Evaluation of Pitch Detection Algorithms in Adverse Conditions

Bojan Kotnik[1], Harald Höge[2], Zdravko Kacic[1]

[1]University of Maribor, Slovenia
[2]Siemens AG, Corporate Technology, Germany
bojan.kotnik@uni-mb.si, harald.hoege@t-online.de, kacic@uni-mb.si

Abstract

Robust fundamental frequency estimation in adverse conditions is important in various speech processing applications. In this paper a new pitch detection algorithm (PDA) based on the autocorrelation of the Hilbert envelope of the LP residual [1] is compared to another well established algorithm from Goncharoff [2]. A set of evaluation criteria is collected on which the two PDA algorithms are compared. In order to evaluate the algorithms in adverse conditions a suited reference database was constructed. This reference database consists of parts of the SPEECON speech database [3] where recordings of 60 speakers were selected and manually pitch marked. The recordings cover several adverse conditions as noise in the car cabin and reverberations of office rooms. The evaluation highlights the good performance of the new algorithm in comparison but shows, that low SNR conditions and strong reverberation are still a demanding challenge for future pitch detection algorithms.

1. Introduction

In the past various pitch detection algorithms have been developed mainly for speech synthesis and speech coding [4, 5, 6]. Recently new applications as linguistic disambiguation in speech to speech translation systems [7], recognition of tonal languages [8], voice conversion [9] and speaker characterization [10] pup up leading to a growing demand of robust pitch detection algorithms (PDA) working in adverse conditions caused by noise and reverberation. Till now there was a lack in suited speech databases to test the performance of PDAs. In chapter 2 of this paper a new reference database is described which is suited to evaluate PDAs. The database is available for research use for ECESS [11] members. The next chapter describes the used evaluation criteria (chapter 3) on which the new PDA algorithm (chapter 4) is evaluated (chapter 5).

2. PDA evaluation reference database

The reference database constructed to evaluate PDAs consists of parts of the SPEECON speech database [3]. The acoustical environments found in this database comprise those of the car interior, the office, and living rooms. The office environment is mostly quiet, and slightly affected by stationary and white noises from computer fans or air-conditioning devices. However, in some of the offices the recordings contain also background voices. The living room recordings (entertainment environment) contain a wider range of noises, less stationary and more colored than the office noises. In some utterances, the radio or TV set is on; consequently, voices can be found in the recordings, as well as music, etc. The reverberations are mostly present in office and entertainment environments. The durations of the reverberation effect measured in the SPEECON rooms vary from 250 ms up to 1.2 seconds (T_{60} measure) [13]. The in-car recordings contain medium to high noise levels, which are of both stationary (engine) and instantaneous nature (wipers) [3]. Thus, the SPEECON database contains a large variety of distortions like additive noises, reverberations and channel distortions.

The database was recorded at 16 kHz sampling frequency and quantized using 16-bit linear coding. From this database the recordings of 60 speakers was selected (30 male + 30 female speakers, speaker age from 19 to 79 years). The signal was acquired simultaneously by four microphones; each microphone was placed at a different distance from the speaker: a head mounted close-talk microphone (recording channel C0), a Lavalier microphone (a microphone placed just below the chin of the speaker, channel C1), a directional microphone situated at 1 meter from the speaker (channel C2), and an omni-directional microphone situated at 2-3 meters from the speaker (channel C3). Note that in the car environment the C3 microphone was placed at the most distant front corner of the car interior. The SNR measured at the close-talk microphone is around 30dB, which indicates this microphone provides nearly clean speech signal. On the opposite side, the omni-directional microphone is strongly affected by the background noise, and therefore the corresponding speech signal has low SNR (around 0dBs). Similar SNR values were observed for the entertainment environment, except for the omni 2-3m microphone, where the entertainment environment SNR is about 4dB higher than the office environment SNR [13].

In order to manually construct the reference pitch-marked database under low noise conditions and without reverberation the close talking microphone recordings in the amount of 1 minute per speaker were selected. Thus the reference database comprises 60 minutes of pitch-marked speech signal. In the first step, the 60 minutes of selected close-talking channel C0 speech signal were automatically pitch-marked (epoch marked). In the next step accurate manual rechecking and correcting of pitch marks is performed thus resulting in reference pitch-marked database.

Due to the simultaneous recordings of the 4 channels C0-C3 the pitch marks of C0 could be transferred to the channels C1, C2, C3. Since there is a substantial physical distance between the microphones located at distant locations, there exists a time delay misalignment between the signals from different input channels. In order to perform the time delay alignment, the cross-correlation based algorithm is applied between the time reference signal from channel C0 and particular distant channel (C1-C3).

3. PDA evaluation criteria

A variety of the PDA evaluation criteria were already established in the literature. The gross error high and the gross error low were introduced in [14]. The voiced error, unvoiced error, the absolute difference between the mean values, and the absolute difference between the standard deviations of reference and estimated pitch were noticed first in [12]. The full set of definitions of evaluation criteria used are:

- **Gross error high (GEH) and gross error low (GEL)**

The gross error high (GEH) presents the percentage of voiced speech segments for which the detected pitch is more than 20% higher than the reference pitch (*Estimated_Pitch* > 1.2**Reference_Pitch*). The gross error low (GEL) presents the percentage of voiced speech segments for which the detected pitch is more than 20% lower than the reference pitch (*Estimated_Pitch* < 0.8**Reference_Pitch*).

- **Voiced error (VE) and unvoiced error (UE)**

The voiced error (VE) presents the percentage of voiced speech segments which are misclassified as unvoiced. The unvoiced error (UE) presents the percentage of unvoiced speech segments which are misclassified as voiced. Both, the VE and UE are used to evaluate the performance of the voiced/unvoiced detection stage of the PDA algorithm.

- **Absolute difference between the mean values (AbsMeanDiff)**

The absolute difference (in Hz) between the mean values of the reference pitch and the estimated pitch: *AbsMeanDiff* [Hz] = abs{ *MeanRefPitch*[Hz] – *MeanEstPitch*[Hz] }.

- **Absolute difference between the standard deviations (AbsStdDiff)**

The absolute difference (in Hz) between the standard deviations of reference pitch and the estimated pitch: *AbsStdDiff*[Hz] = abs{ StdRef[Hz] – StdEst[Hz] }. The mentioned mean values and standard deviations are computed on whole reference and estimated F0 data respectively.

4. A new PDA algorithm

The proposed pitch determination procedure was inspired by [1]. Since the PDA is initially developed as a fundamental frequency determination stage for the pitch marking algorithm (PMA), only the PDA processing blocks of the complete PMA is presented in Figure 1. The related processing steps are described in the following subsections.

4.1. Speech signal preprocessing

Prior to the PDA analysis, the signal $x[n]$ is segmented into overlapping frames with the frame length of N = 780 samples (48 ms) and with frame shift interval of S = 16 samples (1 ms). The sampling frequency of f_S = 16 kHz is presumed throughout the paper. Through the following sections the frame index will be denoted by m, whereas n will present the sample index. In order to enhance the periodic structure of voiced segments of the input speech signal, the short time energy contour $e[n]$ of the input speech signal is multiplied with the input signal, thus producing signal $x_2[n]$. As presented in [2], the short time energy contour is produced by low-pass filtering of the squared speech signal $(x[n])^2$. This is accomplished by convolving the squared speech signal with a smoothing window $w_{\mathrm{SMOOTH}}[n]$ that is found as the convolution of the Hann window $w_{\mathrm{Hann}}[n]$ with itself:

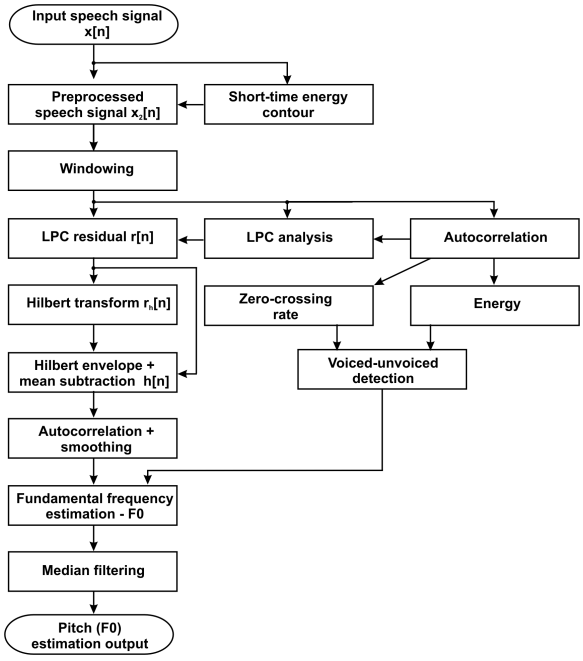

Figure 1: Schematic diagram of the proposed PDA.

$$\begin{aligned} \mathbf{w}_{SMOOTH} &= \mathbf{w}_{Hann} * \mathbf{w}_{Hann} \\ \mathbf{e} &= \mathbf{w}_{SMOOTH} * \mathbf{x}^2 \\ \mathbf{x}_2 &= \mathbf{x} \cdot \mathbf{e} \end{aligned} \quad . \qquad (1)$$

For a sampling frequency chosen the following parameters are used:

- Length of the Hann window ($\mathbf{W}_{\mathrm{Hann}}$): 54 samples,
- Length of the smoothing window ($\mathbf{W}_{\mathrm{Smooth}}$): 107 samples

4.2. LPC analysis

Prior to the LPC analysis each overlapped frame m is multiplied with the Hann window (780 samples) thus producing overlapped windowed signal $x_2[n,m]$. Next, the LPC analysis of each frame m of the $x_2[n,m]$ using autocorrelation principle is performed. The order of the LPC analysis is set to p = 16. The LPC residual $r[n,m]$ is computed as follows:

$$r[n,m] = x_2[n,m] + \sum_{k=1}^{p} a_k x_2[n-k,m], \qquad (2)$$

where a_k are the LPC coefficients computed using Levinson-Durbin recursion.

4.3. Hilbert envelope computation and autocorrelation of the HE

The Hilbert transformation is defined by the transfer function $H(\omega) = -j\, sign(\omega)$. Denoting the Hilbert transform of the LPC residual by $r_h[n,m]$, then the Hilbert envelope $h[n,m]$ of the linear prediction residual is computed as follows [4]:

$$h[n,m] = \sqrt{r_h^2[n,m] + r^2[n,m]} \ . \qquad (3)$$

Afterwards, the mean value of the $h[n,m]$ is subtracted from $h[n,m]$, and finally the autocorrelation of the mean-subtracted

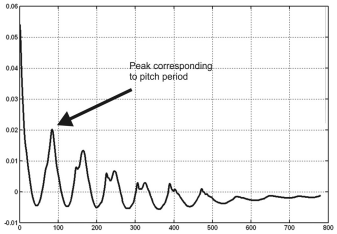

Figure 2: Right-hand side of the mean subtracted Hilbert envelope of the linear prediction residual.

Hilbert envelope $h[n,m]$ is computed. The lag $T0[m]$ of the first peak on the right-hand side of the autocorrelation function of the Hilbert envelope corresponds to the pitch period (see Figure 2).

4.4. Voiced/unvoiced

A voiced/unvoiced detector (V/UV) based on zero-crossing rate and energy of the preprocessed signal autocorrelation $x_{\mathrm{ACORR}}[n,m]$ is applied in order to prevent false computation of *F0* in unvoiced or noise-only region of the input speech signal. The V/UV is based on the *energy-to-zero crossing rate* ratio $EZR[m]$ which is computed as follows:

$$EZR[m] = \frac{\overline{E}[m]}{ZCR[m]}\ , \qquad (4)$$

where $ZCR[m]$ presents the zero crossing rate of the frame m and $\overline{E}[m]$ is time-smoothed energy of the $x_{\mathrm{ACORR}}[n,m]$. The presented parameter $EZR[m]$ is applied as voiced/unvoiced decision criterion. Namely, there is in the voiced regions of the speech the signal energy relatively high and zero-crossing rate relatively low. The $EZR[m]$ will therefore be relatively high. The opposite is true in the unvoiced speech regions, where the signal energy is usually low and the zero-crossing rate is high. The $EZR[m]$ will be therefore relatively low in the unvoiced regions of the speech signal. Voiced/unvoiced decision ($VUV[m]$) is estimated by comparison of $EZR[m]$ with some database-dependant threshold ϑ:

$$\begin{aligned} &\textit{If}\quad EZR[m] > \vartheta \quad \textit{Then} \\ &\quad VUV[m] = 1 \\ &\textit{Else} \\ &\quad VUV[m] = 0 \end{aligned}\ . \qquad (5)$$

The optimal value of the threshold ϑ is determined with parameter determination subset of the speech database.

4.5. F0 estimation and median filtering

The peak picking interval for the $T0[m]$ (fundamental period of the voiced speech) estimation is limited in the region from the lag 32 (equal to 500 Hz) to the lag 250 (equal to 64 Hz). The fundamental speech frequency (*F0* in Hz) of the frame m ($F0[m]$) is then computed with the following equation:

$$F0[m] = \frac{f_s}{T0[m]} \quad [Hz]\ . \qquad (6)$$

In mixed-excitation regions of the speech signal some $F0[m]$ determination errors may occur. Especially *F0* doubling or halving errors are most frequent. In order to compensate the effect of these errors, the median filtering of the *F0* estimates is applied for each frame m as follows:

$$F0[m] = \underset{k=m-M}{\overset{k=m+M}{Median}}\{F0[k]\}\ , \qquad (7)$$

where M represents the width of the median filtering interval in frames. In proposed PDA procedure the value of $M = 4$ is applied. The filtering operation produces final F0 estimates.

5. PDA performance evaluation discussion

For the evaluation of PDA algorithms the new algorithm described in chapter 4 is compared with the algorithm of V. Goncharoff [2]. Table 1 presents the overall PDA evaluation performance achieved with above mentioned algorithms. The best PDA performance of the proposed algorithm is achieved with C0 channel of the SPEECON database. 4.48 % of voiced frames have the estimated pitch more than 20 % higher than the reference (GEH), whereas 0.30 % of voiced frames have the estimated pitch more than 20 % lower than the reference pitch of corresponding frames (GEL). The absolute mean difference of 3.27 Hz was observed between the reference and estimated pitches. The achieved VE of 0 % at C0 means that none of the voiced frames was misclassified as unvoiced. However, 1.47 % of unvoiced frames were misclassified as voiced. The performance of the proposed PDA deteriorates with decreasing SNR conditions (SPEECON channels C2, C3). The most prominent error is observed with voiced/unvoiced detection performance. The most problematic was found to be the voiced error (VE). Namely, with decreasing conditions more and more voiced frames are misclassified as unvoiced by the proposed voiced/unvoiced detector. The UE error is increased with channels C1, C2, and C3 but remains more or less constant. The fundamental frequency estimation also becomes less accurate with noisier channels of the SPEECON database (see *AbsMeanDiff*).

The PDA evaluation results using Goncharoff's algorithm are presented in the right-hand side of the Table 1. The Goncharoff's algorithm tends to produce higher *GEL* than *GEH*, which is the reversed behavior as with the proposed PDA. With Goncharoff's algorithm, the VE tend to steeply rise at adverse SNR conditions. Moreover, the sum of voicing errors (*VE+UE*), and the sum of gross errors (GEL + GEH) within particular SPEECON channel are in most cases higher with the Goncharoff's algorithm than with the proposed PDA procedure. However, the *AbsMeanDiff* and *AbsStdDiff* are a bit lower with Goncharoff's algorithm than with proposed PDA. The anomaly observed with GEL and GEH at C3 are probably due to the specific characteristic of the omni-directional microphone used to capture the signal of the channel 3.

Table 2 presents the SPEECON F0 performance evaluation of the proposed PDA at channels C2 and C3 for different environments. It can be observed that the performance of the presented PDA varies not only from channel to channel but also between different environmental conditions. In order to be able to compare different numbers, the PDA performance will be discussed in the form of the sum GEL+GEH. The best PDA performance of the channel C1 is achieved in the car environment (GEL+GEH = 3.95 %), while in the other two remaining environments (office, entertainment) quite similar (9.73 % and 10.35 % respectively) performance is achieved. The reason for such performance is in the nature and in the SNR of the environmental noise. The noise levels are in the car environments higher than in the

Table 1: *Overall F0 estimation performance evaluation results of the proposed (left) and Goncharoff's (right) PDA.*

	Proposed PDA				Goncharoff's PDA			
Speecon Condition → PDA eval. Criteria ↓	**C0**	**C1**	**C2**	**C3**	**C0**	**C1**	**C2**	**C3**
GEL (%)	4.48	7.36	14.98	8.84	3.93	2.54	7.31	2.67
GEH (%)	0.30	1.51	0.72	0.74	8.22	8.66	22.19	9.06
VE (%)	0.00	17.80	41.91	66.51	1.40	25.83	32.83	75.60
UE (%)	1.47	14.89	14.17	11.67	0.00	9.91	18.48	6.33
AbsMeanDiff (Hz)	3.27	16.08	38.17	42.68	5.99	15.24	30.65	35.92
AbsStdDiff (Hz)	1.26	11.97	11.53	13.80	3.59	6.41	8.91	10.40

Table 2: *F0 estimation performance evaluation results of the proposed PDA in different SPEECON environments.*

Environment	**Car**		**Office**		**Entertainment**	
Speecon Condition → PDA eval. Criteria ↓	**C1**	**C2**	**C1**	**C2**	**C1**	**C2**
GEL+ GEH (%)	3.95	30.29	9.73	16.13	10.35	9.27
VE (%)	52.93	4.74	3.72	35.57	12.48	61.75
UE (%)	2.73	20.20	28.37	12.23	24.18	7.25
AbsMeanDiff (Hz)	15.13	57.36	12.27	33.90	15.71	23.54
AbsStdDiff (Hz)	31.66	22.22	7.55	1.15	10.64	5.01

other two environments (this resulted in poor V/UV performance in the car environment). However, the colored noises of the office and entertainment environment consist also of human voices (speech in the background) which can easily degrade performance of the PDA. Strong decrease in the PDA performance is observed with C2 in the car environment (GEL+GEH = 30.29 %), while the PDA performance in the entertainment environment remains similar than in C1. The PDA performance in the office environment also decreases in C2 (GEL+GEH = 16.13 %). The V/UV performance is decreased within the office and entertainment environment, while it improves in the car environment.

6. Conclusion

In the presented paper the problems of pitch detection in adverse environments are described. A PDA algorithm based on Hilbert envelope is described and compared to the PDA proposed by Goncharoff [2]. A pitch-marked reference database based on Spanish SPEECON speech database was constructed and used in the PDA evaluation procedure. The F0 estimation results show that with the proposed PDA better pitch detection performance in adverse conditions is achieved when compared to the Goncharoff's algorithm. However, strong SNR dependence of the F0 estimation performance is still observed. In order to improve the F0 estimation performance more robust voiced/unvoiced detectors along with a denoising stage should be integrated.

7. Acknowledgement

We thank all partners of the ECESS [11] consortium, who participated on the discussion of epoch detection. This work was partly funded by the European Union under the Integrated Project 'TC-STAR – Technology and Corpora for Speech to Speech Translation' (IST-2002-FP6-506738, http://www.tc-star.org).

8. References

[1] Mahadeva, S. R., and Yegnanarayana, B., 2004. Extraction of Pitch in Adverse Conditions. *Proc. ICASSP 2004*.

[2] Goncharoff, V., and Gries, P., 1998. An Algorithm for Accurately Marking Pitch Pulses in Speech Signals. *IASTED International conference SIP '98*, Nevada, USA.

[3] Iskra, D. J. et al., 2002. SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation. *Proc. LREC'2002*.

[4] Hess, W. , 1983. Pitch Determination of Speech Signals. Berlin, Germany: Springer Verlag

[5] Markel, J. D., 1972. The SIFT algorithm for fundamental frequency estimation. *IEEE Trans. Audio Electroacoust.* 20,357-377.

[6] Ertan, A. E., Barnwell, T.P. 2005. Improving the 2,4KB/s Military Standard MELP (MS-MELP) Coder using Pitch-Synchronous Analysis and Synthesis Techniques. IEEE Proc. *ICASSP2005*

[7] Harbeck, S., Kießling, A., Kompe, R., Niemmann, H, Nöth, E. 1995. Robust pitch period detection using dynamic programming with an ANN cost function. *Proc. EUROSPEECH* Madrid 2,1337-1340.

[8] Li Ming, Yu Tiecheng, 2001. Robust and Efficient Pitch Tracking Method for Tonal Feature Extraction. IEEE Proc. *ICASSP2001*

[9] Sündermann, D., Bonafonte, A., Ney, H., Höge, H. 2004. A First Step Towards Text-Independent Voice Conversion. *Proc. ASRU2004*

[10] Kim, S., Eriksson, T., Kang, H.-G., and Youn, D. H., 2004. Pitch Synchronous Feature Extraction Method for Speaker Recognition. *Proc. ICASSP 2004*.

[11] www.ecess.org

[12] Martino, J., Yves, L., 1999. An Efficient F0 Determination Algorithm Based on the Implicit Calculation of the Autocorrelation of the Temporal Excitation Signal. *Proc. EUROSPEECH'99*. Budapest, Hungary.

[13] Pujol, P., Nadeu, C., Macho, D., Padrell, J., 2004. Speech Recognition Experiments with the SPEECON Database Using Several Robust Front-Ends. *Proc. ICSLP*, 2004.

[14] Ying, G., Jamieson, H., Mitchell, C., 1996. A Probabilistic Approach to AMDF Pitch Detection. *Proc. ICSLP 1996*, Philadelphia, PA.