Exploring Abilities of Merged Normalized Forward-Backward Correlation for Speech Pitch Analysis

Jan Bartošek, Václav Hanžl

Department of Circuit Theory, FEE CTU in Prague, Czech Republic
{bartoj11;hanzl}@fel.cvut.cz

Abstract – **The article deals with usage of time-domain merged normalized forward-backward correlation (MNFBC) for pitch estimation of speech signals. This method should prevent from shortcomings of other methods commonly used in pitch detection algorithms (PDA). The text also presents comparison of possible improvements for voicing decision stage of MNFBC and also puts mind to final fundamental frequency (F0) smoothing with Viterbi algorithm. The precision and voiced-unvoiced (VUV) decision was compared against pitch reference database (part of Spanish Speecon). Results show that F0 estimate precision of MNFBC in connection with Viterbi smoothing using cents conversion in transition probability function is comparable to PRAAT cross-correlation. Although with additional signal energy thresholding unvoiced errors for close-talk channel 0 are lowered, the results are still better in PRAAT algorithm, but the difference gets even for channel 1 (lavaliere microphone). Noise robustness of the algorithm could be improved by pre-ordering a noise reduction block.**

I. INTRODUCTION

The need of accurate and noise robust pitch detection algorithm (PDA) is there since early 1970s, when first algorithms were described in literature. The initial motivation was the idea of automated music transcriptions. Nowadays, 40 years later, pitch detection still plays a key part in speech and music processing. As part of prosodic information, the pitch of speech helps speech recognizers to achieve better success-rates [1] in real-life application (e.g. automatic speech transcription with inserting of punctuation marks). It is also used in speech synthesis that made huge progress in past ten years. Also in tasks like speaker identification or real-time speech re-synthesis (voice conversion) is the use of decent PDA inevitable.

The paper is organized as follows: Section II presents brief overview of pitch detection algorithms and also touches the input signal pre-processing and possibilities for pitch contour smoothing. Section III brings the description of merged normalized backward-forward correlation (MNFBC) and its voicing decision capabilities. Section IV deals with post-processing with Viterbi algorithm. Reference database and evaluation criteria are described in Section V. Results and discussion can be found in Section VI.

II. PITCH DETECTION ALGORITHMS OVERVIEW

PDAs can be generally divided according to the domains where are computed.

In time domain (in which acoustical signals are initially represented and stored) exist basic methods trying to find local extreme of the waveform (peak picking). Main advantage of this approach is really low complexity. Auto-correlation (ACF) [2] is relatively noise robust time-domain algorithm formerly used with success in many pitch tracking application. Signal is compared with itself shifted by time lag that corresponds to tested fundamental frequency (F0). For pitch candidate we look for maximum of the function. But for rapidly changing F0 over the time window the main peak of function could disappear. The biggest shortcoming is also the need of two whole periods of signal to make reasonable estimate even for lowest F0 frequency. That is why cross-correlation function (CCF) has been developed with ability of reasonable estimate with window size of a single typical length glottal period. An average magnitude difference function (AMDF) [3] is a time-domain alternative to ACF with less computational requirements (difference is computationally cheaper than product). In contrast to ACF, minimum of the function is in the point of interest. Accurate pitch tracker called Direct Frequency Estimation (DFE) working in time-domain with really low computational requirements was presented in [4].

In frequency domain we use periodicity of spectrum, each F0 in real-world has its harmonics as natural multiplies of this fundamental frequency, although not all of them are present in most cases (in speech they are suppressed by formant frequencies of head cavities). Sub-harmonic summation (SHS) [5] is typical representative of this approach. Cepstral method [6] is inverse Fourier transform of short-time log magnitude spectrum.

Example of multi-domain complex algorithm can be found in [7]. The core method MNFBC operates in time-domain and is complemented by SHS. Overall PDA has been found to be noise robust and accurate.

Problem of voiced-unvoiced (VUV) decision is related to the usage of PDAs on speech signals. These consist of voiced and unvoiced parts according to whether glottis tend to vibrate when the speech is created (voiced parts) or not (unvoiced parts). Good VUV decision is crucial step in pitch detection. Some of pitch detection functions are able to directly make these decisions by simple thresholding their peaks, other need to gather more speech characteristics (zero crossing rate or energy) to do right decision. But best results are achieved by special VUV block, where for example

statistical approach can take place (for example parametrization as an input for artificial neural network [7]).

Pre-processing stage is aimed to remove interfering signal components. Often includes DC offset removal, noise removal (by spectral subtracting) and filtering the signal on both frequency borders. High pass (low cut) filter is usually set to value around 60Hz and is meant to suppress low frequency content from electrical network that could be present in recordings. It is also used for eliminating a low frequency noise such as room rumble, unwanted very low frequency components of speech or wind noise in outdoor applications. Low pass (set above the highest end of vocal range) can help to emphasis true pitch period by removing loss of periodicity at higher frequencies in voiced speech. But this at the same time destroys higher harmonic content of signal that could be important for frequency-domain methods such as SHS. In general any bandwidth limitation of signal increases correlation between samples and thus should be applied wisely when VUV decision is done by thresholding correlation peaks.

The main task for post-processing stage of PDAs is smoothing the final pitch contour that is often needed duet to an existence of out-layer values (octave errors or not precious estimate). These artifacts are in praxis eliminated by median filtering (usually 5-point) or dynamic programming approaches (Viterbi algorithm) finding the cheapest path. The latter method needs to have more F0 candidates in each step to be able to choose from them the most probable through-pass.

III. MERGED NORMALIZED FORWARD-BACKWARD CORRELATION

The MNFBC method [7] itself is improvement of well-known and for pitch analysis widely used normalized cross-correlation (NCFF) [8]. Main idea of MNFBC is taking one signal frame $x_w[n]$ with length of *4*MPP* (where *MPP* stands for Maximal Pitch Period as the reciprocal of the lowest F0) and computing forward normalized correlation (NFC) on first half-frame and backward normalized correlation (NBC) on the second half. In equation (1) is basic correlation term for one frame, which is used for expressing the computation of NFC[t] and NBC[t] in equations (2) and (3).

$$\langle x_{w_k}[n], x_{w_l}[n]\rangle = \sum_{n=0}^{2*MPP-1} x_w[n+k]\,x_w[n+l] \qquad (1)$$

$$NFC[t] = \frac{\langle x_{w_0}[n], x_{w_t}[n]\rangle}{\sqrt{\langle x_{w_0}[n], x_{w_0}[n]\rangle\langle x_{w_t}[n], x_{w_t}[n]\rangle}} \qquad (2)$$

$$NBC[t] = \frac{\langle x_{w_{2MPP}}[n], x_{w_{2MPP-t}}[n]\rangle}{\sqrt{\langle x_{w_{2MPP}}[n], x_{w_{2MPP}}[n]\rangle\langle x_{w_{2MPP-t}}[n], x_{w_{2MPP-t}}[n]\rangle}} \qquad (3)$$

In the Figure 1 can be seen typical course of two opposite normalized correlations on voiced frame. Peaks are not exactly on the same lag index for both halves of the processed frame saying there was slight shift in pitch period during the whole frame.

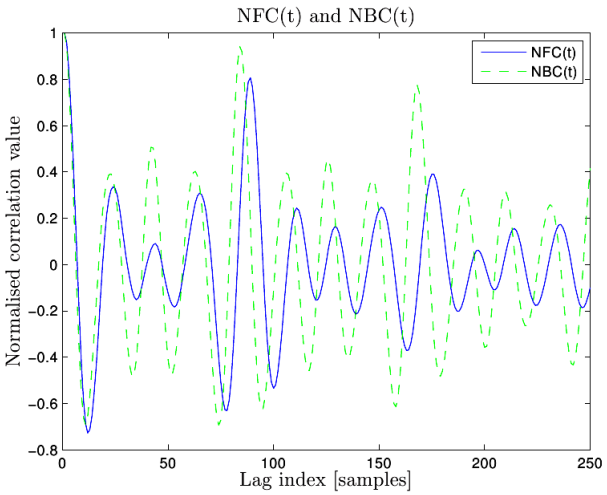

Figure 1: Courses of NFC and NBC functions on reference voiced frame

Both correlations are then half-rectified producing NFC'[t] and NBC'[t]. From these slightly modified correlations is then computed merged normalized backward-forward correlation MNFBC[t] for given frame according to equation (4). The function is depicted in the Figure 2. Second peak corresponds to first harmonic.

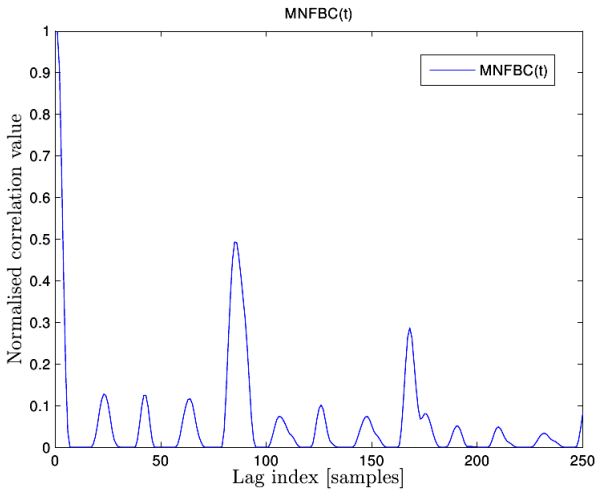

Figure 2: Course of final MNFBC function on reference voiced frame

The MNFBC[t] function should solve problems of weak periodicity of mixed-excitation regions. Also with prolonged frame, in contrast to standard length of two maximal periods, an estimation should be more accurate. On the other hand the length of frame designates the method to not to be used in real time-critical applications.

Generally the ability of voicing decision by thresholding correlation functions peaks is not best and this was unfortunately found to be true for MNFBC too (see result section for details). That is why few tweaks were suggested to improve this shortcoming: thresholding of sum of 3 best candidates, establishing third voicing state for unsure frames (that copies the decision from last sure frame) and additional thresholding by energy of actual frame. For energy computation, the frame was firstly pre-processed by short-time energy envelope to emphasize its period content [9].

$$MNFBC[t]=\frac{\langle x_{w_0}[n],x_{w_0}[n]\rangle (NFC'[t])^2+\langle x_{w_{2MPP}}[n],x_{w_{2MPP}}[n]\rangle (NBC'[t])^2}{\langle x_{w_0}[n],x_{w_0}[n]\rangle +\langle x_{w_{2MPP}}[n],x_{w_{2MPP}}[n]\rangle} \quad (4)$$

In general, amplitude of MNFBC function should not depend on the amplitude of input frame of signal, that is why main voicing threshold for MNFBC can be constant. But considering the energy of the frame, it is directly influenced by input acoustic pressure of speaker's voice that can vary in time even for same speaker. This leads to adapting the energy threshold by getting to know the energy range of current utterance (minimal and maximal energy dB level is corrected with each incoming frame).

Logic of voicing is than as follows: if the frame satisfies MNFBC threshold along with energy threshold, it is considered as voiced, otherwise unvoiced.

IV. VITERBI POST-PROCESSING

Viterbi algorithm [10] belongs to category of dynamic programming. It is used in many technical fields, but for years has been the core of speech recognition systems (Viterbi decoder). It finds the cheapest path through the multi-state space. When using it for final pitch contour smoothing, the PDA core method is needed to give more candidates per voiced frame. Three candidates are considered in this article. Each candidate k is then evaluated with an emission and transition probabilities b_k and a_{kl} . Having these probabilities the cumulative probabilities can be computed as new candidates in next steps are coming. Exact equations can be found for example in [7]. For getting good results it is the most important how instant probabilities are computed. Emission probability b_k is simply value of MNFBC function for candidate k. In this work transition probability a_{kl} (that the candidate k is followed by candidate l in next step) is derived from musical distance of candidates in semitones according to equation (5), where f_k is frequency estimate in Hz for candidate k.

$$a(k,l)=e^{-0.12*12|\log_2(\frac{f_k}{f_l})|} \quad (5)$$

V. REFERENCE DATABASE AND EVALUATION CRITERIA

A part of Spanish Speecon manually pitch-marked reference database [9] was used to compute VUV decision and precision success-rates. Results introduced in this paper are shown on its close talk channel 0 (highest SNR) and lavaliere microphone Channel 1 (lower SNR, more noise). Standard PDA evaluation criteria were used (voiced errors VE, unvoiced errors, UE, 20% tolerance gross errors GEH and GEL, see for example [7] for details) with addition to gross errors with only 10% tolerance (GEH10, GEL10) and octave errors statistic (doubling and halving errors DE and HE) [11].

VI. RESULTS AND DISCUSSION

Results comparing different possibilities for VUV decision are shown in Table 1. First variant V1 is pure thresholding the MNFBC peak on value 0.5. Second variant brings thresholding a sum of three highest peaks. V3 stands for "DK (don't know)" in specified range of MNFBC (this uncertain state is copying last certain VUV decision). In versions 4-6 a decision is made with help of utterance log energy range thresholding (second parameter). When considering optimum VUV decisions according to lowest VE+UE sum, this optimum parameter settings is somewhere between 0.27 and 0.3 for MNFBC voicing threshold and energy dB threshold of 1/3. It can be seen that by lowering the energy threshold more, the VE slightly decrease, but UE tend to raise a little and thus VE+UE sum can not be lowered more by this change of parameters. Same situations occurs when we try to lower more MNFBC peak threshold.

TABLE 1: VOICING DECISION PARAMETERS FOR MNFBC, CHANNEL 0

MNFBC VUV	VE [%]	UE [%]	VE+UE [%]
V1 (0.5 thr.)	22.0	12.7	24.7
V2 (3PS, 0.6 thr.)	16.0	25.6	41.6
V3 (0.55-0.6 DK state)	19.3	11.8	31.1
V4 (0.3 thr., 0.33E)	11.4	10.4	21.8
V5 (0.3 thr., 0.26E)	11.0	11.4	22.4
V6 (0.27 thr., 0.33E)	10.2	11.1	21.3

TABLE 2: SMOOTHING COMPARISON OF MNFBC OUTPUT

MNFBC smooth.	GEH [%]	GEL [%]	GEH10 [%]	GEL10 [%]	DE [%]	HE [%]
V0	0.50	2.41	1.80	3.50	0.05	2.00
V1	0.31	2.26	1.60	3.90	0.01	1.80
V2	0.42	0.99	1.80	2.10	0.04	0.70

Table 2 compares the smoothing possibilities – V0 for no smoothing, V1 is 5-point median filtering and V2 Viterbi algorithm with cent based transition probability function. As entry point to smoothing the MNFBC version with VUV peak thresholding at 0.5 was chosen. Median filtering can decrease high gross errors and doubling errors, but in GEL10 criterion is worse than V0 without any smoothing. Although Viterbi algorithm brings more demand on computational power, it is worth to compute it for significant eliminating gross errors low and halving errors.

Tables 3 and 4 bring overall results and compare MNFBC (with frame energy improved VUV decision and Viterbi traceback smoothing) to other PDAs mostly mentioned in Section II. They also show mean pitch difference Δ and standard deviation of the pitch difference σ in semitone cents [4]. PRAAT cc stands for implementation of cross-correlation method in phonetic tool PRAAT [12]. Abbreviation Kotnik2009 stands for complex method described in [7] and the partial results presented are taken from there. Except for the unvoiced errors (UE) criterion, MNFBC is comparable to

TABLE 3: PDA COMPARISONS - CHANNEL 0 OVERALL RESULTS

PDA	VE [%]	UE [%]	GEH [%]	GEL [%]	GEH10 [%]	GEL10 [%]	DE [%]	HE [%]	Δ [cents]	σ [cents]
MNFBC	10.20	11.10	0.85	1.80	2.50	3.60	0.10	1.10	-27	208
DFE	26.60	15.50	8.40	4.20	16.50	8.90	0.20	1.30	22	268
PRAAT CC	9.06	3.04	0.77	2.30	2.30	3.25	0.08	1.80	-13	225
KOTNIK2009	7.28	2.25	1.39	0.15	N/A	N/A	N/A	N/A	N/A	N/A

TABLE 4: PDA COMPARISONS - CHANNEL 1 OVERALL RESULTS

PDA	VE [%]	UE [%]	GEH [%]	GEL [%]	GEH10 [%]	GEL10 [%]	DE [%]	HE [%]	Δ [cents]	σ [cents]
MNFBC	24.00	9.50	14.70	1.80	16.30	3.30	1.90	1.10	283	844
DFE	45.40	11.10	8.50	8.10	17.90	13.10	0.05	4.30	-23	355
PRAAT CC	22.00	7.70	9.70	2.50	11.10	3.30	1.30	1.90	190	722
KOTNIK2009	9.28	5.48	1.62	3.17	N/A	N/A	N/A	N/A	N/A	N/A

PRAAT cc on channel 0 and in all criteria on channel 1. The advantage of neural network VUV decision block is remarkable on both channels for Kotnik2009 PDA.

VII. CONCLUSIONS

The usability of merged normalized backward-forward correlation (MNFBC) with Viterbi smoothing has been studied on speech signals in terms of voicing decision ability and F0 estimation accuracy. Results computed on pitch reference database show that as other correlation based methods it is not possible to achieve good voicing decision results without help of frame energy information. The benefit of Viterbi smoothing over median filtering has been also verified. Results also show that presented PDA with MNFBC as core method is generally comparable with PRAAT cc algorithm in VUV decision and also in accuracy. For low SNR signals noise reduction along with special VUV block could help to achieve better VUV and accuracy success-rate.

ACKNOWLEDGMENT

Research described in the paper was supported by the Czech Grant Agency under grant No. 102/08/0707 "Speech Recognition under Real-World Conditions" and grant No. 102/08/H008 "Analysis and modeling of biomedical and speech signals".

REFERENCES

[1] K. Vicsi and G. Szaszák, "Using prosody to improve automatic speech recognition," Speech Commun., vol. 52, pp. 413–426, May 2010. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2010.01.003

[2] L. Rabiner, "On the use of autocorrelation analysis for pitch detection," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 25, no. 1, pp. 24 – 33, Feb. 1977.

[3] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 22, no. 5, pp. 353 – 362, Oct. 1974.

[4] H. Bořil and P. Pollák, "Direct time domain fundamental frequency estimation of speech in noisy conditions," in Proceedings of EUSIPCO 2004 (European Signal Processing Conference, Vol.1), pp. 1003–1006, 2004.

[5] D. J. Hermes, "Measurement of pitch by subharmonic summation," The Journal of the Acoustical Society of America, vol. 83, no. 1, pp. 257–264, 1988. [Online]. Available: http://link.aip.org/link/?JAS/83/257/1

[6] A. M. Noll, "Cepstrum pitch determination," The Journal of the Acoustical Society of America, vol. 41, no. 2, pp. 293–309, 1967. [Online]. Available: http://link.aip.org/link/?JAS/41/293/1

[7] B. Kotnik and et al., "Noise robust f0 determination and epoch-marking algorithms," Signal Processing 89., pp. 2555–2569, 2009. [Online]. Available: DOI: 10.1016/j.sigpro.2009.04.017

[8] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," Speech Coding and Synthesis, Elsevier Science, pp. 495–518, 1995.

[9] B. Kotnik, H. Hoge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions," Proc. 3rd International Conference on Speech Prosody, Dresden, Germany, pp. 149–152, 2006.

[10] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," Information Theory, IEEE Transactions on, vol. 13, no. 2, pp. 260 – 269, Apr. 1967.

[11] J. Bartošek, "Pitch detection algorithm evaluation framework," 20th Czech – German Workshop on Speech Processing, Prague, pp. 118–123, 2010.

[12] Boersma, Paul & Weenink, David (2011). Praat: doing phonetics by computer [Computer program]. Version 5.2.21, retrieved 29 March 2011 from http://www.praat.org/