Comparing Pitch Detection Algorithms for Voice Applications

Jan Bartošek Václav Hanžl
Department of Circuit Theory
FEE CTU in Prague
Technická 2, 166 27 Praha 6 -
Dejvice, Czech Republic
[bartoj11,hanzl@fel.cvut.cz]

Abstract — *The article deals mainly with objective comparisons of pitch-detection algorithms (PDAs) in area of speech signals processing. For this purpose evaluation framework was developed using for comparisons reference pitch database. A set of objective criteria was established too. All tested algorithms are also briefly described, new method MNFBC is presented in detail. Results show the biggest bottleneck in voice/unvoiced decision stage for the most of tested algorithms. Optimal time resolution for PDA is discussed too.*

1 INTRODUCTION

Intonation as a term for change of pitch (fundamental frequency, F0) of voice in time is one of most important prosodic features of our speech. Extraction of pitch contour can play an indispensable role in speech processing and recognition [1]. This task is not as easy as it may seem because in comparison with singing or prolonged fonation not all section of speech are voiced (have F0). That is why the pitch detection algorithm should not only estimate F0 as accurate as possible, but it should also detect correctly if the section of speech is voiced or unvoiced (V/UV). Although the research made in PDAs area is over 40 years old, we still do not have well-working one in previously mentioned aspects in conjunction with noise robustness. Objective comparisons between PDAs can be achieved by use of pitch reference database and suitable set of criteria. This article presents design and realization of such evaluation framework. Additionally it describes some od PDAs in more detail (especially MNBFC method) and according to achieved results it deals also with realization of simple voiced/unvoiced detector.

2 Voiced or unvoiced

Human voice originates in vocal tract that is depicted in figure 1. Our breath created in lungs goes initially through vocal cords muscle and pitch of voice is driven there by glottal pulses. Finally it is filtered by head cavities (nasal and oral) that act as resonators with formant frequencies. In this way voiced sections of speech having F0 are created (e.g. vowels). When vocal cords do not move the final sound shaped only by head cavities is similar to coloured noise and unvoiced speech is generated (most of consonants). Vocal folds cycle is also depicted in 1. The duration of cycle determines the fundamental frequency of voice and can be controlled by our will (thus we are able to sing).

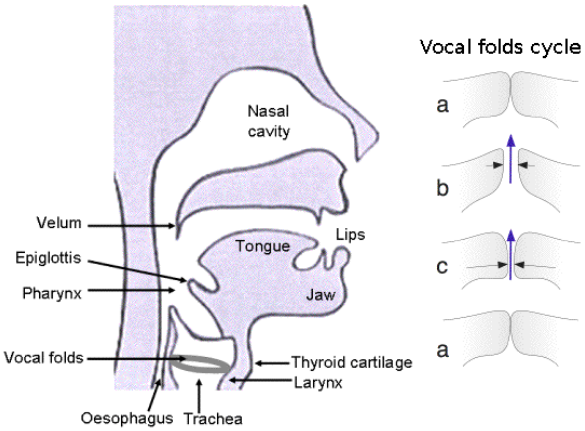


Figure 1: Human vocal tract and vocal folds

3 Audio processing

The basic block scheme of detecting F0 of an audio signal can be seen in the 2. The left grey areas of diagram present most general parts of voice or audio processing applications or algorithms. On the input is either whole recorded audio file (this approach is called offline processing) or direct data stream from microphone (online processing). Both types of processing finally end with selection of a frame of samples and let the algorithm itself do the job on it. Online processing has to be used in real-time use and is globally more difficult to deal with, because we do not know the data that will come in future (e.g. no statistics can be computed on

overall utterance etc.).

In our case of pitch detection algorithm it takes time frame of samples as input and on its output is estimated frequency in Hz for voiced frames or some info that the frame is not voiced for non-voiced frames. If certain PDA is not capable of voiced/unvoiced decision by itself, the optional V/UV block can be pre-ordered in the chain.

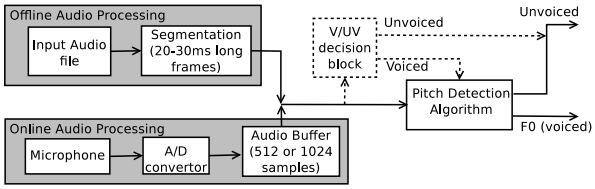


Figure 2: Basic block diagram of finding F0

4 Tested PDAs

4.1 Common implemented PDAs

Most of implemented PDA methods are theoretically described i [6], namely these are autocorrelation in frequency domain(ACF_freq), autocorrelation computed in time domain (ACF_time), Average Magnitude Difference Function (AMDF) and cepstral method (Ceps). Equations (1), (2), (3) and (4) describe these methods.

$$ACF_{time}(\tau) = \frac{1}{N}\sum_{n=0}^{N-n-1} x(n)x(n+\tau) \quad (1)$$

$$AMDF(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} |x(n) - x(n+\tau)| \quad (2)$$

$$ACF_{freq}(n) = IFFT\{[abs(FFT(x(k)))]^2\} \quad (3)$$

$$Ceps(n) = IFFT\{log(abs(FFT(x(k))))\} \quad (4)$$

4.2 Real-time time domain pitch tracking using wavelets

Method is described in [5] in detail and its presented results seemed very good. During implementation was found ,however, that mentioned used multi-level wavelet transform is in the method narrowed in each level into low-pass filter with subsequent decimation, as showed in equation (5). Then test for F0 candidate is done (peak-picking and searching for most central mode of time differences). If there is no candidate in current level of transform, the transformed signal goes into next level. The work [5] also presents idea of voiced/unvoiced detector based on energy ratios of thirds of actual frame. Such detector was tried but with its success rate was far behind usability and did not really meet reference results at all. That is why this PDA was knowingly tested without the V/UV stage (however algorithm itself is able to do very rough unvoiced evaluation where no candidate is found in last level).

$$a(n) = [x(2*n-1) + x(2*n)]/2 \quad (5)$$

4.3 Merged Normalized Forward-Backward Correlation (MNFBC)

This method of digital signal processing working in time domain was defined in [3] as base part of very complex PDA. Its core is computation of two correlations going against each other. Equations (7)/(8) show formulas for computing forward/backward normalised correlation, where constant MAX_PER refers to time period of lowest detectable frequency. The functions are always computed from frame with length of 4*MAX_PER. The courses of both functions applied on reference voiced part of utterance are depicted in 3a. Both of functions are then half-way rectified and used for computation of merged normalised forward-backward correlation MNFBC (9), its course is in the figure 3b. Equation (6) shows formal expression for used correlation term.

$$< x_{w_k}[n], x_{w_l}[n] > = \sum_{n=0}^{2*MAX\_PER-1} x_w[n+k]x_w[n+l] \quad (6)$$

$$NFC[t] = \frac{< x_{w_0}[n], x_{w_t}[n] >}{\sqrt{< x_{w_0}[n], x_{w_0}[n] >< x_{w_t}[n], x_{w_t}[n] >}} \quad (7)$$

4.4 Direct Frequency Estimation (DFE)

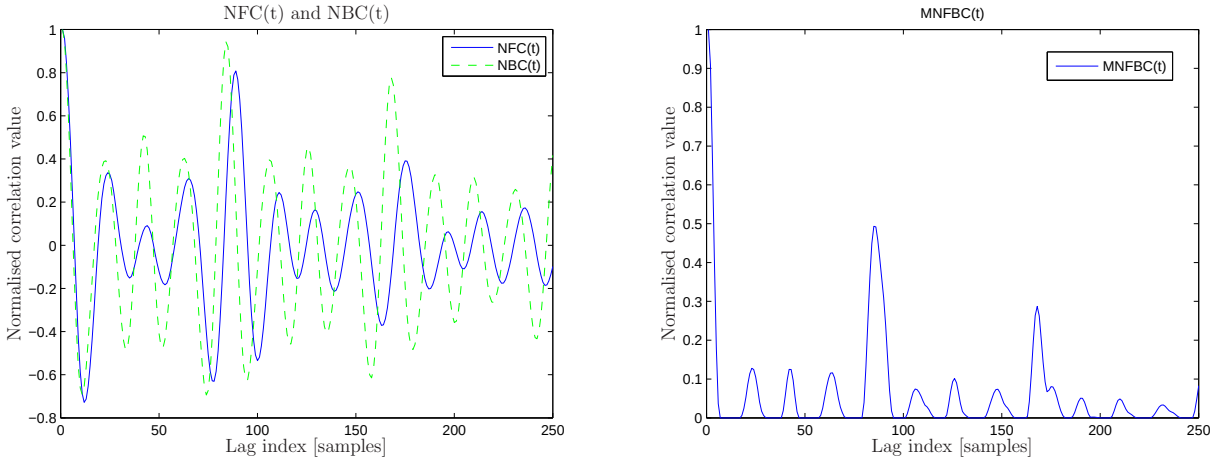
The DFE method works purely in time domain and is in detail described in [2] and the algorithm was overtaken in its binary form. It contains V/UV detection stage and is quite often used in various speech analysis related project in our department.

5 Optimal PDA Time Resolution Question

This part of paper presents some facts about biological capabilities of human voice tract leading to answer the question about convenient time resolution of PDA. This value says how often new F0 is computed. On one hand our aim is to have detailed information about course of F0, on the other hand there is by physical bases of voice tract certain

$$NBC[t] = \frac{< x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER-t}}[n] >}{\sqrt{< x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER}}[n] >< x_{w_{2MAX\_PER-t}}[n], x_{w_{2MAX\_PER-t}}[n] >}} \quad (8)$$

$$MNFBC[t] = \frac{< x_{w_0}[n]], x_{w_0}[n]] > (NFC'[t])^2 + < x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER}}[n] > (NBC'[t])^2}{< x_{w_0}[n], x_{w_0}[n] > + < x_{w_{2MAX\_PER}}[n], x_{w_{2MAX\_PER}}[n] >} \quad (9)$$



(a) Function NFC(t) and NBC(t)

(b) Function MNFBC(t)

Figure 3: Courses of forward and backward correlation functions on reference voiced frame

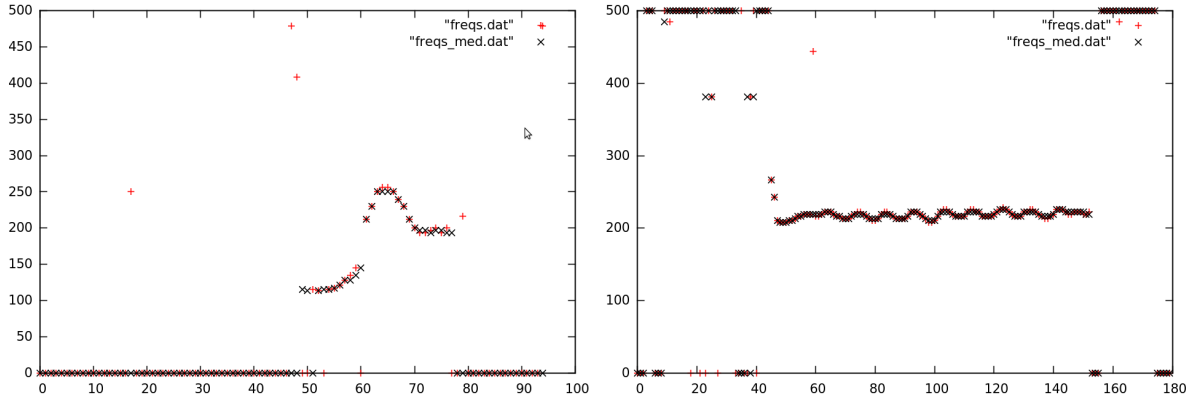
limit from which better resolution is not needed because whole information about F0 we already have and better time resolution leads only to increase in computation costs. This plays role especially in real-time applications with efficient computing resources in terms of electrical power and such lower computation power. Used reference database (see section 3) uses time step 1ms which is quite high resolution and the question is if we need it.

An answer can be found for example in [7], where a speed of pitch change of human vocal tract was studied. According to it the fastest pitch movement in Dutch speech is 50 semitones per second (50 cents per 10ms). This is experimental limit number of our physiology and is rarely achieved in real speech and intonation. Also 50cents (half of semitone) is very good frequency resolution for our purposes. For illustration two sample courses of F0s are included, both are results of tested ACF in frequency domain PDA. In Fig. 4a a time course of intonation of question with very fast intonation is depicted. The time resolution in this case was 23ms (sampling frequency 11kHz, 256 samples shift of frames). We can see that in place of fastest change there could be more detected frequencies. That is why time resolution of 16ms (sampling frequency 16kHz, 256 samples shift of frames) was tested in Fig. 4b on fast vibrato voice of singer. From this picture is obvious that 16ms time step is enough. Study [1] also presents the fact, that rate of pitch change is faster for a larger pitch interval than for smaller one. The conclusion of the section is that we do not need as high time resolution as reference database offers and in tested algorithms time step of 16ms will be sufficient.

6 Pitch Reference Database

When we want to evaluate the PDA, we need to know correct outputs for sample data. In this work a manually pitched-marked part of Speecon Spanish database was used. This pitch-marked part of database is quite known across the PDA creators all over the world. The reference part was created as part of work described in [4] as a result of final utilization of pitch-marking algorithm (pitch-mark is a defined start of glottal cycle) and then also manually corrected. Having these pitch-marks[1] we can easily compute the F0s from them as inverted value of their time distances. The used database has following specification: raw audio data format with sampling frequency of 16kHz, 2B/sample, linear-coding, mono. In recordings there are 60 speakers (30 males, 30 females). An overall length of

[1]Pitch-mark is well defined time instant in glottal cycle detectable in speech signal



(a) Time resolution of 23ms, very fast question melodeme

(b) Time resolution of 23ms, vibrato sing

Figure 4: Influence of time resoultion choice onto detected intonation course

speech signal is about 1 hour which means that there is 1 minute of speech material per speaker. The database is simultaneously recorded by 4 microphones varying in distance from speaker so there are 4 channels varying in SNR. It also contains recordings varying according to environments varying in type and level of background noise (Car, Office, Public places). Except F0 reference data the database includes mentioned pitch-marks and also silence/voiced/unvoiced information.

7 Pitch Evaluation Framework

7.1 Motivation

Main motivation for rising the pitch detection algorithms evaluation framework is not only possibility of their objective mutual comparisons against known reference, but also finding optimal settings for parameters of certain PDA. Various evaluation criteria and evaluation across different categories allow us to pick most suitable PDA for needs of certain application (e.g. error rate in V/UV decisions compared to accuracy in frequency estimation of F0).

7.2 PDA File Formats

There are three basic formats commonly used to store a pitch information of acoustic signal in time as follows:

Type 1 refers to native type of .pda files of reference pitch-marked database containing pitch information. There is no special information about the time step in it (time step is considered to be known a priori, e.g. 1ms in used database) and thus the file starts directly with pitch frequency one per each single line. There is also silence/unvoiced/voiced information encoded in the values. Value 0 means silence, value 1 means unvoiced part of speech and values higher than 1 should be interpreted as valid F0 frequencies.

Type 2 is very close to type1 with the only difference occurring at very first line of file. There is saved time step information which says for how long time period is each F0 valid.

Type 3 does not match to any of so far mentioned types. The main difference is that there is no constant time step for F0s and there is a couple of numbers on each line. First number describes F0 and second one is time in seconds when this F0 ends in signal. Type 3 is the most efficient in memory requirements because if compresses the information.

7.3 Design and implementation of evaluation framework

Fig. 5 presents global pitch evaluation framework architecture block scheme. The core of framework is pitch reference database which consists of testing audio files and their correct PDA reference files. List of tested audio files goes on the input of “PDA run script” box, which is responsible for calling the PDA algorithm on single audio files. It is capable of calling various implementations of PDAs – native operation system binaries (C,C++) and also can call PDA M-file in Matlab environment from shell. The only requirement on PDA is that it needs to be capable of creating the output .pda file in one of known formats. The special note is needed to be done to V/UV (Voiced/Unvoiced) decision box, that is not implemented in current stage of framework and should be preceding as an optional part of PDA if PDA is not written with the ability of doing this decision by itself. If PDA produce some other type of .pda file than type 1 (most common are types 2 and 3), the convert script needs to be called

to create type 1 .pda file. Having this file we can run single report script that evaluates the output of PDA in comparison to reference .pda file. Many evaluation criteria are computed, but only on single audio files. Then having set of single report files we are able to run global report scripts that firstly compute global report file for certain PDA and secondary many other report files are computed across all categories and their combinations (e.g. channel0 only in car environment).

The framework was implemented under UNIX type OS as combination of script around reference database. Scripts were mostly written as combination of multi-platform interpreted PERL language (main logic) and BASH (basic file operations). The whole environment is thus very easy to port to another platforms. Instead of used reference database could be with minimal effort used completely different pitch reference database and whole framework could help in areas outside speech technologies systems (e.g. musical segment).

7.4 Set of evaluation criteria

There are a few criteria commonly used in the area of evaluating pitch detection algorithms [3], but one of aims of the work was also reasonably suggest some new ones. The voiced error VE (unvoiced error UE) rate is proportion of voiced (unvoiced) frames misclassified as unvoiced (voiced). Gross error high GEH (gross error low GEL) is rate of F0 estimates (correctly classified as voiced) which does not meet the 20% upper (lower) tolerance of frequency in Hz. GEH and GEL 20% tolerance range is quite large and thus can not distinguish clearly between two precise PDAs. That is why GEH10 and GEL10 were established analogically to GEH and GEL but with only 10% tolerance ranges. These new criteria are also expected to result in higher error rates than older ones, but might be useful in applications where precision matters. Sometimes UE+VE and GEH+GEL criteria are used to summarize errors of PDA. Halving errors (HE - estimated frequency is half of reference) and doubling errors (DE) were also brought in with a tolerance of 1 semitone range from half or double of reference F0). These kind of errors are special type of gross errors and occur often on real PDA outputs for noisy signals or transitions from voiced to unvoiced parts of speech. Sometimes we could need to watch errors not in entire frequency band but e.g. within 5 smaller frequency sub-bands individually (2/3 octave bands were used to cover range of 60 to 560 Hz). Statistical data based on frequency values (absolute difference between reference and estimate means, standard deviations) can be also seen in literature computed over whole reference and estimated F0 data set. But these statistics do not have very predicative value thanks to logarithmic course of our hearing. That is why modified statistical cireria according to [2] were used - mean difference $\overline{\Delta}_{\%}$ (10) and standard deviation $\delta_{\%}$ (11) both computed in semitone cents.

$$\overline{\Delta}_{\%} = \frac{1200}{N} \sum_{n=1}^{N} \log_2 \frac{F_{est}(n)}{F_{ref}(n)} \tag{10}$$

$$\delta_{\%} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} [1200 \log_2 \frac{F_{est}(n)}{F_{ref}(n)} - \overline{\Delta}_{\%}]^2} \tag{11}$$

For explanation is needed to be added that criterion VE+UE is not sum of error rates, but is defined as ratio of count of all wrong classified sections to count of all sections. Not to accumulate errors into next criteria are for further processing passed only those voiced sections that were correctly classified as voiced. That could in certain situations advantage F0 accuracy of PDAs that have high rate of VE, because for these PDAs the accuracy is computed only from frames that were classified as voiced and thus problematic frames (that could apparently decrease accuracy in general) may not be present in accuracy computation.

7.5 Results

All the PDAs mentioned in section 4 were tested. Results on this set of PDAs on channel 0 (highest SNR value, close-talk microphone) and channel 1 (lavalier microphone) are presented in tables 1 and 2. Some of algorithms (ACF time, AMDF, CEPS a MNBFC) were implemented without decision thresholds for voiced/unvoiced classification nor any V/UV detector was pre-ordered before them. This is why they classify all unvoiced segments as voiced and UE criterion reaches value of 100 percent. This makes them more difficult to compare to the rest with V/UV decision stage, but on the other hand these PDAs are directly comparable in accuracy without any further discussion needed. Interesting thing is that algorithms AMDF and CEPS have reached (although working on really different approach) almost same results in all criteria. From results point of view they seem to be equivalent what is in contrast with claims of some articles presenting them as complements and building more advanced PDAs on their combination. New method MNBFC has unfortunately showed results worse then ACF_freq. Also can be seen which PDAs tend to do more GEH or GEL and

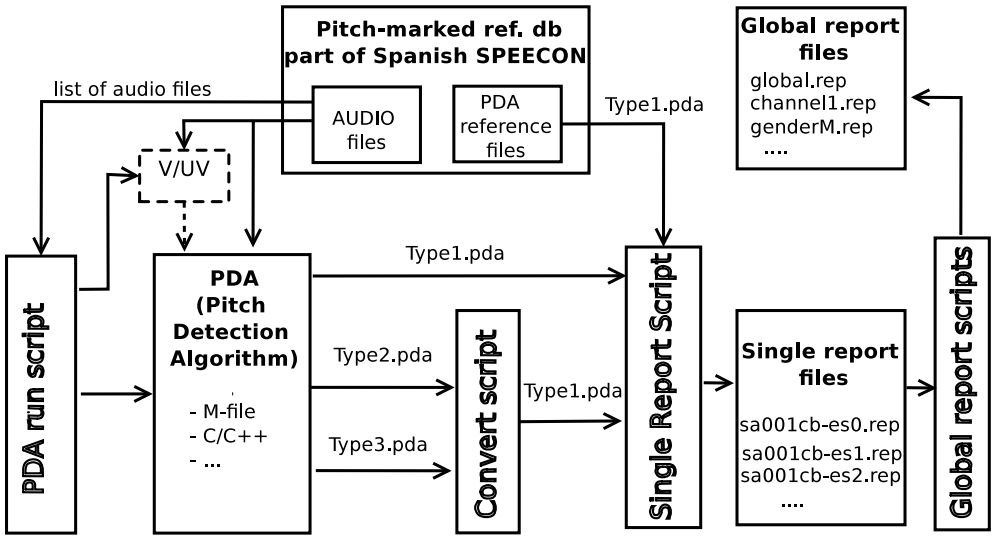


Figure 5: Pitch Evaluation Framework architecture – block scheme

also their corresponding halving and doubling error rates. Next confirmed fact is decrease in accuracy and also V/UV stage with increasing noise level (channel 0 compared to channel 1, biggest downgrade in GEH for Acf_freq). Most robust method from tested test is claimed to be DFE.

7.6 Additional experimental V/UV block

Based on preceding results the basic detector of voiced and unvoiced parts of speech was implemented according to [4]. It is based on ratios of signal energy (E) to zero-crossing rate (ZCR). The energy is computed from preprocessed frame, when there is emphasised periodical structure of voiced frames by applying short-time energy envelope. Formula for detector output function EZR is in equation (12). It can be said that for voiced segments value of EZR is high, because energy of signal is quite big and ZCR is compared to noise lower. On the other hand for unvoiced segments with high ZCR values and low energy will the final EZR value be low too. Evaluation framework was enriched by module allowing separate valuation of success rate of V/UV detectors. The detector base on EZR function with empiric threshold reached on channel 0 results worse than DFE method (UE+VE: 24,3 % for EZR versus 20,4 % for DFE), but for basic tasks it could be pre-ordered to PDAs without this function V/UV block and could so increase their global results.

$$EZR[m] = \frac{\overline{E}[m]}{ZCR[m]} \tag{12}$$

8 Results

Except implementations of basic PDAs there were studied also more advance pitch detecting algorithms. Experimentally was verified that time resolution of 16ms is suitable for needs of following intonation of speech. For objective PDAs evaluation the framework was designed and implemented based on existing pitch reference database. Set of various PDA evaluation criteria was proposed enabling detailed analysing of PDA behaviour in various conditions. According to our expectations overall error rate of all PDAs increases with lower SNR rapidly. Method based on merged normalised correlations (MNFBC) unfortunatelly did not bring expected results in F0 estimation accuracy. Results also show, that weakest point of all algorithms is voiced/unvoiced detection phase.

Acknowledgments

The research was supported by grants GAČR 102/08/0707 "Speech Recognition under Real-World Conditions", GAČR 102/08/H008 "Analysis and modelling biomedical and speech signals".

Reference

[1] Bartošek, J. Prozodie, zjištění a využití základního tónu v rozpoznávání řeči. *Semináře katedry teorie obvodů, analýza a zpracování řečových a biologických signálů - sborník prací 2009* (2009), 1–8.

[2] Bořil, H.; Pollák, P. Direct time domain fundamental frequency estimation of speech in noisy

PDA	**VE** [%]	**UE** [%]	**VE+UE** [%]	**GEH** [%]	**GEL** [%]	**GEH10** [%]	**GEL10** [%]	**DE** [%]	**HE** [%]
ACF freq	44,4	23,5	31,6	1,2	0,1	1,5	0,18	0,4	0,06
ACF time	0	100	61,9	4,7	2,3	6,2	3,5	0,8	1,3
AMDF	0	100	61,9	0,6	27,2	1,4	28,3	0,1	16,2
CEPS	0	100	61,9	0,6	27,1	1,4	28,1	0,1	16,0
DFE	26,6	15,5	20,4	8,4	4,2	16,5	8,9	0,2	1,3
Wavelets	67,7	11,3	32,7	2,5	4,9	3,7	6,0	1,1	3,9
MNBFC	0	100	61,9	4,8	4,4	6,6	6,6	0,4	2,8

Table 1: Overall channel 0 results

PDA	**VE** [%]	**UE** [%]	**VE+UE** [%]	**GEH** [%]	**GEL** [%]	**GEH10** [%]	**GEL10** [%]	**DE** [%]	**HE** [%]
ACF freq	52,7	34,1	41,3	23,3	0,1	23,5	0,2	3,2	0,03
ACF time	0	100	61,9	28,8	2,5	29,8	3,4	3,6	1,5
AMDF	0	100	61,9	10,8	44	11,3	45,2	1,3	21,4
CEPS	0	100	61,9	10,1	43,4	10,5	44,7	1,3	21,4
DFE	45,4	11,1	25,9	8,5	8,1	17,9	13,1	0,05	4,3
Wavelets	70,4	9,5	32,6	14,3	9,9	17,4	11,6	4,3	6,7
MNBFC	0	100	61,9	29,1	4,9	30,4	6,5	2,1	3,1

Table 2: Overall channel 1 results

conditions. *in Proceedings of EUSIPCO 2004 (European Signal Processing Conference, Vol. 1)* (2004), 1003–1006.

[3] Kotnik, B.; et al. Noise robust f0 determination and epoch-marking algorithms. *Signal Processing 89.* (2009), 2555–2569.

[4] Kotnik, B.; Höge, H.; Kacic, Z. Evaluation of pitch detection algorithms in adverse conditions. *Proc. 3rd International Conference on Speech Prosody, Dresden, Germany* (2006), 149–152.

[5] Larson, E. Real-time time domain pitch tracking using wavelets. *Journal of the Acoustical Society of America* (2005), 111(4).

[6] Uhlíř, J. *Technologie hlasových komunikací.* ČVUT Praha, 2007.

[7] Xu, Y.; Sun, X. Maximum speed of pitch change and how it may relate to speech. *Journal of Acoustical Society of America, Vol. 111, No. 3* (2002), 1399–1413.