PITCH DETECTION ALGORITHM EVALUATION FRAMEWORK

J. Bartošek

Department of Circuit Theory, FEE CTU in Prague
Technická 2, 166 27 Praha 6 - Dejvice, Czech Republic
email: bartoj11@fel.cvut.cz

Abstract

When implementing an existing or designing a brand new pitch detection algorithm (PDA) there is a need to evaluate it and maybe compare it to another one. We could make some comparisons or rough estimates based on graphical output, but this approach is ineffective and definitely inconvenient for larger file-bases. In this paper design and implementation of PDA speech-oriented evaluation framework is presented. It is based on existing ECESS PMA/PDA manually pitch marked reference database. The database consists of four-channel recordings in various environments with varying SNR and thus is suitable for testing robustness of pitch detection algorithms. Standard evaluation criteria widely used by many authors in the field are discussed and few improvements are suggested in this area. Finally results of comparing several PDA algorithms are presented and further idea of using the best one in a real-time punctuation detector is presented too.

1 Introduction

A pitch of a human voice (further also mentioned as F0 or base frequency) is fundamental frequency of glottal pulses when we are pronouncing voiced parts of speech. It is base of intonation of our speech as an important part of prosody information and could have reasonable utilization in some areas of speech processing and generation. A pitch detection algorithm (PDA) is such kind of algorithm that has sampled audio data on its input and tries to identify base frequencies of the part of that audio signal with some time step. On its output is usually a data file with time series of F0s in one of common PDA file formats.

Motivation of creating pitch detection algorithm evaluation framework (PDAEF) composes of two basic tasks in which could existence of the framework help a lot. The first one is possibility for automatic finding of optimal parameters for certain PDA by evaluating its separate runs varying in parameters values. The second utilization is comparing different PDAs to each other, because comparing large series of F0s only by sight is impossible, plots made from output files are better, but also very time consuming and thus very inefficient. PDAEF with suitable evaluation criteria allows us to make comparisons of algorithms on highest level from few resulting numbers. Another important point in this kind of evaluation framework is need of pitch reference database, which are we PDAs comparing to. More about used part of Spanish Speecon Database can be found in section 3.

2 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 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 [1], 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. 1 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. 2 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.

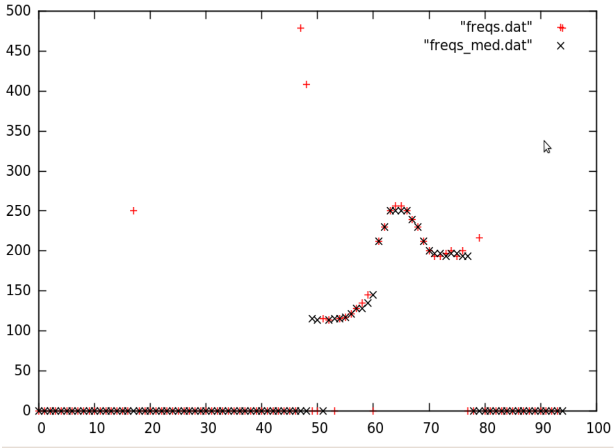

Fig. 1: Intonation of very fast question, resolution 23ms

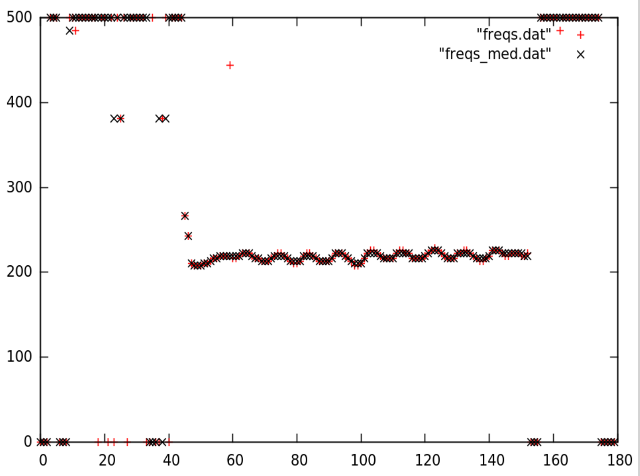

Fig. 2: Fast vibrato voice, resolution 16ms

3 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 [2] 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 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 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.

4 Pitch Evaluation Framework

4.1 PDA File Formats

There are three formats commonly used to store a pitch information of acoustic signal in time. In the following text I will use certain name conventions for types, which are 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 most efficient in memory requirements because if compresses the information.

4.2 Framework Architecture

Fig. 3 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 file. 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 env).

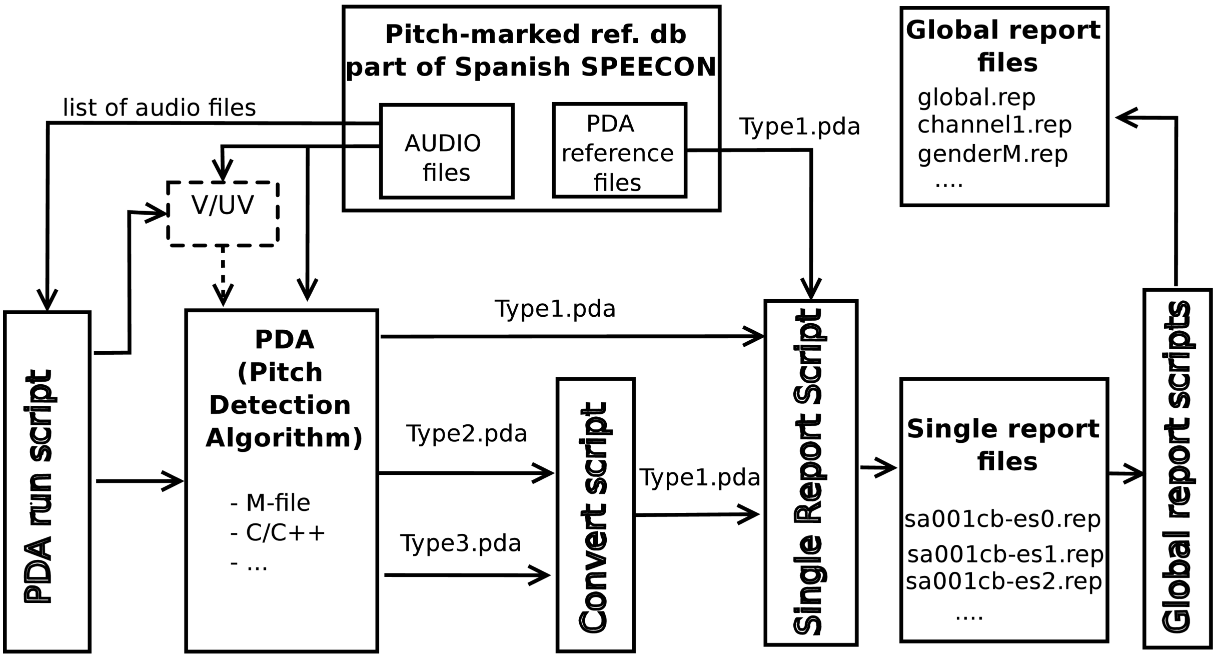

Fig. 3: Pitch Evaluation Framework architecture – block scheme

4.3 Framework Implementation

An aim of an implementation of the framework was to give final user a possibility to change the code to his own needs without a need to recompile it. That is why interpreted languages were used. PDA evaluation framework is implemented as a set of bash and perl scripts around reference DB and PDA executables. Short bash scripts were used only for very basic operations, main logic is written in multi-platform perl language, which is in its basic version very powerful for writing any kind of scripts and is very well documented. Used bash scripts could be with little effort very easily replaced with windows batch files and thus the framework can support windows platform too.

4.4 Common PDA Evaluation Criteria and their improvements

There are a few criteria commonly used in the area of evaluating pitch detection algorithms. 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. Sometimes UE+VE and GEH+GEL criteria are used to summarize errors of PDA. Statistical data based on frequency values (mean, standard deviation) can be also seen in literature computed over whole reference and estimated F0 data set:

- AbsMeanDiff: Absolute difference between the mean values of the reference and estimated pitch over the whole signal (in Hz).
- AbsStdDiff: Absolute difference between the standard deviations of the reference and estimated pitch over the whole signal (in Hz).

Besides these well-established criteria some improvements are suggested in this paper. Firstly the GEH and GEL 20% tolerance range is quite large and thus can not distinguish clearly between two precise PDAs. So let GEH10 and GEL10 be analogical to GEH and GEL but with only 10% tolerance instead of 20% one. These new criteria are also expected to result in higher error rates than older ones, but might be useful in applications where precision matters. Secondly halving errors (HE - estimated frequency is half of reference) and doubling errors (DE) were 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. Thirdly 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). Finally according to [3] statistics are computed based on pitch in cents of semitones, not in frequency in Hz. This change compensates logarithmic scale of human perception of frequencies and gives sense to statistical data computed over the whole signal:

- mean difference (in semitone cents): $\bar{\Delta}_{\%}=\frac{1200}{N}\sum_{n=1}^{N}\log_2\frac{F_{est}(n)}{F_{ref}(n)}$
- standard difference (in semitone cents): $\delta_{\%}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}[1200\log_2\frac{F_{est}(n)}{F_{ref}(n)}-\bar{\Delta}_{\%}]^2}$

In implementation one-pass Knuth-Welford algorithm was used for computing standard deviation.

4.5 Tested PDAs

In initial state after implementing the framework four basic and two more advanced pitch detection algorithms were tested. Autocorrelation in time domain (ACF time), autocorrelation in frequency domain (ACF freq), average magnitude difference function (AMDF) and cepstral (CEPS) method were considered as basic. Real-time time domain pitch tracking using wavelets (Wavelets) [4] and Direct Frequency Estimation (DFE) [3] are considered as more advanced PDA methods. Some of them should be also able to decide whether processed audio signal is voiced or unvoiced, this could be seen from results tables in next section.

4.6 Results

Results of only a few most important evaluation criteria are presented in this section. Table 1 presents overall global results computed over all 4 channels. To see how tested PDAs can deal with more noiseless signals, Table 2 presents same results but only on channel 0 (closest microphone) with highest signal to noise ratio. That is why a lot better results are expected in this second table. At first sight there is noticeable that in both tables PDAs ACF time, AMDF and CEPS do not do voicing detection stage by themselves well, from unvoiced errors value around 100% is clear, that these PDAs classify all frames of signal as voiced. ACF freq PDA is capable of rough voicing decision and reaches good values in GEH and GEL, especially on channel 0. There is also remarkable decrease in gross error rates comparing global results and channel 0 results as expected. This is really noticeable in every algorithm except DFE, which is quite robust and thus the difference is not as big. From gross errors point of view the ACF freq gives best result on channel 0 but note that algorithm tends to do more than 10 times more high errors then low ones, AMDF and CEPS have an opposite problem. With additional voicing decision stage ACF freq PDA could give really good results on signals with high SNR. Our future aim is to get PDA with low VE/UE error rates and also low GEH/GEL by adding stand-alone V/UV decision box in front of existing or implementing new PDAs.

PDA	VE[%]	UE[%]	GEH[%]	GEL[%]	DE[%]	HE[%]
ACF freq	57,1	26,3	15,6	0,09	3,4	0,03
ACF time	0	99,99	22,1	4,4	4,1	2,2
AMDF	0	100	5	50	0,8	21,7
CEPS	0	100	4,9	48,7	0,8	21,5
DFE	47	12,5	8,6	7,6	0,1	3,2
Wavelets	70	13	16,3	13,9	4,4	7,4

Table 1: Global results over all 4 channels

PDA	VE[%]	UE[%]	GEH[%]	GEL[%]	DE[%]	HE[%]
ACF freq	44,4	23,5	1,2	0,1	0,4	0,06
ACF time	0	99,99	4,7	2,3	0,8	1,3
AMDF	0	100	0,6	27,2	0,1	16,1
CEPS	0	100	0,6	27,1	0,1	16
DFE	26,6	15,5	8,4	4,2	0,2	1,3
Wavelets	67,7	11,3	2,5	4,9	1,1	3,9

Table 2: Results in Channel 0 (with highest SNR)

4.7 Use of Framework

The framework will be used to determine the best PDA from large set of implemented PDAs considering not only gross error rates but also algorithm robustness. The winner is then going to be used in a real-time punctuation detector of continuous speech.

The use of framework is not strictly limited on speech data only. It can be also with little effort adjusted to be applied to any other pitch reference database (e.g. testing musical oriented PDAs with a reference database for musical signals).

5 Conclusion

An introduction to pitch detection algorithms and its evaluation were presented in the paper. Question about time resolution of PDA was discussed and answered. Pitch Evaluation Framework and its architecture were introduced and new evaluation criteria were suggested. Few initial results were presented and explained, results are not as good as expected and show, that mainly more work has to be done in voice decision stage for basic PDAs. Possible utilization of the framework was also brought in.

Thanksgiving

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”.

References

[1] 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, pp. 1399–1413, March 2002

[2] Kotnik B. et.al., Noise robust F0 determination and epoch-marking algorithms. Signal Processing 89. 2009, pp. 2555-2569, doi:10.1016/j.sigpro.2009.04.017.

[3] Bořil H., Dissertation Thesis, FEE CTU in Prague, 2008

[4] Larson E., Real-time Time Domain Pitch Tracking Using Wavelets, In Proceedings of the University of Illinois at Urbana Champaign Research Experience for Undergraduates Program, 2005.