Detection of Voiced Speech and Pitch Estimation for Applications with Low Spectral Resolution

Simon Graf[a,b], Nabeel Zaidi[a,b], Tobias Herbig[a], Markus Buck[a] and Gerhard Schmidt[b]
[a] Nuance Communications Deutschland GmbH, E-mail: simon.graf@nuance.com
[b] Christian-Albrechts-Universität zu Kiel, Germany

Abstract

Speech enhancement algorithms are employed in many applications, such as hands-free telephones, or speech recognizers, to recover a speech signal that is recorded in a noisy environment. In automotive environments, the noise particularly affects the low frequencies that are relevant for voiced speech. Detection of voiced speech sections and estimation of the pitch frequency help to reconstruct the harmonic structure of voiced speech and to enhance the speech signal. Many algorithms were introduced to detect voiced speech and to estimate the pitch. Most of them rely on a high spectral resolution that is achieved by employing long window lengths. However, some applications, such as in-car-communication (ICC) systems, have to deal with short windows in order to reduce computational costs and to ensure low system latencies. Resolving the pitch is difficult in this case. Spectral refinement techniques have been introduced to increase the spectral resolution by combining multiple consecutive low-resolution spectra. Using these techniques, standard pitch estimation algorithms can be applied even though the resolution of the original spectrum was too low. In this paper, we analyze the performance of pitch estimation using spectral refinement techniques and introduce an alternative approach that explicitly takes into account the short windows of ICC applications.

Introduction

Speech is an intuitive way for human communication that is employed in more and more applications. Devices, such as the car navigation system or smartphones, can be controlled conveniently via voice commands. Other applications facilitate the voice communication between humans, e.g., via hands-free telephone. In particular, in-car-communication systems amplify the driver's voice and support the communication with passengers on the backseat. By employing these systems, conversations are possible even in noisy conditions at higher velocities [1].

Voiced speech portions, e.g., vowels are important for correct recognition of human speech. However, the background noise in automotive environments masks especially these low-frequent components. The unvoiced speech portions in higher frequencies are masked less but are also less important for recognition. Therefore, robust detection of voiced speech and estimation of the pitch frequency are important problems in speech enhancement algorithms [2].

Detection of voiced speech can be used to distinguish speech from noise, e.g., for robust noise estimation. The pitch frequency can be employed to reconstruct speech that is masked by noise.

To capture the pitch information, long window lengths are required that exceed the pitch period. Some applications, however, need shorter windows in order to reduce the processing delay and the computational complexity. To overcome these contradicting requirements, techniques that approximate a long window by a combination of multiple shorter windows have been introduced in literature.

In this paper, two approaches will be discussed in more detail:

- *Spectral refinement* [3] combines multiple complex-valued spectra in order to recreate a spectrum with a higher frequency resolution.
- *Extended ACF* [3] combines multiple cross-correlations between short frames to approximate a longer auto-correlation function (ACF).

Both techniques gain information from some previous frames in addition to the current frame. By employing this temporal context, pitch information can be extracted even for very short windows.

In this contribution, the detection of harmonic components, as well as pitch estimation will be summarized. A conventional approach based on the auto-correlation function is employed. Afterwards, we will consider shorter windows and discuss the two approaches to deal with this challenge. We will briefly summarize spectral refinement and provide a more detailed description of the extended ACF.

Our analyses focus on the comparison of the different approaches. In particular, the detection performance of voiced speech and the estimated pitch are assessed.

Pitch Estimation using ACF

First, we describe the basic principle of ACF-based pitch estimation. Based on a frame of an audio signal

$$\tilde{\mathbf{x}}(\ell) = [x(\ell R - \tilde{N} + 1), \cdots, x(\ell R - N + 1), \cdots, x(\ell R)]^T, \quad (1)$$

the ACF is determined. Here, the number of samples $\tilde{N}$ that are taken into account is chosen much longer than the expected pitch periods. The shift between two succeeding frames is denoted by R and the frame index by

ℓ. Later in this paper, shorter frames of length N will be considered that are too short to resolve the pitch.

The frames in time-domain are converted into the spectral-domain

$$\tilde{\mathbf{X}}(\ell) = \tilde{\mathbf{D}} \cdot \left(\tilde{\mathbf{h}} \circ \tilde{\mathbf{x}}(\ell)\right) \tag{2}$$

by applying a window $\tilde{\mathbf{h}}$ followed by a discrete Fourier-transform (DFT) $\tilde{\mathbf{D}}$. The windowing is based on an element-wise multiplication "$\circ$" of the two vectors.

In order to determine the pitch period, the power spectral density is estimated and transformed back to the time-domain to get the auto-correlation function

$$\mathbf{r}_{\tilde{x}\tilde{x}}(\ell) = \tilde{\mathbf{P}} \cdot \tilde{\mathbf{D}}^{-1} \cdot \left(\tilde{\mathbf{X}}^*(\ell) \circ \tilde{\mathbf{X}}(\ell)\right) \tag{3}$$

$$= \left[r_{-\tilde{N}/2+1}(\ell), \cdots, r_0(\ell), \cdots, r_{\tilde{N}/2}(\ell)\right]^T \tag{4}$$

where a permutation matrix $\tilde{\mathbf{P}}$ is employed to ensure that the zeroth element is placed in the middle of the vector.

The position of the maximum of the ACF

$$\hat{\tau}_{\text{pitch}}(\ell) = \underset{\tau \in \{\tau_{\text{low}}, \cdots, \tau_{\text{high}}\}}{\arg\max} \{r_\tau(\ell)\} = \frac{f_\text{s}}{\hat{f}_{\text{pitch}}(\ell)} \tag{5}$$

is interpreted as the estimated pitch period. It is limited to the range of human pitch periods $\{\tau_{\text{low}}, \cdots, \tau_{\text{high}}\}$. The presence of pitch can be detected by comparing the normalized ACF maximum value

$$\hat{p}_{\text{pitch}}(\ell) = r_{\hat{\tau}_{\text{pitch}}(\ell)}(\ell)/r_0(\ell) \tag{6}$$

to a threshold.

Shorter Windows and Combination

For some applications, shorter windows have to be employed

$$\mathbf{x}(\ell) = [x(\ell R - N + 1), \cdots, x(\ell R)]^T \tag{7}$$

where the window length N is too short to capture the long pitch period τ_{high}.

To achieve a long window of length $\tilde{N}$, $M = \frac{\tilde{N}-N}{R} + 1$ consecutive frames have to be combined. In the following sections, two different strategies to exploit the temporal context are described.

Spectral refinement directly combines multiple low-resolution spectra

$$\mathbf{X}(\ell) = \mathbf{D} \cdot (\mathbf{h} \circ \mathbf{x}(\ell)) \tag{8}$$

to approximate the high resolution spectrum $\tilde{\mathbf{X}}(\ell)$ whereas the extended ACF approach approximates the long ACF $\mathbf{r}_{\tilde{x}\tilde{x}}(\ell)$ by means of multiple shorter correlations.

Spectral Refinement

For spectral refinement, multiple low-resolution spectra $\mathbf{X}(\ell)$ are combined to approximate the high-resolution spectrum $\tilde{\mathbf{X}}(\ell)$. For this, a spectral refinement matrix $\mathbf{S} \in \mathbb{C}^{\tilde{N} \times MN}$ is found that maps the stacked low-resolution spectra to a longer vector

$$\hat{\tilde{\mathbf{X}}}(\ell) = \mathbf{S} \cdot \left[\mathbf{X}^T(\ell), \mathbf{X}^T(\ell-1), \cdots, \mathbf{X}^T(\ell-(M-1))\right]^T \tag{9}$$

of the approximated high-resolution spectrum. The spectral refinement matrix

$$\mathbf{S} = \tilde{\mathbf{D}} \cdot \mathbf{A} \cdot \mathbf{D}_{\text{Block}}^{-1} \tag{10}$$

comprises a transformation $\mathbf{D}_{\text{Block}}^{-1}$ of the stacked low-resolution spectra back into the time-domain, a combination $\mathbf{A}$ of multiple time-domain signals to a longer time-domain signal, and a transformation $\tilde{\mathbf{D}}$ of the long signal back into the frequency domain. Due to the sparseness of the $\mathbf{S}$-matrix, the refinement can be implemented very efficiently as described in [3].

Afterwards, the ACF can be calculated using (3) based on the approximated high-resolution spectrum.

Extended ACF

Now, we approximate the long auto-correlation $\mathbf{r}_{\tilde{x}\tilde{x}}(\ell)$ by a combination of shorter cross-correlations (CCF)

$$\mathbf{c}_{xx}(\ell, \Delta\ell) = \mathbf{P} \cdot \mathbf{D}^{-1} \cdot (\mathbf{X}^*(\ell) \circ \mathbf{X}(\ell - \Delta\ell)) \tag{11}$$

$$= \left[c_{-N/2+1}(\ell, \Delta\ell), \cdots, c_0(\ell, \Delta\ell), \cdots, c_{N/2}(\ell, \Delta\ell)\right]^T. \tag{12}$$

In contrast to spectral refinement, the element-wise multiplication in (11) is a non-linear operation that cannot perfectly be reverted using a linear matrix multiplication. However, we know which elements of the CCFs are relevant for the ACF and can compensate the envelope caused by the window functions.

For this, we calculate a weighted sum of normalized CCFs

$$\tilde{r}_\tau(\ell) = \beta_\tau \cdot \sum_{\tilde{\ell}=0}^{(M-1)/2} \frac{c_{\tau-\tilde{\ell}R}(\ell, \tilde{\ell})}{\sqrt{c_0(\ell, 0) \cdot c_0(\ell - \tilde{\ell}, 0)}} \cdot \alpha_{\tau-\tilde{\ell}R} \tag{13}$$

where the weighting coefficients α are chosen in a way that the envelope after the summation is flat. The coefficients β then recreate the desired envelope of the long ACF as illustrated in Figure 1.

To determine the envelopes, we consider a constant excitation $x(n) = 1$. Then, we get a short envelope

$$\mathbf{e} = \mathbf{P} \cdot \mathbf{D}^{-1} \cdot (\mathbf{D}^* \mathbf{h}^* \circ \mathbf{D}\mathbf{h}) \tag{14}$$

$$= \left[e_{-N/2+1}, \cdots, e_0, \cdots, e_{N/2}\right]^T \tag{15}$$

based on the short window $\mathbf{h}$ and analogously a long envelope $\tilde{\mathbf{e}}$ based on the long window $\tilde{\mathbf{h}}$.

The coefficients β_τ directly correspond to the desired long envelope $\tilde{e}_\tau$.

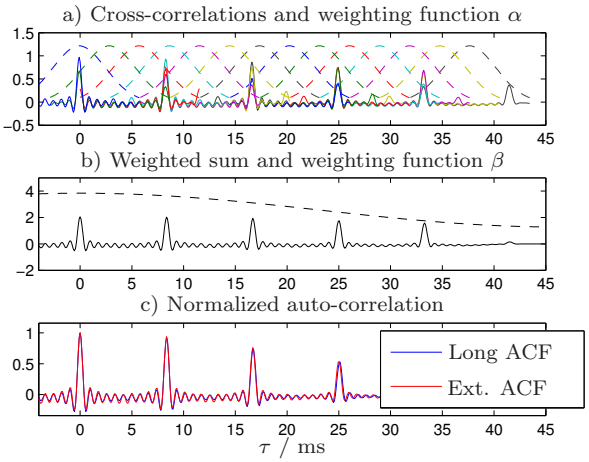

Figure 1: Example of the weighting functions for extended ACF: a) The different cross-correlations (11) (solid lines) are weighted with α (16) (dashed lines) and summed up. b) This weighted sum (black solid line) with a flat envelope is then reshaped by β (black dashed line) to approximate the ACF. c) The extended ACF (13) (red solid line) approximates the long ACF (3) (blue solid line) well.

In contrast, the weighting coefficients α are chosen such that

$$\mathbf{H} \cdot \boldsymbol{\alpha} = \mathbf{H} \cdot \left[\alpha_{-N/2+1}, \cdots, \alpha_0, \cdots, \alpha_{N/2}\right]^T = \mathbf{1}_{R\times 1} \tag{16}$$

to provide perfect reconstruction of a flat envelope. For this, a matrix

$$\mathbf{H} = \left[\mathbf{d}_{-N/2+1}, \mathbf{d}_{-N/2+R+1}, \cdots, \mathbf{d}_{N/2-R+1}\right] \tag{17}$$

is defined that is composed of multiple diagonal matrices

$$\mathbf{d_i} = \begin{bmatrix} e_i & 0 & 0 & 0 & 0 \\ 0 & e_{i+1} & 0 & 0 & 0 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & 0 & e_{i+R-2} & 0 \\ 0 & 0 & 0 & 0 & e_{i+R-1} \end{bmatrix} \tag{18}$$

containing the values of the short envelope.

To solve (16) for $\boldsymbol{\alpha}$, the pseudo-inverse $\mathbf{H}^+$ of $\mathbf{H}$ is employed. Additional constraints guarantee a symmetric weighting coefficient vector and a continuous shape.

Using this technique, only the CCFs between the current frame and some previous frames are taken into account. To capture also the information from CCFs between previous frames, temporal smoothing

$$\hat{r}_\tau(\ell) = \frac{1}{\tilde{L}} \sum_{\tilde{\ell}=0}^{\tilde{L}-1} \tilde{r}_\tau(\ell - \tilde{\ell}) \tag{19}$$

can be applied. Choosing $\tilde{L} = M/2$, almost the same context is considered as for the ACF of a long window. Alternatively, the smoothing can be realized with a recursive filter to save memory and computational costs.

Further simplifications can be achieved by calculating only the relevant CCFs that cover the range of human pitch periods.

Experiments

For our analyses, we consider a configuration that is typical for real-time applications with critical latency requirements, such as ICC applications. For a sampling rate $f_s = 16\,\text{kHz}$, short Hann windows of 128 samples with an overlap of 75% are chosen. Using this configuration, a single frame is not sufficient to resolve the pitch. We therefore target on extending the search range for the pitch period by considering some previous frames. Both techniques, spectral refinement and extended ACF, are applied in order to achieve an effective window length of 1024 samples.

First, the performance is illustrated for an artificial signal. A harmonic signal is sweeped in the typical range of the human pitch frequencies between 300 Hz and 60 Hz. For this signal, the ACF is estimated by means of the different approaches.

The estimated ACFs for a long and a short window, as well as the approximations using spectral refinement and extended ACF are shown in Figure 2. As expected, the short window does not capture the relevant frequency range of human pitch periods. In contrast, the long ACF and both approximations cover the full range.

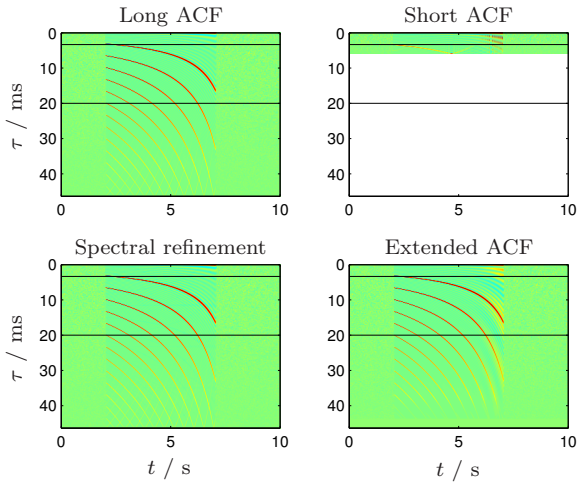

Figure 2: Example comparison of different methods for ACF estimation for a harmonic frequency sweep: ACF based on a long window of 64 ms and a short window of 8 ms as well as estimated ACFs with an effective length of 64 ms using spectral refinement and extended ACF. The typical range of human pitch periods is indicated by black lines.

To get an impression of the performance for the detection of harmonic components and pitch estimation, both features are determined for the four variants as depicted in Figure 3. Again, it is obvious that the short window does not reasonably capture the pitch: the voicing feature does not follow the correct shape. All other approaches, however, provide the same results for the voicing feature and the pitch estimate. For this artificial example therefore all approaches with long effective windows are applicable.

A second experiment targets on the detection performance in a realistic noise scenario. Speech data from the TIMIT database [5] was mixed with automotive noise

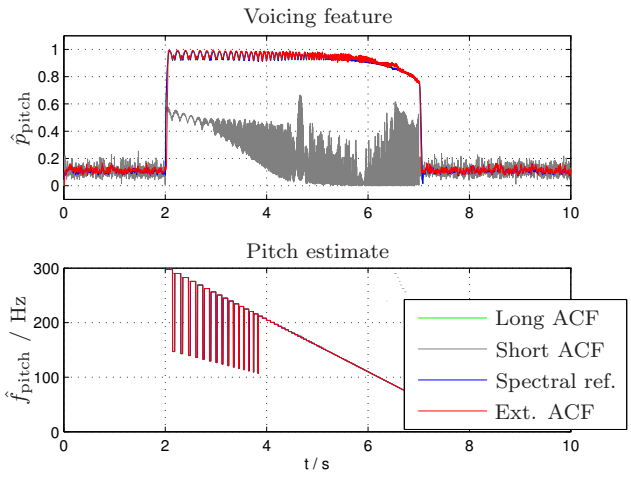

Figure 3: Voicing feature and pitch estimate of the harmonic frequency sweep. Using a short window, the pitch is not reasonably captured whereas all other approaches provide almost the same results.

taken from UTD-CAR-NOISE [6]. A variety of noises and SNRs was taken into account to investigate realistic conditions.

The receiver operation characteristic (ROC) curve in Figure 4 illustrates the results. The curve for a short window is close to the diagonal which indicates again an insufficient detection performance. All other approaches show the same performance which underlines that spectral refinement and extended ACF both are capable to increase the effective window length.

Comparing the computational costs of the approximations, both approaches appear to be on a similar level. Spectral refinement requires $M \cdot N/2 + M \cdot N$ operations [4] in addition to a long IFFT of order $\tilde{N}\,\mathrm{ld}(\tilde{N})$. In contrast, $M/2$ shorter IFFTs of order $N\,\mathrm{ld}(N)$ have to be calculated for the extended ACF.

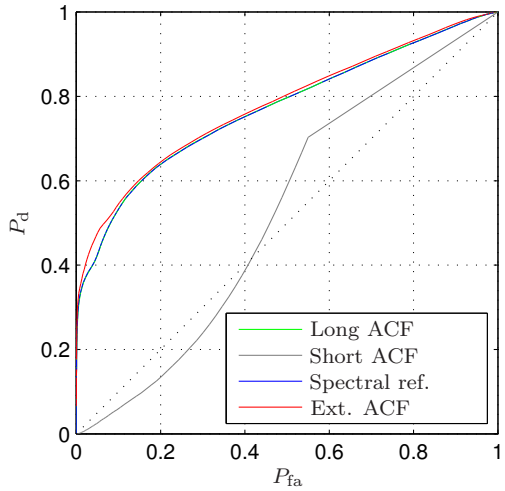

Figure 4: ROC curve: detection of voiced speech in automotive noise. The performance of all approaches with a long effective window is almost the same.

Conclusions

In this paper, two approaches to extend the effective window length for detection of voiced speech and pitch estimation have been summarized and discussed. Spectral refinement targets on extending the resolution of a spectrum by incorporating information from the past. In contrast, extended ACF considers the temporal context by combining multiple short cross-correlations between current and previous frames. Our analyses confirmed that both approaches for combining short windows are capable to approximate an ACF for a longer window. Almost the same detection and estimation performance was achieved for all the approaches with a long effective window.

References

[1] G. Schmidt, T. Haulick *Signal processing for in-car communication systems*, Signal processing, vol. 86, no. 6, pp. 1307–1326, 2006.

[2] A. de Cheveigné, H. Kawahara *YIN, a fundamental frequency estimator for speech and music*, The Journal of the Acoustical Society of America, vol. 111, no. 4, p. 1917, 2002.

[3] M. Krini, G. Schmidt *Spectral refinement and its application to fundamental frequency estimation*, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, USA, 2007.

[4] M. Krini, G. Schmidt *Refinement and Temporal Interpolation of Short-Term Spectra: Theory and Applications*, in Smart Mobile In-Vehicle Systems: Next Generation Advancements, G. Schmidt, H. Abut, K. Takeda, and J. H. L. Hansen, Eds. Springer New York, 2014, pp. 139–166.

[5] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, N. L. Dahlgren *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM* National Institute of Standards and Technology, 1993.

[6] N. Krishnamurthy, J. H. L. Hansen *Car noise verification and applications* International Journal of Speech Technology, 2013.