DIRECT TIME DOMAIN FUNDAMENTAL FREQUENCY ESTIMATION OF SPEECH IN NOISY CONDITIONS

Hynek Bořil and Petr Pollák

Czech Technical University in Prague, Faculty of Electrical Engineering
CTU FEE K13131, Technická 2, 166 27 Prague, Czech Republic
phone: +420 22435 2048, email: borilh@feld.cvut.cz, pollak@feld.cvut.cz

ABSTRACT

A new algorithm of direct time domain fundamental frequency estimation (DFE) and voiced/unvoiced (V/UV) classification of speech signal is presented in this paper. The DFE algorithm consists of spectral shaping, detection of significant extremes based on adaptive thresholding, and actual frequency estimation under several truth criteria. We propose a majority criterion for V/UV classification based on the detected frequencies consistency evaluation. Performance of the algorithm is tested on the Speecon database and compared to the Praat modified autocorrelation algorithm. In comparison to the Praat, the results indicate better properties of the DFE for clean speech and speech corrupted by additive noise to SNR about 10 dB. For lower SNR, sensitivity of the DFE to the speech component decreases rapidly while Praat fails to differentiate noise and unvoiced parts of speech from voiced parts.

1. INTRODUCTION

Fast and reliable estimation of fundamental frequency (f_0) is an essential task in the speech signal processing. Due to the large range of dynamic and voice color variations of speech produced in diverse noisy environments, robust frequency and voicing detection represents a complex problem.

Various algorithms based on different approaches have been developed recently. Generally, analyses are performed in the time domain, frequency domain or combine both domains. Time domain estimators are mostly based on autocorrelation function [1] and its modifications, e. g. autocorrelation analysis of the LPC residual signal [2]. In the frequency domain, modifications of short-time Fourier transform (STFT) [3, 4], cepstral analysis [5] and wavelet transform [6] are frequently used. The disadvantage of methods derived from STFT is problematic time/frequency resolution. Third group of algorithms utilizes advantages of both time and frequency domain analyses [7].

In despite of continual development, algorithms based on modified autocorrelation fuction still prove to be very robust and able to compete the other approaches. In consequence, the Direct Time Domain Fundamental Frequency Estimation algorithm (DFE) proposed in this paper was compared to the Praat modified autocorrelation function detection algorithm [8], which is respected and used widely in works involved in speech analysis and synthesis as referential, e. g. [9, 10]. Tests were performed on the Czech Speecon database [11].

The DFE was originally developed for monophonic pitch detector unit of the guitar MIDI converter [12]. The goal is to find an algorithm allowing real-time tone detection with relatively high time and frequency resolution and low detection delay. Since autocorrelation methods require signal segmentation and high number of variable by variable multiplications, it appears reasonable to examine the possibility of frequency detection directly from the shape of the signal in the time domain. Considering typical harmonic structure of speech signal, amplitude of the spectral component related to f_0 can be found significantly lower then amplitudes of the higher harmonics. Hence, conventional zero-crossing or peak-to-peak period measurement detection algorithms cannot be successfully used. However, once the f_0 component is emphasized to certain level by spectral preshaping, adaptive peak-to-peak detection followed by appropriate classification criteria can bring good detection results – and that is the idea of the DFE.

2. THE DFE ALGORITHM

Complete DFE chain, shown in Fig. 1, consists of envelope detector, pitch detector and evaluation part.

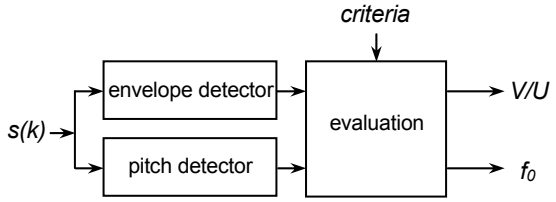

Figure 1: The DFE chain.

The envelope is determined as a short-time moving average of the signal energy, realized by low-pass FIR filtering of the squared signal. The filter order is chosen as a compromise between envelope smoothing and ability to follow fast energy changes on the boundaries of voiced/unvoiced parts of the speech signal.

The actual detected frequency from the pitch detector and corresponding value of the energy envelope are processed in the evaluation part by applying truth criteria and the most probable value of f_0 is estimated.

2.1 Pitch Detector

Actual f_0 candidate is evaluated from the distance between neighboring significant peaks – such local extremes that there is only one peak representing the absolute maximum and one the absolute minimum in the quasi-period of the signal. Structure of the pitch detector is shown in Fig. 2.

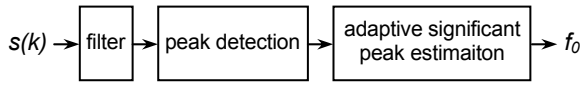

Figure 2: Pitch detector chain

2.1.1 Spectral Shaping

From the standpoint of f_0 quasi-period, strong higher harmonic components of the speech signal produce additional "false" peaks and zero-crossings in the time domain. To reduce them, spectral shaping by appropriate filter is used. It proved to work well to use a low-pass filter with significant tilt of the transfer function modulus over frequency range of f_0 typical occurence (60 – 600 Hz) to assure sufficient suppression of higher harmonics. To minimize transient distortion on fast amplitude changes of the filtered signal, low order IIR filter was chosen.

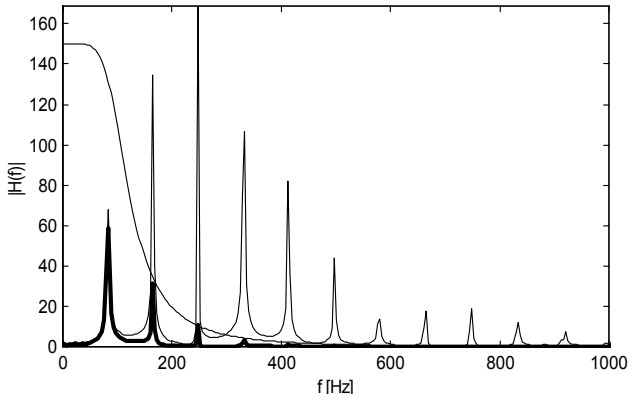

Figure 3: Spectral shaping by low-pass IIR.

An example of spectral shaping by low-pass filter is shown in Fig. 3, thin lines represent a spectrum of the original signal and transfer function of the filter (scaled 150x) and solid bold line a spectrum of the signal after spectral shaping.

2.1.2 Adaptive Significant Peak Estimation

After spectral shaping, all local extremes are detected. Due to the low order of the filter, some "false" peaks and zero-crossings still may remain in the signal. To identify locations of singificant extremes, the adaptive significant peak estimation based on neighboring peaks thresholding is performed, see Fig. 4.

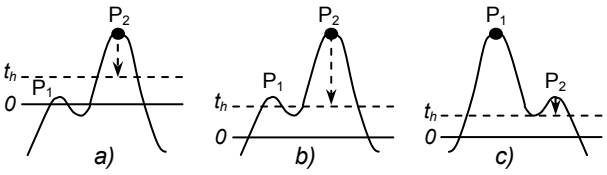

Figure 4: Adaptive significant peak thresholding.

P_1 is not significant peak related to the maximum if:

$$P_1 < 0 \ \cup \ ZC(P_{last}, P_1) = 0 \ \cup \ P_1 < P_2 \cdot th \ \cup \ \cup \ (P_1 < P_2 \ \cap \ ZC(P_1, P_2) = 0), \quad (1)$$

where $ZC(X, Y) = 1$ if there is at least one zero-crossing between peaks X and Y, else 0; P_{last} represents last detected significant peak. In other cases P_1 is significant peak related to the maximum. Afterward, P_2 is shifted to P_1, new peak becomes P_2 and the test is repeated. Significant peak related to the minimum is obtained by reversing the signs of inequality in (1). Finally, the frequency is determined from the distance between neighboring significant peaks related to maxima or minima.

As shown in Fig. 5, the peak estimation is robust to quasi-stationary additive noise in case the amplitude of additive noise is significantly lower than the amplitude of the speech signal.

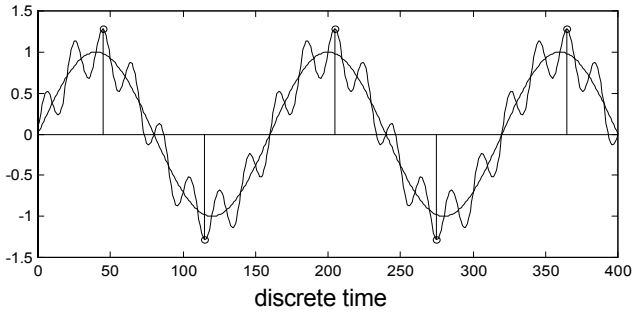

Figure 5: Significant peak estimation in harmonic additive noise.

2.2 Evaluation Part

Since the pitch detector returns all detected frequencies – even for speech silence and unvoiced parts of speech – and there still may appear some wrong f_0 estimations usually corresponding to frequency halving and doubling, several criteria are used to select voiced parts only and eliminate estimation errors. First criterion is related to the level of the signal – no frequency estimations are performed for levels of signal lower than the threshold E_{th}.

$$E(k) < E_{th} \ \Rightarrow \ f_{est} \neq f(k).$$

The actual level of energy $E(k)$ is evaluated by the envelope detector.

Second criterion – expected frequency range of f_0 – accepts no frequency out of specified range (60 – 600 Hz) as a valid estimation.

$$f(k) \notin (f_{floor}; f_{ceiling}) \ \Rightarrow \ f_{est} \neq f(k).$$

Third – M-order majority criterion – says that more than a half of M consecutive detected frequencies must lie in the same frequency band of chosen width. Let $\{f_m\}$ are M sequently detected frequencies, $\text{count}_{\text{fk}}(\{f_m\})$ – number of f that

$$f \in \{f_m\} \ \cap \ f \in \left(\frac{f_k}{\sqrt[24]{2}}; f_k \cdot \sqrt[24]{2} \right). \quad (2)$$

The interval in (2) equals to the frequency bandwith of 1 halftone – centered to f_k.

$$\begin{aligned} &\text{Let } p = \max_k(count_{fk}(\{f_m\})), \\ &\quad q = \arg\max_k(count_{fk}(\{f_m\})), \qquad k = 1,...,M. \end{aligned} \tag{3}$$

$$\text{If } p > \left\lfloor \frac{M}{2} \right\rfloor,\ M > 1 \quad \Rightarrow \quad f_{est} = f_q, \tag{4}$$

braces $\lfloor\ \rfloor$ represent round down. If more than one f_k satisfies (3) and (4), $f_{est} = f_{\min(k)}$. If majority criterion is satisfied, actual signal is evaluated as voiced. An example is shown in Fig. 6.

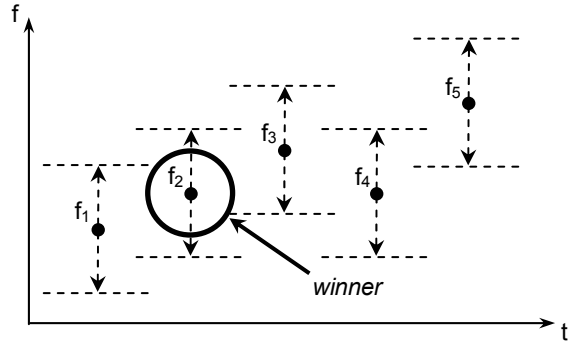

Figure 6: Majority criterion, M = 5.

3. THE DFE AND PRAAT TESTS

The DFE and Praat were tested on a selected part of the Speecon Database of Czech language [11]. Testing data consist of 366 sentences read by adult men and women in the office environment. Speech was recorded by 4 microphones placed in different distances from the speaker, hence 4 signals with different SNR are available. The nonstationary environmental noise consists of outside traffic noise and usual office sounds produced by computer fans, chair creaking, corridor door opening etc.

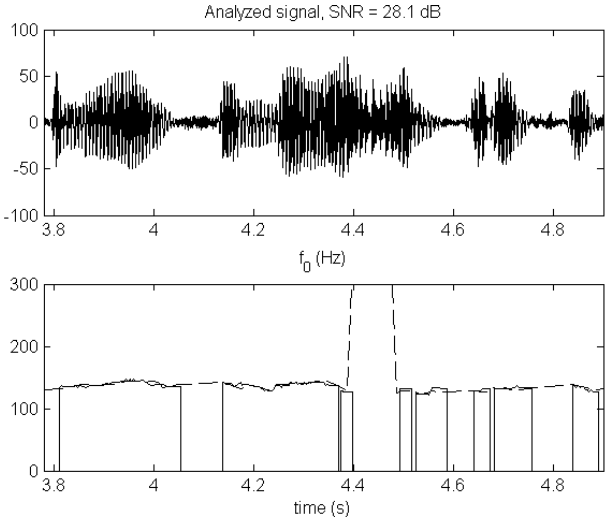

Figure 6: Example of detected frequencies by DFE and Praat.

Narrow band components that appear in the noise can often corrupt the f_0 detection more than white noise of the same SNR level artifically added to clean speech. The SNR is determined as a mean value and does not bring information about short-time noises of relatively high energy that often appear in the signal. Moreover, in the distant microphone channels a strong room echo component appears. This convolutional noise causes phase shifted signals sum and also increases the f_0 detection errors.

3.1 An example of DFE and Praat f_0 estimation results

An example of estimated frequencies by DFE and Praat is shown in Fig. 6. Since there is no explicit information about boundaries of voiced parts in the Praat output, Praat frequencies are connected by continuous dash line, frequencies detected by the DFE are plotted by solid line. In the time interval 4.4 – 4.5 s a typical Praat detection error is shown. In this case, unvoiced part of the speech signal is considered to be voiced, even though the detection is performed on the channel of SNR 28.1 dB and thus not much affected by additive noise.

3.2 Test evaluation criteria

During the test, following criteria were evaluated. *Compared frequencies* – number of detected frequencies in the channel compared to the refferential. First, detected frequencies in channel of SNR 28.1 dB by DFE and Praat were compared, Praat was understood to be refferential in this case. Consequently, mentioned channels became refferential for channels of lower SNR.
Average difference is defined:

$$\overline{\Delta} = \frac{1}{N}\sum_{n=1}^{N}\Delta_n, \quad \Delta_n = 1200 \cdot \log_2 \frac{f_2}{f_1} \quad (\%).$$

Tone frequencies in the musical scale are distributed exponentially. Difference 100 % represents a halftone distance, e. g. for 120 Hz and 300 Hz, halftones one step higher are 127.14 Hz and 317.84 Hz respectively.
Octave errors – number of differences equal or greater than one octave.
Standard deviation is defined:

$$\sigma = \sqrt{\frac{1}{N}\left(\Delta_n - \overline{\Delta}^*\right)^2}, \quad n = 1,...,N, \quad \Delta_n < 1200 \quad (\%),$$

where $\overline{\Delta}^*$ is an average error with octave errors excluded.
Voiced error is defined:

$$VE = \left| \frac{T_{ref} - T}{T_{ref}} \cdot 100 \right| \quad (\%),$$

where T_{ref} and T are total voiced times in the referential and compared channel. Since there is no explicit information about position of the voiced parts of speech in the Praat program, the voiced/unvoiced performace was not evaluated and could not have been compared to DFE.

SNR/SNR$_{ref}$ (dB)		compared freqs	$\overline{\Delta}$ (%)	octave err. (%)	σ (%)	voiced err. (%)
D/P	28.1/28.1	188734	39.69	1.11	64.31	N/A
D/D	17.9/28.1	147545	33.14	0.25	60.06	0.47
P/P	17.9/28.1	76957	80.44	3.16	66.50	N/A
D/D	9.6/28.1	94516	103.47	4.98	102.66	21.53
P/P	9.6/28.1	72742	133.64	5.48	92.12	N/A
D/D	4.9/28.1	5100	246.43	15.01	141.48	92.24
P/P	4.9/28.1	48096	1157.76	51.36	206.76	N/A

Table 1: The DFE (*D*) and Praat (*P*) comparison results.

3.3 Test results

Results of the DFE and Praat comparison are shown in Tab. 1. Symbols *D* and *P* represent frequencies detected by DFE and Praat, following numbers are SNR's of evaluated and refferential channel. In the first row of the results DFE and Praat (*D*/*P*) are compared on the channel of SNR 28.1 dB. In the next rows, channels of lower SNR are compared to the refferential channels.

As shown in the table, for SNR higher or equal to 9.6 dB, more information about frequency (*compared freqs*) is available in case of DFE due to detection both for significant maxima and minima of the signal. An *average difference* is lower in case of DFE for all SNR channels. In case of Praat, with lower SNR than 9.6 dB the difference grows rapidly, algorithm fails to differentiate noise and unvoiced parts of speech from voiced parts. The *voiced error* of the DFE is still less than 22% for SNR 9.6 dB. For lower SNR, the DFE sensitivity to voiced speech decreases rapidly in consequence of rising inability to locate the voiced parts of speech in the noise. For SNR 4.9 dB becomes DFE almost insensitive to the voiced speech and Praat returns values with *average difference* close to one octave.

4. Conclusions

New methods of direct fundamental frequency estimation (DFE) and voiced/unvoiced classification were proposed in this paper. The DFE algorithm is based on f_0 detection in the time domain and consists of spectral shaping, significant extremes detection realized by adaptive thresholding, and actual frequency estimation under truth criteria. For V/UV classification, a majority criterion based on the detected frequencies consistency evaluation is used.

In comparison to autocorrelation methods, DFE requires no signal segmentation, performs sample by sample f_0 estimation and preserves its phase, while the computation costs are significantly lower.

DFE was tested and compared to widely used Praat modified autocorrelation algorithm on selected part of Czech Speecon database. Test results approved that DFE brings better detection results (better time-frequency resolution, lower error ratios) for SNR to 9.6 dB (considering real environmental additive and convolutional noise). For lower SNR, sensitivity of the DFE to the speech component decreases rapidly while Praat fails to differentiate noise and unvoiced parts of speech from voiced parts.

REFERENCES

[1] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 727 – 730, Oct. 2001.

[2] H. Fujisaki, S. Narusawa, S. Ohno and D. Freitas, "Analysis and modeling of f_0 contours of Portuguese utterances based on the command-response model," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 2003, pp. 2317–2320.

[3] D. Arifianto, T. Kobayashi, "Performance Evaluation of IFAS-based Fundamental Frequency Estimator in Noisy Environment," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 2003, pp. 2877–2880.

[4] T. Nakatani, T. Irino and P. Zolfaghari, "Dominance spectrum based V/UV classification and estimation," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 2003, pp. 2313–2316.

[5] S. Ahmadi and A. S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 727 – 730, May 1999.

[6] L. Janer, "A modulated gaussian wavelet transform based speech analyser pitch determination algorithm," in *Proc. EUROSPEECH 1995*, Madrid, Spain, Sept. 1995, pp. 405 –407.

[7] D. J. Liu and C. T. Lin, "Fundamental frequency estimation on the joint time-frequency analysis of harmonic spectral structure," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 609 – 621, Sep. 2001.

[8] P. Boersma and D. Weenik, *Praat – A System for Doing Phonetics by Computer,* Eurospeech CD Software & Courseware, 1999.

[9] F. Tamburini, "Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 2003, pp. 129 –132.

[10] L. Devillers and I. Vasilescu, " Prosodic cues for emotion characterization in real-life spoken dialogs," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 2003, pp. 189 –192.

[11] C*zech Speecon Database*, http://www.speecon.com.

[12] H. Bořil, "Pitch detector for guitar MIDI converter," in *Proc. Poster 2003*, Prague, Czech Republic, May 2003, pp. 6–7.