INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Low-Complexity Pitch Estimation Based on Phase Differences Between Low-Resolution Spectra

Simon Graf[1,2], *Tobias Herbig*[1], *Markus Buck*[1], *Gerhard Schmidt*[2]

[1]Acoustic Speech Enhancement Research, Nuance Communications Deutschland GmbH, Ulm, Germany
[2]Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel, Kiel, Germany

simon.graf@nuance.com

Abstract

Detection of voiced speech and estimation of the pitch frequency are important tasks for many speech processing algorithms. Pitch information can be used, e.g., to reconstruct voiced speech corrupted by noise.

In automotive environments, driving noise especially affects voiced speech portions in the lower frequencies. Pitch estimation is therefore important, e.g., for in-car-communication systems. Such systems amplify the driver's voice and allow for convenient conversations with backseat passengers. Low latency is required for this application, which requires the use of short window lengths and short frame shifts between consecutive frames. Conventional pitch estimation techniques, however, rely on long windows that exceed the pitch period of human speech. In particular, male speakers' low pitch frequencies are difficult to resolve.

In this publication, we introduce a technique that approaches pitch estimation from a different perspective. The pitch information is extracted based on phase differences between multiple low-resolution spectra instead of a single long window. The technique benefits from the high temporal resolution provided by the short frame shift and is capable to deal with the low spectral resolution caused by short window lengths. Using the new approach, even very low pitch frequencies can be estimated very efficiently.

Index Terms: speech detection, voiced speech

1. Introduction

Speech enhancement techniques are employed in many speech-driven applications. Based on a speech signal that is corrupted with noise, these techniques try to recover the original speech. In many scenarios, such as automotive applications, the noise is concentrated at the lower frequencies. Speech portions in this frequency region are particularly affected by the noise.

Human speech comprises voiced as well as unvoiced phonemes. Voiced phonemes exhibit a harmonic excitation structure caused by periodic vibrations of the vocal folds. In time domain, this voiced excitation is characterized by a sequence of repetitive impulse-like signal components. Valuable information is contained in the pitch frequency, such as information on the speaker's identity or the prosody. It is therefore desirable for many applications to detect the presence of voiced speech and to estimate the pitch frequency [1, 2, 3, 4].

Typically, long window lengths are required to resolve the pitch frequency accurately. Multiple excitation impulses have to be captured to extract the pitch information. This is a problem especially for low male voices with pitch periods that may exceed the typical window lengths used in practical applications [5]. Increasing the window length is mostly not acceptable since it also increases the system latency as well as the computational complexity.

Beyond that, the constraints regarding system latency and computational costs are very challenging for some applications. For in-car-communication (ICC) systems, the system latency has to be kept as low as possible in order to ensure a convenient listening experience. Since the original speech and the amplified signal overlay in cabin, delays longer than 10 ms between both signals are perceived as annoying by the listeners [6]. Very short windows have to be employed which obviates the application of standard approaches for pitch estimation.

In this paper, we therefore introduce a pitch estimation technique that is capable to deal with very short windows. In contrast to usual approaches, the pitch information is not extracted based on a single long frame. Instead, our approach considers the phase relation between multiple shorter frames. Using this technique, even very low pitch frequencies can be resolved. Since the approach completely operates in the frequency domain, a low computational complexity is achieved.

2. Algorithm

Typical pitch estimation techniques search for periodic components in a long frame. Using, e.g., the auto-correlation function (ACF), repetitive structures in a long frame can be detected. The pitch period is then estimated by finding the position of a maximum of the ACF.

In contrast, our approach detects repetitive structures by comparing pairs of non-overlapping short frames. We assume that two excitation impulses are captured by two different short frames. Further assuming that both impulses are equally shaped, the signal sections in both frames are equal except for a temporal shift. By determining this shift, the pitch period can be estimated very efficiently.

2.1. Signal model

We formulate two hypotheses (H_0 and H_1) for presence and absence of voiced speech. For presence of voiced speech, the signal is expressed by a superposition

$$\mathsf{H}_0 : x(n) = s_{\mathrm{v}}(n, \tau_{\mathrm{v}}(n)) + b(n) \tag{1}$$

of voiced speech components s_{v} and other components b comprising unvoiced speech and noise. Alternatively, when voiced speech is absent, the signal

$$\mathsf{H}_1 : x(n) = b(n) \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2017-1254

purely depends on noise or unvoiced speech components.

Our goal is to detect the presence of voiced speech components. In case that voiced speech is detected, we further want to estimate the pitch frequency $f_\mathrm{v} = f_\mathrm{s}/\tau_\mathrm{v}$ where f_s denotes the sampling rate and τ_v the pitch period in samples.

Voiced speech is modeled by a periodic excitation

$$s_\mathrm{v}(n, \tau_\mathrm{v}(n)) = g_n(n) + g_n(n + \tau_\mathrm{v}(n)) + g_n(n + 2\tau_\mathrm{v}(n)) + \cdots \tag{3}$$

where the shape of a single excitation impulse is expressed by a function g_n. The distance τ_v between two succeeding peaks corresponds to the pitch period. For human speech, the pitch periods may assume values up to $\tau_\mathrm{max} = f_\mathrm{s}/50\,\mathrm{Hz}$ for very low male voices.

2.2. Pitch estimation using auto- and cross-correlation

Signal processing is performed on frames of the signal

$$\mathbf{x}(\ell) = [x(\ell R - N + 1), \cdots, x(\ell R - 1), x(\ell R)]^T \tag{4}$$

where N denotes the window length.

For long windows $N > \tau_\mathrm{max}$, the maximum of the ACF

$$\mathrm{acf}_{xx}(\tau, \ell) = \frac{1}{N} \sum_{k=0}^{N-1} |X(k, \ell)|^2 \cdot e^{2\pi j k \tau / N} \tag{5}$$

in the range of human pitch periods can be used to estimate the pitch as illustrated in Figure 1(a)-(c). An IDFT is applied here to transform the estimated high-resolution power spectrum $|X(k, \ell)|^2$ to the ACF.

In this paper, however, we focus on very short windows $N \ll \tau_\mathrm{max}$ that are too short to capture a full pitch period. The spectral resolution of $X(k, \ell)$ is low due to the short window length. However, for short frame shifts $R \ll \tau_\mathrm{max}$, a good temporal resolution is achieved. In this case, two short frames $\mathbf{x}(\ell)$ and $\mathbf{x}(\ell - \Delta\ell)$ can be exploited to determine the pitch as illustrated in Figure 1(d). When both frames contain different excitation impulses, the cross-correlation between the frames

$$\mathrm{cc}_{xx}(\tilde{\tau}, \ell, \Delta\ell) = \frac{1}{N} \sum_{k=0}^{N-1} X^*(k, \ell) \cdot X(k, \ell - \Delta\ell) \cdot e^{2\pi j k \tilde{\tau} / N} \tag{6}$$

has a maximum $\tilde{\tau}_\mathrm{v}$ that corresponds to the pitch period $\hat{\tau}_\mathrm{v} = \tilde{\tau}_\mathrm{v} + \Delta\ell \cdot R$. To emphasize the peak of the correlation, the generalized cross-correlation (GCC)

$$\mathrm{gcc}_{xx}(\tilde{\tau}, \ell, \Delta\ell) = \frac{1}{N} \sum_{k=0}^{N-1} \underbrace{\frac{X^*(k, \ell) \cdot X(k, \ell - \Delta\ell)}{|X^*(k, \ell) \cdot X(k, \ell - \Delta\ell)|}}_{\mathrm{GCS}_{xx}(k, \ell, \Delta\ell)} \cdot e^{2\pi j k \tilde{\tau} / N} \tag{7}$$

can be employed instead. By removing the magnitude information in the normalized cross-spectrum GCS_{xx}, the GCC purely relies on the phase. As a consequence, the distance between the two impulses can be clearly identified as shown in Figure 1(f).

In the following, we will introduce a method to estimate the pitch period directly in the frequency domain. The estimation is based on the slope of the phase of GCS_{xx} that is illustrated in Figure 1(e).

2.3. Pitch estimation based on phase differences

When two short frames capture temporally shifted impulses of the same shape, the shift can be expressed by a delay. In frequency domain, this is characterized by a linear phase of the cross-spectrum. In this case, the phase relation between neighboring frequency bins

$$\Delta\mathrm{GCS}(k, \ell, \Delta\ell) = \mathrm{GCS}_{xx}(k, \ell, \Delta\ell) \cdot \mathrm{GCS}^*_{xx}(k - 1, \ell, \Delta\ell) \tag{8}$$

$$= e^{j\Delta\varphi(k, \ell, \Delta\ell)} \tag{9}$$

is constant for all frequencies with a phase difference $\Delta\varphi(\ell, \Delta\ell) = \Delta\varphi(1, \ell, \Delta\ell) = \Delta\varphi(2, \ell, \Delta\ell) = \cdots$. For signals that don't exhibit a periodic structure, $\Delta\varphi(k, \ell, \Delta\ell)$ has a rather random nature over k. Testing for linear phase therefore can be employed to detect voiced components.

A weighted sum along frequency

$$\overline{\Delta\mathrm{GCS}}(\ell, \Delta\ell) = \frac{\sum_{k=1}^{K-1} w(k, \ell, \Delta\ell) \cdot \Delta\mathrm{GCS}(k, \ell, \Delta\ell)}{\sum_{k=1}^{K-1} w(k, \ell, \Delta\ell)} \tag{10}$$

can be employed to detect speech and estimate the pitch frequency. For harmonic signals, the magnitude of the weighted sum yields values close to 1 due to the linear phase. Otherwise smaller values result. The weighting coefficients $w(k, \ell, \Delta\ell)$ are used to emphasize frequencies that are relevant for speech. The coefficients can either be set to fixed values or can be chosen dynamically, e.g., using an estimated signal-to-noise power ratio (SNR). We set them to

$$w(k, \ell, \Delta\ell) = \begin{cases} |X(k, \ell)| & \text{for } 50\,\mathrm{Hz} < k f_\mathrm{s}/N < 4\,\mathrm{kHz} \\ 0 & \text{else} \end{cases} \tag{11}$$

in order to emphasize dominant components in the spectrum in the frequency range of voiced speech.

The weighted sum in (10) relies only on the phase difference between the most current frame ℓ and one previous frame $\ell - \Delta\ell$. To include more than two excitation impulses for the estimate, it is beneficial to apply temporal smoothing

$$\overline{\overline{\Delta\mathrm{GCS}}}(\ell, \Delta\ell) = \alpha \cdot \overline{\overline{\Delta\mathrm{GCS}}}(\ell - \Delta\ell, \Delta\ell) + (1 - \alpha) \cdot \overline{\Delta\mathrm{GCS}}(\ell, \Delta\ell). \tag{12}$$

The temporal context that is employed can be adjusted by changing the smoothing constant α. For smoothing, only frames are considered that probably contain the previous impulse: when we search for impulses with a distance of $\Delta\ell$ frames, we take the smoothed estimate at $\ell - \Delta\ell$ into account.

Based on the averaged phase differences, we define a voicing feature

$$p_\mathrm{v}(\ell, \Delta\ell) = \left| \overline{\overline{\Delta\mathrm{GCS}}}(\ell, \Delta\ell) \right| \tag{13}$$

that represents the linearity of the phase. When all complex values $\Delta\mathrm{GCS}$ have the same phase, they accumulate and result in a mean value of magnitude one indicating linear phase. Otherwise, the phase is randomly distributed and the result assumes lower values.

In a similar way, we can estimate the pitch period. Replacing the magnitude in (13) by an angle operator

$$\widehat{\Delta\varphi}(\ell, \Delta\ell) = \angle \overline{\overline{\Delta\mathrm{GCS}}}(\ell, \Delta\ell) \tag{14}$$

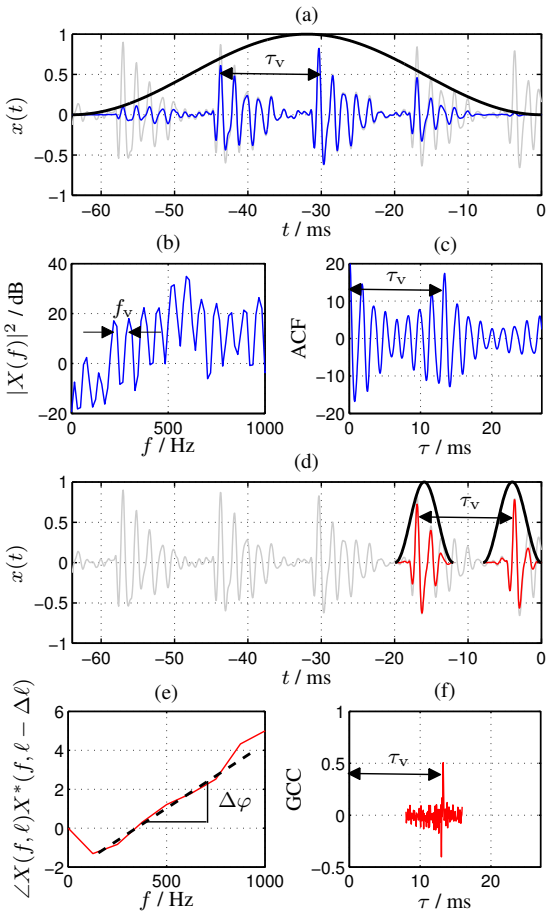

Figure 1: *Pitch estimation using long and short windows: (a) A long window captures multiple excitation impulses. (b) The power spectral density reflects the pitch frequency f_v using only magnitude information. (c) The pitch period τ_v can be determined by means of the auto-correlation function's (ACF) maximum. (d) For shorter windows, two frames are needed to capture the pitch period. (e) The phase differences between the two low-resolution spectra contain all relevant information. (f) The generalized cross-correlation (GCC) between the frames shows the peak more distinctly compared to the ACF in (c). Our approach directly estimates the pitch based on the slope $\Delta\varphi$.*

we get an estimate of the slope of the linear phase. This slope can be converted to an estimate of the pitch period

$$\hat{\tau}_v(\ell, \Delta\ell) = \frac{\widehat{\Delta\varphi}(\ell, \Delta\ell)}{2\pi} N + \Delta\ell \cdot R. \tag{15}$$

In contrast to conventional approaches, the pitch is directly estimated in the frequency domain based on the phase differences. The approach can be implemented very efficiently since there is no need for a transformation back into the time domain and a maximum search that are typically required by ACF-based methods.

2.4. Post-processing and detection

In a post-processing, the results of different short frames are combined to achieve a final voicing feature and a pitch estimate. Since a moving section of the audio signal is captured by the short frames, the most current frame can contain one excitation impulse, however, it might also lie between two impulses. In this case, no voiced speech would be detected in the current frame even though a distinct harmonic excitation is present in the signal. To prevent from these gaps, maximum values of $p_v(\ell, \Delta\ell)$ are held over $\Delta\ell$ frames.

Using Eq. (13), multiple results for different pitch regions are considered. For each phase difference between the current frame ℓ and one previous frame $\ell - \Delta\ell$, a value of the voicing feature $p_v(\ell, \Delta\ell)$ is determined. The different values are fused to a final feature by searching for the most probable region

$$\widehat{\Delta\ell}(\ell) = \underset{\Delta\ell}{\operatorname{argmax}} \left(p_v(\ell, \Delta\ell)\right) \tag{16}$$

that contains the pitch period. Then, the voicing feature and pitch estimate are given by $p_v(\ell) = p_v(\ell, \widehat{\Delta\ell}(\ell))$ and $\hat{f}_v(\ell) = \hat{f}_v(\ell, \widehat{\Delta\ell}(\ell))$ respectively.

Based on the voicing feature p_v we want to take a decision on the presence of voiced speech. To decide for one of the two hypotheses H_0 and H_1 in (1) and (2), a threshold η is applied to the voicing feature. When the feature exceeds the threshold, voiced speech is detected, otherwise absence of voiced speech is supposed.

3. Experiments and Results

For our evaluations, we focus on an automotive noise scenario that is typical for ICC applications. We employ speech signals from the Keele speech database [7] and automotive noise from the UTD-CAR-NOISE database [8]. The signals are downsampled to a sampling rate of $f_s = 16$ kHz. A frameshift of $R = 32$ samples (2 ms) is used for all our analysis. For the short frames, a Hann window of 128 samples (8 ms) is employed.

A pitch reference based on laryngograph recordings is provided with the Keele database. We use this reference as a ground truth for all our analyses.

For comparison, we employ a conventional pitch estimation approach based on ACF. This algorithm is applied to the noisy data to get a baseline to assess the performance of the new approach. Since a long temporal context is considered by the long window of 1024 samples (64 ms), a good performance can be achieved.

First, we illustrate the results for one example. Speech and noise were mixed to an SNR of 0 dB. The detection result and the pitch estimate of both algorithms as well as the reference are depicted in Figure 2. Both algorithms indicate voiced speech by high values of the voicing feature p_v close to one. A threshold can be applied as a simple detector. We set the threshold to $\eta = 0.25$ for the conventional approach and to $\eta = 0.5$ for our new approach and estimate the pitch only when the voicing features exceed the threshold. The resulting pitch estimates for the new algorithm show that it is capable to track the pitch. However, the results are not as precise as the results from the baseline method.

To evaluate the performance for a more extensive database, we mixed the ten utterances (duration 337 s) from the Keele database spoken by male and female speakers with automotive noise and adjusted the SNR. The receiver operating characteristic (ROC) was determined for each SNR value by tuning the threshold η between 0 and 1. The rate of correct detections was found by comparing the detections for a certain threshold to the reference of voiced speech. On the other hand, the false-alarm rate was calculated for intervals where the reference indicated absence of speech. By calculating the area under ROC curve

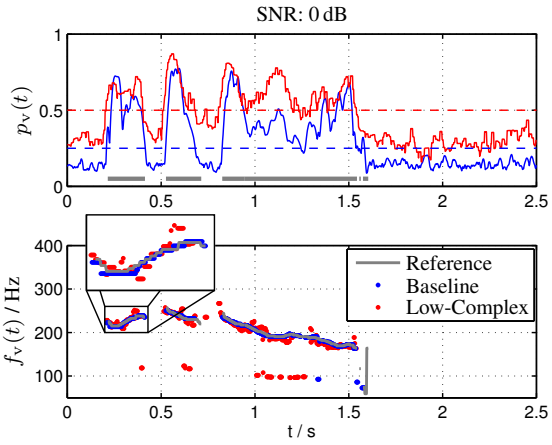

Figure 2: *Example for voicing feature p_v and pitch estimate f_v for a noisy speech signal (SNR=0 dB): the low-complexity feature indicates speech similar to the ACF-based baseline algorithm. Both approaches are capable to estimate the pitch frequency, however, the variance of the low-complexity feature is higher. Some sub-harmonics are observable for both approaches and even for the reference.*

(AUC), the performance curve was compressed to a scalar measure. AUC values close to one indicate a good detection performance whereas values close to 0.5 correspond to random results.

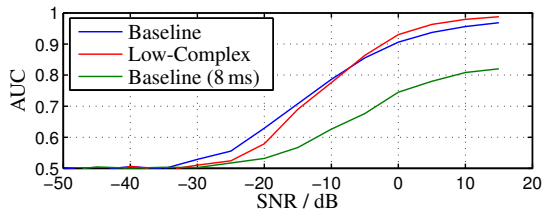

Figure 3: *Performance of voiced speech detection over SNR: the low-complexity feature shows a good detection performance that is similar to the performance of the baseline algorithm with a long context. When applying the baseline algorithm to a shorter window, even for high SNRs the performance is low since low pitch frequencies cannot be resolved.*

The detection performances of the baseline and the new algorithm are shown in Figure 3. As discussed, the baseline approach shows a good detection performance since it captures a long temporal context. Even though the low-complexity approach has to deal with less temporal context, a similar detection performance is achieved. When applying the baseline approach to a short window, even for high SNRs voiced speech is not perfectly detected. Low pitch frequencies cannot be resolved using a single short window which explains the low performance.

In a second analysis, we focus on the pitch estimation performance of both approaches. For this, we consider time instances where both the reference and the algorithm under test indicate presence of voiced speech. The deviation between the estimated and the reference pitch frequency is assessed.

For 0 dB, we observed a good detection performance for both algorithms. We therefore first investigate the estimation

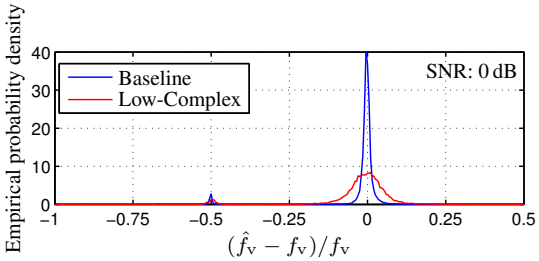

Figure 4: *Distribution of errors of pitch frequency estimates: the majority of errors are in an interval of ± 10% of the reference pitch frequency. Some sub-harmonics are falsely identified as the pitch, reflected by the peak at -0.5.*

performance for this situation. In Figure 4, a histogram of the deviations $\hat{f}_v - f_v$ relative to the reference frequency f_v is depicted. It is observable that the pitch frequency is mostly estimated correctly. However, small deviations in an interval of $\pm 10\%$ of the reference pitch frequency can be noticed for both approaches. The smaller peak at -0.5 can be explained by sub-harmonics that were accidentally selected. By applying a more advanced post-processing instead of the simple maximum search in (16), this type of errors could be reduced.

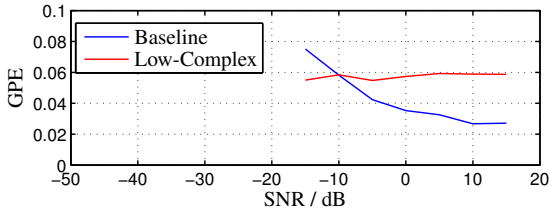

Figure 5: *Gross pitch error: empirical probability of pitch estimation errors with deviations that exceed 20% of the reference pitch frequency. The baseline approach estimates the pitch frequency more accurate than the new low-complexity algorithm.*

Deviations from the reference pitch frequency can be evaluated using the gross pitch error (GPE) [9]. For this, we determine the empirical probability of deviations that are greater than 20% of the reference pitch: $P(|\hat{f}_v - f_v| > 0.2 \cdot f_v)$.

In Figure 5, the GPE is depicted for SNRs where a reasonable detection performance was achieved. For high SNRs, we again observe higher deviations of the new algorithm compared to the conventional approach. Many of these errors can be explained with sub-harmonics that are falsely identified as the pitch frequency.

4. Conclusions

In this contribution, a low-complexity algorithm for detection of voiced speech and pitch estimation was introduced. It is capable to deal with special constraints given by applications where low latency is required, such as ICC systems. In contrast to conventional pitch estimation approaches, the algorithm employs very short frames that capture only a single excitation impulse. The distance between multiple impulses, corresponding to the pitch period, is determined by evaluating phase differences between the low-resolution spectra. Since no IDFT is needed to estimate the pitch, the computational complexity is low compared to standard pitch estimation techniques.

5. References

[1] A. de Cheveigné and 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.

[2] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *Proc. of EUSIPCO*, Barcelona, Spain, 2011.

[3] B. S. Lee and D. P. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in *Proc. of Interspeech*, Portland, Oregon, USA, 2012.

[4] F. Kurth, A. Cornaggia-Urrigshardt, and S. Urrigshardt, "Robust F0 Estimation in Noisy Speech Signals Using Shift Autocorrelation," in *Proc. of ICASSP*, Florence, Italy, 2014.

[5] M. Krini and G. Schmidt, "Spectral refinement and its application to fundamental frequency estimation," in *Proc. of WASPAA*, New Paltz, New York, USA, 2007.

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

[7] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Proc. of EUROSPEECH*, Madrid, Spain, 1995.

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

[9] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.