ROBUST F0 ESTIMATION IN NOISY SPEECH SIGNALS USING SHIFT AUTOCORRELATION

Frank Kurth, Alessia Cornaggia-Urrigshardt and Sebastian Urrigshardt

Fraunhofer FKIE, Communication Systems
Fraunhoferstr. 20, 53343 Wachtberg, Germany
frank.kurth@fkie.fraunhofer.de

ABSTRACT

We present a novel method for robustly extracting the fundamental frequency (F0) of noisy speech signals. Our method uses the recently proposed shift autocorrelation to locally emphasize harmonically distributed energy in the spectrogram. Subsequently, a trajectory extraction algorithm based on an optimization technique is used to determine local F0 contours of voiced segments. Our evaluation shows that the proposed method is capable of estimating F0 even in the presence of severe noises such as in radio communications.

Index Terms— Robust F0 estimation, shift-ACF

1. INTRODUCTION

Robust estimation of the fundamental frequency (F0) of a given speech signal is important in many speech processing applications. In this paper we consider the particular case that the underlying speech signal is corrupted by significant noise, as it is typical when dealing with outdoor recordings, phone calls, or radio communication. In such cases, established techniques for F0 estimation might fail as either voiced speech components may be distorted by the transmission channel or partially masked by secondary signals.

To allow for a reliable F0 estimation even under such adverse conditions, we suggest to use the recently proposed shift autocorrelation (shift-ACF) [1]. The shift-ACF is based on emphasizing multiply repeated signal components within a target signal. In this paper we consider F0 and its harmonics as such repeating components, allowing us to locally detect F0 candidates. Concatenating those candidates we then construct voiced speech segments as F0 trajectories. Hence, for a given speech signal, the proposed method both detects voiced regions and yields corresponding time-variant F0 estimates.

In recent papers on noise-robust F0 estimation [2] and multi-band pitch detection [3], Tan and Alwan divide the current methods for F0 estimation into three larger categories, namely time-domain-based, frequency-domain-based, and time-frequency-domain-based algorithms (see [4] and [5] for a comparative overview). A widely used method of the first kind is YIN, the fundamental frequency estimator for speech and music [6]. This method is based on the autocorrelation function with some additional modifications for error prevention and noise-robustness. Another state-of-the-art time-domain implementation of a pitch detection algorithm, which is included in the Snack Sound Toolkit and in WaveSurfer (http://www.speech.kth.se/snack/), is commonly known as the ESPS or get_f0 method and follows the robust algorithm for pitch tracking (RAPT) by Talkin [7], who uses a cross-correlation function. Both methods are included in the Praat software (http://www.fon.hum.uva.nl/praat/). An example for the frequency-based domain is the subharmonic summation method (shs) introduced in 1988 [8]. A more recent example for this category is SWIPE, the sawtooth waveform inspired pitch estimator for speech and music [9].

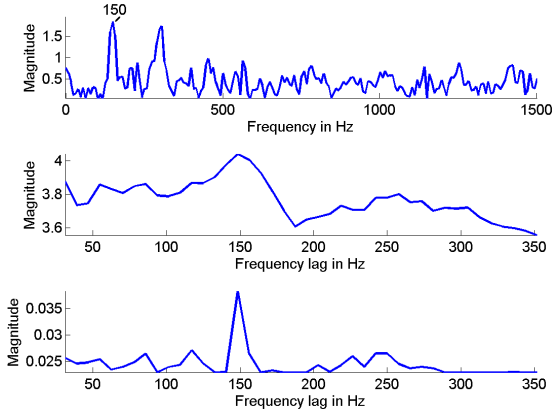

Fig. 1. Short-time magnitude spectrum (top), classical ACF (center), and type 100 shift-ACF (bottom).

In order to obtain a noise-robust pitch detector, Tan and Alwan developed a method [2, 3] that works both in the time and the frequency domain. Their pitch estimation algorithm is based on a correlogram, which is a two-dimensional autocorrelation plot showing correlation statistics. Our proposed approach employing shift-ACF falls as well in the latter category. However, it follows a different approach by applying a modified version of an autocorrelation function on the spectrum of a signal in order to emphasize and detect periodicity

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

in the frequency domain.

Our paper is organized as follows. In Sect. 2 we summarize the idea behind the shift-ACF. Sect. 3 then proposes how to exploit shift-ACF for F0 estimation by first locally detecting F0 candidates which are then concatenated to F0 trajectories using an optimization approach. In Sect. 4 we provide an evaluation and show that the proposed approach outperforms classical techniques in noisy speech scenarios.

2. SHIFT-METHOD AND SHIFT-ACF

The fundamental frequency of voiced speech may be observed as a high energy region within the short time spectrum x around a frequency F0. Characteristically, harmonic frequencies 2·F0, 3·F0, ... are as well of high energy. This motivates an approach to F0 estimation by detecting such repeated high energy regions as local maxima of the autocorrelation $\mathrm{ACF}[x](s) := \sum_{k\in\mathbb{Z}} x(k)\cdot\overline{x(k-s)}$. Fig. 1 shows a short time magnitude spectrum $|x|$ of voiced speech (top) with an F0 of about 150 Hz, producing visible peaks at F0 and the first few harmonics. In $\mathrm{ACF}[x]$ (center), this results in a high energy region at the lag frequency of 150 Hz. As in this example the speech is corrupted by noise, the peaks in x as well as in $\mathrm{ACF}[x]$ are not very pronounced.

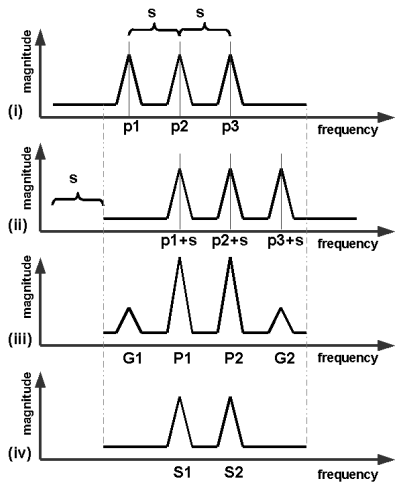

Fig. 2. (i) spectral signal x, (ii) frequency-shifted version x^s, (iii) shift-product $x\cdot\overline{x^s}$, and (iv) shift-minimum $\min(|x|,|x^s|)$

In [1], it has been proposed to exploit the presence of *multiple* signal repetitions to enhance the ACF. Moreover, in addition to comparing a signal x with its s-shifted versions $\overline{x^s(k)} := x(k-s)$ by using shift-products $P_s[x](k) := x(k)\cdot\overline{x^s(k)}$ as in classical ACF, it is proposed to additionally use a shift-minimum operator $M_s[x](k) := \min(|x(k)|,|x(k-s)|)$ to eliminate artifacts caused by non-repeating components. The effects of both operators are illustrated in Fig. 2, showing (i) a synthetic spectral signal x with a component at $p1$ repeated two times (at $p2$ and $p3$) at a lag of s between two

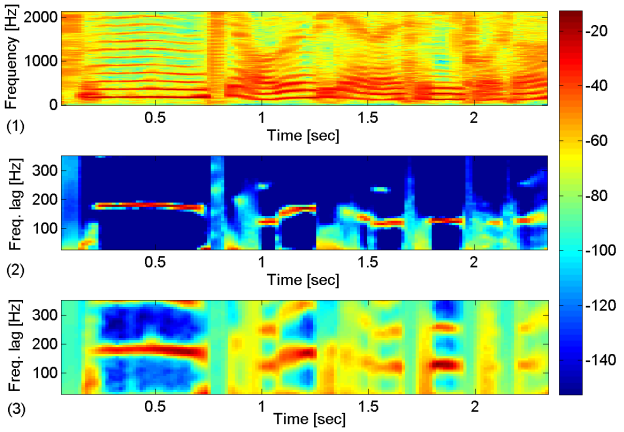

Fig. 3. (1) Speech spectrogram, (2) spectral shift-ACF type 100, (3) spectral ACF.

successive components. In (ii), the shifted version $x^s(k)$ is shown which is the main building block of classical ACF, where, for a lag s, $\mathrm{ACF}[x](s)$ is the sum over the shift-product $P_s[x]$ shown in (iii). Shift-products involving background noise may produce ghost components such as indicated by $G1$ and $G2$. In [1] the usage of the shift-minimum operator is proposed to eliminate such ghost components as illustrated in (iv), and hence avoid possible artifacts within the ACF.

By combining the *type 0* shift-product operator $\mathcal{O}_s^0 := P_s$ to emphasize repeating components and the *type 1* shift-minimum operator $\mathcal{O}_s^1 := M_s$ to suppress non-repeating components, a general shift-method framework is established by operator composition $\mathcal{O}_s^t := \mathcal{O}_s^{t_1}\circ\cdots\circ\mathcal{O}_s^{t_n}$ where $t=(t_1,\ldots,t_n)\in\{0,1\}^n$ specifies the sequence of applied minimum and product operators. The *shift-ACF of type* t is then defined as $\mathrm{ACF}^t[x](s) := \sum_{k\in\mathbb{Z}}\mathcal{O}_s^t[x](k)$. Note that classical ACF is the special case of a type 0 shift-ACF. As noted in [1], an n-fold iteration of shift-operators implies that repeated components are represented by peaks within the shift-ACF where the peak width is decreasing as a function of n, implying an improved sharpness. The latter is illustrated in our previous example: Fig. 1 (bottom) shows the type 100 shift-ACF. Clearly, the peak around 150 Hz is more pronounced as in the classical ACF (center).

3. F0-ESTIMATION USING SHIFT-ACF

To estimate the time-varying F0 we first compute the shift-ACF for successive time frames of a speech signal y. For this, we compute the spectrogram $\mathrm{SG}[y]$, where the $j-$th column $\mathrm{SG}[y]_{:,j}$ is obtained by computing the discrete Fourier transform of a suitably windowed version of the j-th time frame $(y_{jS},\ldots,y_{jS+N-1})$ of length N extracted from y using step size S. Then the *spectral shift-ACF* of type t is defined by $\mathrm{SpACF}^t[y](s,j) := \mathrm{ACF}^t[\mathrm{SG}[y]_{:,j}](s)$, i.e., by independently computing the shift-ACF for each spectrogram column.

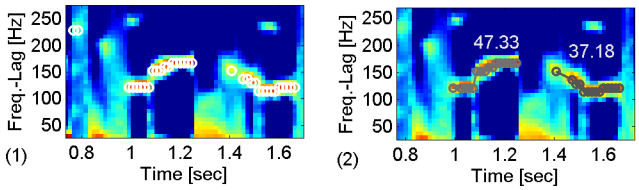

Fig. 4. (1) Peaks extracted from spectral shift-acf and (2) paths extracted by optimization approach.

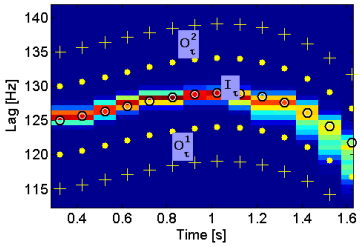

Fig. 5. Regions involved in computing trajectory sharpness.

Fig.3 shows the spectrogram (1) of a clean speech signal (male speaker) of length 2.4 seconds taken from the Kiel corpus [10]. For illustration, only frequencies up to 2 kHz are shown. In the center (2), the type 100 spectral shift-ACF is shown, where columns were postprocessed by normalization and thresholding by the median. The F0 is cleary visible by sharp temporal trajectories between 130 and 190 Hz. For comparison, (3) shows the type 0 spectral shift-ACF, corresponding to the classical ACF. Here, trajectories are more blurred and significant energy is present at harmonic lags.

Now we extract significant time-varying F0 trajectories from the spectral shift-ACF. First, a peak picking step is performed. As F0 trajectories evolve in temporal direction, this is done by successively considering each colum $c_j := \mathrm{SpACF}^t[y]_{:,j}$. After thresholding c_j by a smoothed, median-filtered version, peaks are picked iteratively. Using a greedy approach, in each step a maximum position is selected. In subsequent iterations, the neigborhoods of already chosen positions are ignored. In Fig. 4 (1), peaks extracted from a region of our example in Fig. 3 (2) are shown as white circles.

For trajectory extraction we consider the set of m extracted peaks as nodes in a graph. We then enforce paths by connecting each node to exactly one successor node by computing a bijection $\pi : [1:m] \to [1:m]$ such that the total cost $\sum_{i=1}^{m} C_{i,\pi(i)}$ of connecting nodes is minimized. The costs $C_{i,j}$ of connecting node i to j are chosen to provide reasonable F0 trajectories: $C_{i,j}$ is set to the Euclidean distance between peaks i and j, where $C_{i,j} := \infty$ if peak i temporally occurs after peak j. Furthermore, $C_{i,i} := \infty$ to prohibit 1-cycles. By introducing additional dummy nodes at a suitable maximum distance of each node, we furthermore allow a path to start or end at each node. The resulting optimization problem is a special case of a linear assignment problem (LAP) which can be efficiently solved using, e.g., the algorithm proposed in [11]. A result of the path extraction for our running example is shown in Fig. 4 (2).

Finally, paths which are too short or have only insignificant energy are discarded. For this, we use a trajectory sharpness measure such as in [1]. This measure, as illustrated in Fig. 5, basically computes a logarithmic energy ratio between an inner region I_τ around the estimated trajectory and an outer region $O_\tau := O_\tau^1 \cup O_\tau^2$. By construction, existing trajectories result in positive sharpness values. Fig. 5 shows the sharpness measure evaluated for the finally resulting F0 trajectories in white color.

Table 1. Test material and parameters for F0 annotation.

	Clean Scenario	**Real Scenario**
Database	KIEL-DB	RADIO-DB
Length	$\approx$ 10 minutes	$\approx$10 minutes
Fs	22050 Hz	8000 Hz
Language	German	various
Time Res.	$\approx$ 11 ms	$\approx$ 16 ms
Freq. Res.	$\approx$ 5.4 Hz	$\approx$ 3.9 Hz

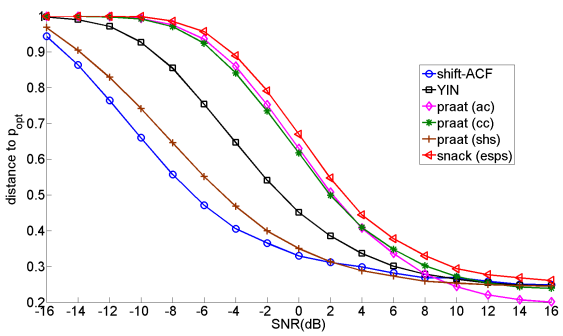

Fig. 6. Performance of the different algorithms on the KIEL-DB for different SNRs of added white noise.

4. EVALUATION

The aim of our proposed algorithm is to detect voiced segments in a noisy signal and provide F0 trajectories, i.e., time-varying F0 estimates, for such segments. To evaluate our method, we have conducted two different kinds of tests: experiments in controlled settings, i.e., with clean speech disturbed by noise and experiments on real audio signals, particularly focussing on a radio communications scenario. The test material consists of two databases of approximately 10 minutes length each. For clean speech we have used files taken from the Kiel corpus [10]: The files (refered to as KIEL-DB) consist of phrases in German language spoken by both women

Table 2. Performance of the different algorithms for clean speech disturbed by noise and for HF radio speech.

	Machine gun	***Factory***	RADIO-DB
Shift-ACF	**0.5143**	**0.7196**	**0.3529**
Praat (ac)	0.5712	0.9125	0.8002
Praat (cc)	0.5413	0.8516	0.6944
Praat (shs)	0.5598	0.8471	0.7111
YIN	0.6373	0.9112	0.8287
Snack (esps)	0.8185	0.8776	0.8054

and men. The sampling frequency (Fs) is 22050 Hz. Our test database of real audio scenarios consists of 8 kHz speech signals from a HF- (high frequency band) radio communication, and is refered to as RADIO-DB.

The ground truth for evaluating F0 estimation performance was annotated from the spectrogram using a Matlab-based annotation software. An overview of the test material we used as well as the time and frequency resolutions we have fixed for the manual annotation is given in Table 1. For annotation, we evaluated the F0 in regular time steps, which correspond to 11 ms for the KIEL-DB and to 16 ms for the RADIO-DB. Each file has been labelled by two persons. The resulting label files have been compared point by point. In cases where the annotated F0 differed by more than 15 Hz, annotation was reconsidered and adjusted manually.

In both experiments, our algorithm has been compared to other commonly used F0 estimation methods:

- Praat (ac), based on an ACF [12] and available with the Praat software http://www.fon.hum.uva.nl/praat/.
- Praat (cc), based on a cross-correlation analysis.
- Praat (shs), based on subharmonic summation [8].
- YIN, based on ACF and some modifications [6], available at http://audition.ens.fr/adc/sw/yin.zip.
- Snack (esps): A standard pitch tracking software using the Entropic Signal Processing Software (ESPS) algorithm which goes back to RAPT [7], a method also used in Wavesurfer [13]. An implementation can be found at http://www.speech.kth.se/snack/.

In order to compare all F0 estimation methods to the ground truth, we run all the algorithms using the same time resolution. In each step, we get an F0 estimate for the corresponding time interval. Each estimate is compared to the ground truth. A box of width equal to the step size and variable height is built around each F0-point of the ground truth. If the estimated F0 lies inside this box, a true positive (TP) is assumed, which means that the estimation is correct. F0 estimates outside the box regions are assumed to be false positives (FP). For all the results reported in this paper we have used an interval of ± 15 Hertz around the ground truth points to build the boxes. Running an estimation algorithm thus results in a *performance point* $p = (\text{FP-rate}, \text{TP-rate})$. Performance of an algorithm is measured by the Euclidean distance of p to the optimal point $p_{\text{opt}} = (0, 1)$, where a smaller distance means better performance.

Tests on clean speech were performed on the KIEL-DB. For each file we have run all the combinations of length two and three shift-ACFs, i.e., for all operators of types $t \in \{0,1\}^2 \cup \{0,1\}^3$, in order to find out the one performing best. In our experiments, the type 010 shift-operator yields the best results, closely followed by the type 100 operator. This optimum shift type was then used for F0 estimation on all of the noisy signals. To do so, noise with different signal-to-noise ratios (SNR) has been added to each file. In particular we have considered SNRs in the interval from -16 to 16 dB. The results on the whole database for added white noise are shown in Fig. 6. For each SNR value, the distance from its performance point to p_{opt} is indicated. Clearly with increasing noise level the shift-ACF method performs best.

In addition to adding white noise, we have added several other kinds noises taken from the NOISEX corpus to the KIEL-DB. Table 2 shows corresponding results for the cases of a machine gun noise with SNR ≈ -3dB and factory noise with SNR ≈ -8dB. Also in these cases the proposed shift-ACF leads to improved results. Table 2 furthermore shows the F0 estimation performance on the radio communication files. Here, noises are usually more severe depending on the characteristics of the radio channel. Also, in the HF radio band, *time varying* noise is usually considerable. In this case the improvement given by the shift-ACF is even more significant as in the artificial noise scenarios.

5. CONCLUSIONS

In this paper we proposed to use the recently introduced shift-ACF for estimating the F0 of noisy speech signals. The shift-ACF is used for emphasizing the harmonic parts of a speech signal based on the assumption that, for each voiced segment, at least a few adjacent harmonics are present. Extraction of a sequence of F0 trajectories is performed using a greedy peak picking technique with a subsequent path extraction step which is based on solving an optimization problem. In our experiments we compare the proposed method to classical approaches and show that significant improvements in F0 estimation may be obtained for the case of noisy signals.

Regarding future work we note that the selection of optimum types for shift-ACF up to now was experimental. However, suitable operator lengths (in our case 2 or 3) were motivated by an assumed minimum number of available harmonics. Furthermore, our theoretical investigations have shown that shift operators can be compared by a partial order, which can help to simplify operator selection in the future.

6. REFERENCES

[1] F. Kurth, "The Shift-ACF: Detecting Multiply Repeated Signal Components," in *Proc. IEEE WASPAA*, 2013.

[2] L. N. Tan and A. Alwan, "Noise-robust F0 estimation using SNR-weighted summary correlograms from multi-band comb filters," in *Acoustics, Speech and Signal Processing ICASSP*. IEEE, 2011, pp. 4464–4467.

[3] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech Communication*, vol. 55, 2013.

[4] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," in *Acoustics, Speech and Signal Processing ICASSP*. IEEE, 1976, vol. 24, pp. 399–418.

[5] W. Hess, *Pitch determination of speech signals: algorithms and devices*, Springer-Verlag Berlin and Heidelberg, 1983.

[6] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of The Acoustical Society of America*, vol. 111, pp. 1917–1930, 2002.

[7] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, pp. 518, 1995.

[8] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of The Acoustical Society of America*, vol. 83, 1988.

[9] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, pp. 1638, 2008.

[10] Institute for Phonetics and digital Speech Processing, University of Kiel, Germany, "The Kiel Corpus of Read Speech," http://www.ipds.uni-kiel.de/forschung/kielcorpus.en.html.

[11] R. Jonker and A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," *Computing*, vol. 38, no. 4, pp. 325–340, Nov. 1987.

[12] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *IFA Proceedings 17*, 1993, pp. 97–110.

[13] K. Sjölander and J. Beskow, "Wavesurfer - an open source speech tool," in *Proceedings INTERSPEECH*, 2000, pp. 464–467.