter i in Eq. 24 specifies the row and the quantity $mN + l$ the column of the SR matrix.

4. SIMPLIFIED VERSION OF SPECTRAL REFINEMENT

Once the general solution for the SR matrix was formulated it is now checked due to sparsely population in the following. In doing so, first it is assumed that the weighting coefficients for each m-th window function are identical, meaning that $a_k^{(\tilde{\mu},m)} = a^{(\tilde{\mu},m)}$. In order to analyse the SR matrix quantitatively, the formulation of Eq. 24 is rewritten as follows:

$$S_{i,mN+l} \;=\; \frac{a^{(i,m)}}{N}\,e^{-j\frac{2\pi}{\widetilde{N}}imr}\sum_{k=0}^{N-1} e^{-j2\pi\left(\frac{i}{\widetilde{N}}-\frac{l}{N}\right)k}\,. \quad (25)$$

[1] For the sake of simplicity the superscript parameter μ has been omitted.

Further on, the geometric series on the left hand side of Eq. 25 can be dissolved in the following way:

$$S_{i,mN+l} = \frac{a^{(i,m)}}{N} \frac{\sin\left(\pi\left(\frac{iN-l\widetilde{N}}{\widetilde{N}}\right)\right) e^{-j\pi\left(\frac{iN-l\widetilde{N}}{N}\right)}}{\sin\left(\pi\left(\frac{iN-l\widetilde{N}}{N\widetilde{N}}\right)\right) e^{-j\pi\left(\frac{iN-l\widetilde{N}}{N\widetilde{N}}\right)}} e^{-j\frac{2\pi}{\widetilde{N}}imr}. \tag{26}$$

If the condition holds, that the higher filter order is a multiple of the lower filter order: $\widetilde{N} = \widetilde{k}_0\, N$ with $\widetilde{k}_0 \in \{2, 3, \ldots\}$, then specific rows and columns of the SR matrix can be further simplified to:

$$S_{i,mN+l} = \begin{cases} 0\,, & \text{if } \left(\frac{i}{\widetilde{k}_0} \in \mathbb{Z}\right) \wedge \left(l \neq \frac{i}{\widetilde{k}_0}\right), \\ a^{(i,m)}\; e^{-j\frac{2\pi}{\widetilde{N}}imr}\,, & \text{if } \left(\frac{i}{\widetilde{k}_0} \in \mathbb{Z}\right) \wedge \left(l = \frac{i}{\widetilde{k}_0}\right), \\ S_{i,mN+l}\,, & \text{else}\,. \end{cases} \tag{27}$$

The symbol $\mathbb{Z}$ denotes the set of integers. Thus, each $\widetilde{k}_0^{\text{th}}$ row of $\boldsymbol{S}$ is sparsely populated, i.e., the elements of each $\widetilde{k}_0^{\text{th}}$ row are zero except of the column indices that are multiples of N. Furthermore, if the filter order N is chosen to be a multiple of the used frameshift, e.g. $2\,r$ or $4\,r$, then those elements of the sparsely populated rows of the SR matrix are either real or imaginary.

For illustration purposes a simple example of the SR matrix is shown in Fig. 2 with $M = 3$, $r = 2$, and $N = 4$. As a result, each second row (even indices of the SR matrix)

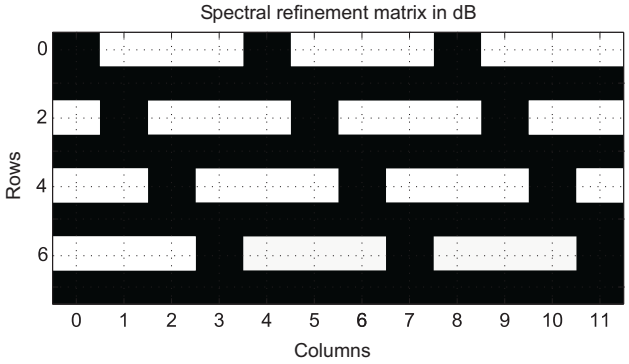

Fig. 2. SR matrix: White color indicate values equal zero and black elements values unequal zero.

is sparsely populated. The elements in white color indicate values equal zero, whereas the ones in black values unequal zero. However, these rows are related to that frequency supporting points, which would be computed with a basic DFT of order N as well as with a higher order DFT of $\widetilde{N}$.

4.1. Realization of spectral refinement

The proposed method for spectral refinement can either be applied to refine only the original frequency resolution of the input signal or to compute additional frequency supporting points in between:

4.1.1. Refinement of the original frequency resolution

If it is desired to calculate a spectral refinement of the original frequency resolution – i.e., each $\tilde{k}_0^{\text{th}}$ frequency supporting point of the vector $\widetilde{Y}(e^{j\widetilde{\Omega}_\mu}, n)$ is refined – the realization of the spectral refinement can be performed in a efficient and reliable manner. Due to the sparsely population of the matrix $\boldsymbol{S}$ the SR can be realized by short FIR-filters applied in each subband after the frequency decomposition of the input signal $y(n)$. The FIR filter coefficients

$$\boldsymbol{g}_{i,i\cdot\tilde{k}_0} \;=\; \left[g_{i,i\cdot\tilde{k}_0,0},\, g_{i,i\cdot\tilde{k}_0,1},\, \ldots,\, g_{i,i\cdot\tilde{k}_0,M-1}\right]^{\mathrm{T}} \tag{28}$$

are extracted from the sparsely populated SR matrix by:

$$g_{i,i\cdot\tilde{k}_0,m} \;=\; S_{i\cdot\tilde{k}_0,i+mN}\,. \tag{29}$$

The refined spectrum for the i-th subband is determined by:

$$\begin{aligned}\widetilde{Y}\left(e^{j\widetilde{\Omega}_{i\cdot\tilde{k}_0}}, n\right) \;&=\; g_{i,i\cdot\tilde{k}_0,0}\, Y\left(e^{j\Omega_i}, n\right) + \ldots \\ &\quad + g_{i,i\cdot\tilde{k}_0,M-1}\, Y\left(e^{j\Omega_i}, n-(M-1)\right).\end{aligned} \tag{30}$$

Often analysis-synthesis schemes use a frameshift which is a multitude of the DFT order. For such cases the filter coefficients $g_{i,i\cdot\tilde{k}_0,m}$ are either real or imaginary which in turn results in a further reduction of the computational cost.

Fig. 3 shows a realization of an analysis filterbank with SR as a post-processor by means of FIR-filters for $\tilde{k}_0 = 2$. The "auto" FIR filter $\boldsymbol{g}_{i,i\cdot\tilde{k}_0}$ applied to refine the original frequency supporting points are depicted in black color.

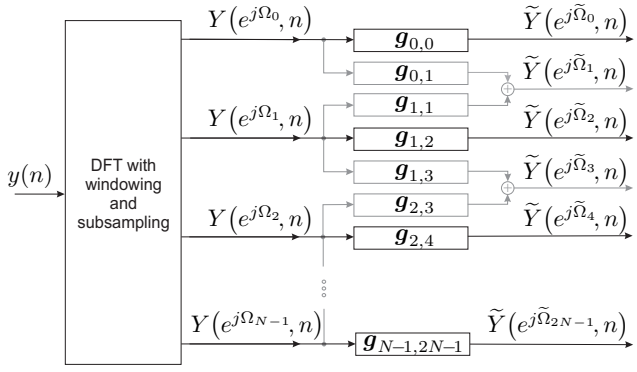

Fig. 3. Analysis filterbank with SR as a post-processor by means of FIR-filters for $\tilde{k}_0 = 2$.

4.1.2. Computation of additional frequency supporting points

Beside refinement of the original frequency resolution it is also possible to calculate frequency points in between of the

original spectrum. At a first glance, however, it's computationally intensive due to the non-sparseness of the remaining rows of the SR matrix. In order to reduce the computational complexity, one can approximate the non-sparse rows of the SR matrix by the M largest coefficient pairs. The largest coefficient pairs correspond exactly to that weighting values around the desired frequency supporting points of the short-term spectrum. Analyses have confirmed that the resulting spectrum astonishingly shows low errors even if only $M = 3 \dots 5$ filter coefficients are used. The complete system of spectral refinement for $\tilde{k}_0 = 2$ is depicted in Fig. 3. The refinement of the original frequency resolution is accomplished using "auto" FIR filters (drew in black color) and the computation of the additional frequency supporting points are performed using "cross" FIR filters (drew in grey color). The "cross" as well as the "auto" filters can be calculated as:

$$g_{i,l,m} \;=\; S_{l,i+mN}\,, \tag{31}$$

and the refined STS, $\widetilde{Y}\big(e^{j\widetilde{\Omega}_l}, n\big)$, is finally determined by:

$$\widetilde{Y}\big(e^{j\widetilde{\Omega}_l}, n\big) \tag{32}$$

$$\approx \begin{cases} \sum\limits_{m=0}^{M-1} g_{l/\tilde{k}_0,l,m}\, Y\big(e^{j\widetilde{\Omega}_{l/\tilde{k}_0}}, n-m\big), & \text{if} \quad \frac{l}{\tilde{k}_0} \in \mathbb{Z}, \\ \sum\limits_{m=0}^{M-1} g_{\lfloor l/\tilde{k}_0 \rfloor,l,m}\, Y\big(e^{j\widetilde{\Omega}_{\lfloor l/\tilde{k}_0 \rfloor}}, n-m\big) & \text{else}, \\ \quad + \sum\limits_{m=0}^{M-1} g_{\lceil l/\tilde{k}_0 \rceil,l,m}\, Y\big(e^{j\widetilde{\Omega}_{\lceil l/\tilde{k}_0 \rceil}}, n-m\big), \end{cases}$$

where $\lfloor\ \rfloor$ and $\lceil\ \rceil$ denote rounding to the next smaller and larger integer, respectively.

4.2. Computational complexity of spectral refinement

After the simplified version of SR and its efficient realization were described, as a next stage we analyze its overall performance. Hence, the computational complexity of a 256-FFT order with additional SR is compared with a 512-FFT order by means of complex multiplications and additions as shown in Tab. 1. Remarkable to this result is the need of only few operations for refining the original frequency supporting points. Using SR as a post processing stage of a basic 256-FFT only about 2688 complex multiplications and additions are required while doubling the basic 256-FFT order to 512-FFT about 4608 operations are needed. It has to be mentioned that in many applications a basic FFT is already available needed for estimating several parameters, like pitch frequency used for speech recognition. In such situations, however, only minor additional operations are required for performing SR. If it is desired to calculate also additional frequency supporting points as presented in Sec. 4.1.2 only few complex multiplications and additions have to be added to the refinement system as seen in Tab. 1.

Complex multiplications and additions for $\widetilde{N} = 512,\ N = 256,\ M = 3 \dots 5$	
$\widetilde{N}$-order FFT	$\widetilde{N}\,\mathrm{ld}(\widetilde{N}) = 4608$
N-order FFT with spectral refinement (SR)	$N\,\mathrm{ld}(N) + MN/2$ $= 2432 \dots 2688$
N-order FFT with SR and additional frequencies	$N\,\mathrm{ld}(N) + MN/2 + MN$ $= 3200 \dots 3968$

Table 1. Computation complexity of a higher order FFT and of a basic FFT with additional SR.

5. APPLICATIONS

The SR method can be applied in a variety of audio signal processing applications. In this contribution two applications will be presented in more detail: The principle idea of using SR as a pre-processing stage for enhanced fundamental frequency estimation and the employment of SR for echo cancelation to achieve a higher steady-state convergence.

5.1. Spectral Refinement for Pitch Estimation

A broad variety of different algorithms for estimating the fundamental frequency of speech signals exists, like methods based on the harmonic product-spectrum [14] or on short-term auto-correlation [1]. For the following evaluations a method based on the last-mentioned approach has been employed: At the first stage the corrupted speech signal $y(n)$ is divided into overlapping blocks and subsequently windowed. Once the FFT as well as the SR method are applied to the input signal block according to Fig. 3, the short-term power spectral density (PSD) is estimated. Applying the IFFT to a normalized version of the PSD the auto-correlation function (ACF) results. Performing a maximum search of the ACF in a selected range of indices, the normalized pitch period is estimated. Finally, the pitch frequency is obtained using the inverse of the pitch period. Further details can be found in [10].

To show the performance and the accurateness of the proposed method, the estimated fundamental frequencies without and with SR at different SNR conditions have been compared with a clean speech laryngograph database. The reference database consists of a multitude of pitch frequencies out of the interval $\hat{f}_\mathrm{p}(n) \in [50\,\mathrm{Hz}, 350\,\mathrm{Hz}]$. For the evaluation the correctness and the false detection have been considered. To analyze the correctness of the estimated pitch frequencies three range of values have been defined: the estimated error lies within a tolerance range of ± 3 %, ± 10 % and ± 20 %. False detection means that the algorithm under test detects a pitch frequency while no reference pitch is available.

Tab. 2 depicts the correctness of the pitch estimation method without and with SR for high, medium, and low SNR. A

	Accepted tolerance	Correctness [%]		
		High SNR	Mean SNR	Low SNR
Standard method	$< 3\,\%$	62.1	70.1	47.2
	$< 10\,\%$	65.1	70.9	48.4
	$< 20\,\%$	65.5	71.4	49.3
Method with SR up to 1 kHz	$< 3\,\%$	82.1	80.3	55.9
	$< 10\,\%$	88.1	85.3	58.2
	$< 20\,\%$	88.8	86.2	58.7
Method with SR up to 3 kHz	$< 3\,\%$	83.2	80.1	53.4
	$< 10\,\%$	89.8	85.3	56.9
	$< 20\,\%$	90.6	86.3	57.5

Table 2. Correctness of estimated fundamental frequencies without and with SR for different tolerance ranges.

pitch was only detected if the normalized ACF at maximum lag exceeds a predefined treshold of $p_0 = 0.25$. Note, that the refinement was only performed at lower frequencies up to 1 kHz and 3 kHz, respectively. The results show that by applying SR an increase of correctness by about $20 - 25\,\%$ (abs.) at high SNR is achieved, approx. $10 - 15\,\%$ (abs.) at medium SNR, and about $8 - 10\,\%$ (abs.) at low SNR. Moreover, it can be observed that nearly the same performance is achieved when using SR up to 1 and 3 kHz. Hence, for pitch estimation it is sufficient to refine the input spectrum only at lower frequencies up to 1 kHz which in turn results in a significant reduction of the computational complexity.

The measured results for miss detections are listed in Tab. 3. From the evaluations one can see that the miss detection rates can nearly be kept constant for all SNR considered while the correctness rates are increased at the same time.

	Missdetection [%]		
	High SNR	Mean SNR	Low SNR
Standard method	18.1	11.4	8.6
Method with SR up to 1 kHz	18.2	11.3	8.2
Method with SR up to 3 kHz	17.2	11.4	8.1

Table 3. Miss detection of pith frequency without and with SR up to 1 kHz and 3 kHz for different SNR's.

5.2. Spectral Refinement for Echo Cancellation

The use of echo cancellation by means of adaptive filters offers the possibility of a full-duplex communications. Due to computational complexity often adaptive filters in the subband domain are used as a digital replica of a loudspeaker-enclosure-microphone (LEM) system [8].

For estimating the subband echo signal first the microphone signal as well as the reference signal have to be segmented into overlapping blocks of appropriate sizes and afterwards each block is transformed into the frequency domain. Usually successive segments are overlapping out of the range of $50 - 75\%$. However, increasing the subsampling rate r the computational load is decreased while at the same time aliasing components within subband signals are increased. It is well known that subband echo cancellation requires nearly aliasing free subband signals. Therefore, a compromise between subsampling and computational cost has to be found.

To overcome a low steady-state convergence of echo cancellation at large subsampling rates, the proposed SR method can be implemented in the microphone as well as in the reference path. In doing so, the refined reference subband signal is convolved with an unknown LEM subband impulse response to estimate the subband echo signal. Afterwards the echo signal is subtracted from the refined microphone subband signal to determine the error subband signal for the filter update.

To show the performance of echo cancellation without and with SR a simulation example is introduced in Fig. 4.

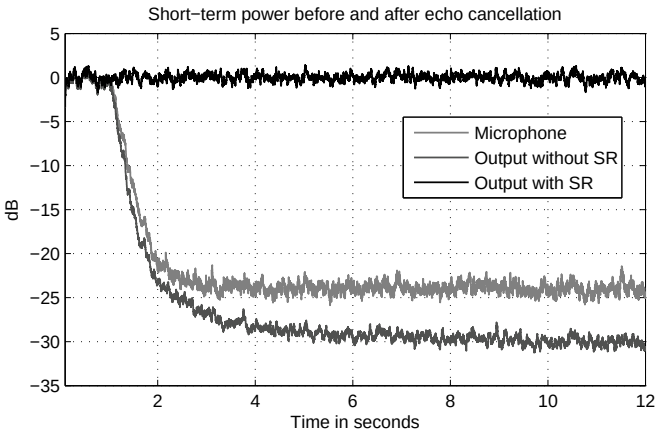

Fig. 4. *Performance of echo cancellation without and with additional SR (white noise signal as excitation)*

The black curve shows the short-term power of the excitation signal, whereas the second and third graph depict the outcome without and with SR (white noise is used as excitation signal, local speech and noise are not considered in this simulation). As setup for the evaluation a basic FFT of order $N = 256$, a subsampling rate of $r = 100$, a Hann-window, and echo cancellation filter of length $K = 8$ at a sampling rate of $f_s = 11025$ Hz are utilized. For the filter update the normalized least mean square (NLMS) algorithm with a step-size of $\beta = 0.3$ has been employed. Experiments have shown that the

echo reduction performance starts to decrease at $r \geq 90$ using a standard method because of aliasing effects. For the refinement only short FIR filter of order $M = 3$ are employed. Furthermore, only the original frequency supporting points have been refined, no computation of additional frequency supporting points in between were calculated. Meaning that the structure depicted in Fig. 3 has been employed except for the "cross" filters. Moreover, equivalent subband filters have been employed for each subband channel: ($\tilde{h}_k^{(\tilde{\mu})} = \tilde{h}_k$). From the simulation one can observe that using SR about 30 dB echo attenuation can be achieved which is appropriate for echo cancellation. Compared to a standard method without SR a faster initial convergence and an improved echo attenuation of approx. 6 dB (after the echo cancellation filter has converged) are achieved. It has to be noticed that only minor operations by means of multiplications and additions are added to a standard method and an insignificant additional delay is inserted in the signal path.

6. SUMMARY AND OUTLOOK

In this paper a generalized method for spectral refinement applied as a post-processing stage of an analysis filterbank for speech signals was presented. At a first stage a general solution how to individually refine subband signals was derived. Afterwards a computational efficient method for spectral refinement was proposed based on a linear combination of weighted subband signal vectors – the refinement procedure can easily be implemented using short FIR-filters in each subband channel. The SR method is particular suitable for speech processing systems with already integrated analysis filterbanks or DFT's – thus applying SR as a post-processing specific feature estimation such as pitch frequency or noise power can further be improved. The calculation of SR introduces an additional delay in the signal which has been kept low using short FIR filters for the refinement. In this contribution the SR method has been applied for fundamental frequency estimation as well as for echo cancellation. Evaluations demonstrated that pitch frequency estimation was improved considerably for all considered SNR levels. For pitch estimation only a refinement of the input signal at lower frequencies (up to 1 kHz) is needed and thus results in a very low computational complexity. The application of SR for echo cancellation have shown that the echo reduction especially for higher subsampled systems can be enhanced when utilizing SR.

7. REFERENCES

[1] A. de Cheveigne, H. Kawahara: *Yin, a Fundamental Frequency Estimator for Speech and Music*, J. Acoust. Soc. Am., vol. 111, no. 4, pp. 1917-1930, 2002.

[2] J. M. Cioffi, J. A. C. Bingham: *A Data-Driven Multitone Echo Canceller*, IEEE Globecom, pp. 2.4.1-2.4.5, 1991.

[3] R.E.Crochiere, L.R.Rabiner: *Multirate Digital Signal Processing*, USA: Prentice-Hall, 1983.

[4] Y. Ephraim, D. Malah: *Speech Enhancement Using a MMSE Short-Time Spectral Amplitude Estimator*, IEEE Trans. Acoust. Speech Signal Process., vol. 32, no. 6, pp 1109–1121, 1984.

[5] A. Gilloire, M. Vetterli: *Adaptive Filtering in Subbands with Critical Sampling*, IEEE Trans. Acoust. Speech Signal Process., vol. 40, no. 8, pp. 1862–1875, 1992.

[6] ETS 300 903 (GSM 03.50): *Transmission Planning Aspects of the Speech Service in the GSM Public Land Mobile Network (PLMS) System*, ETSI, France, 1999.

[7] N. Grbic, J.M. de Haan, I. Claesson, S. Nordholm: *Design of Oversampled Uniform DFT Filter Banks With Reduced Inband Aliasing and Reduced Inband Aliasing and Delay Const.*, ISSPA, vol. 1, pp 104–107, 2001.

[8] E. Hänsler, G. Schmidt: *Acoustic Echo and Noise Control – A Practical Approach*, John Wiley & Sons, Hoboken, NJ, USA, 2004.

[9] W. Kellermann: *Analysis and Design of Multirate Systems for Cancellation of Acoustic Echoes*, ICASSP, vol. 32, no. 2, pp. 2570–2573, 1988.

[10] M. Krini, G. Schmidt: *Model-based Speech Enhancement*, in E. Hänsler, G. Schmidt (eds.), Speech and Audio Processing in Adverse Environments, Berlin, Germany: Springer, pp. 89–134, 2008.

[11] M. Krini, G. Schmidt: *Spectral Refinement and its Application to Fundamental Frequency Estimation*, WASPAA, New York, USA, 2007.

[12] D. R. Morgan, J. C. Thi: *A Delayless Subband Adaptive Filter Architecture*, IEEE Trans. Acoust. Speech Signal Process., vol. 43, no. 8, pp. 1819–1830, 1995.

[13] P. A. Naylor, O. Tanrikulu, A. G. Constantinides: *Subband Adaptive Filtering for Acoustic Echo Control Using Allpass Polyphase IIR Filterbanks*, IEEE Trans. Acoust. Speech Signal Process., vol. 6, no. 2, pp. 143–155, 1998.

[14] M. R. Schroeder: *Period Histogram and Product Spectrum: New Methods for Fundamental Frequency Measurements*, J. Acoust. Soc. Am., vol. 43, no. 4, pp. 829–834, 1968.

[15] P. J. VanGerwen, F. A. Van de Laar, and J. Kotmans: *Digital Echo Canceller*, U.S. Patent 4,903,247, 1990.

[16] G. Wackersreuther: *On the Design of Filters for Ideal QMF and Polyphase Filter Banks*, AEU, vol. 39, no. 2, pp. 123–13, 1985.