EVALUATION OF PITCH DETECTION ALGORITHMS UNDER REAL CONDITIONS

Iker Luengo, Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, Jon Sanchez, Iñaki Sainz
{iker, ibon, eva, inma, ion, inaki}@bips.bi.ehu.es

Aholab-Signal Processing Laboratory – Faculty of Engineering
University of the Basque Country
Urkijo zum. z/g 48013. Bilbao-Spain.

ABSTRACT

A novel algorithm based on classical cepstrum calculation followed by dynamic programming is presented in this paper. The algorithm has been evaluated with a 60-minutes database containing 60 speakers and different recording conditions and environments. A second reference database has also been used. In addition, the performance of four popular PDA algorithms has been evaluated with the same databases. The results prove the good performance of the described algorithm in noisy conditions. Furthermore, the paper is a first initiative to perform an evaluation of widely used PDA algorithms over an extensive and realistic database.

Index Terms— Speech analysis, Pitch detection

1. INTRODUCTION

Pitch detection and marking is a recurrent topic in published papers inside the speech research community. The interest arises naturally from the enormous range of applications and technologies that need and use a pitch detection algorithm (PDA). Precise calculation of the fundamental frequency in the speech signal has demonstrated to be a basic task in almost all areas of speech research, from traditional areas such as speech coding to more recent areas of research like novel speech synthesis techniques or speaker emotional state characterization.

Improving on the first proposed methods based on the periodicity of the speech spectrum at voiced segments [1], a great variety of algorithms have been proposed (see [2] for a revision on classic methods). Some of them are very popular, either because they are publicly available or because they come packaged with some software analysis tool [3][4][5][6]. Considering that many users of these packages are not necessarily part of the speech research community (linguists, educators, speech therapists…), setting references and standards for the evaluation of their quality becomes a necessary task.

The perfect pitch detector should perform well under any reasonable noise or bandwidth condition. In that respect, several robust pitch detection algorithms have been proposed and claim to perform well under different noise conditions [7][8].

However, up to now no work has been published describing the evaluation of any algorithm with a significant amount of data. Some papers describe the performance of the algorithm only in a qualitative way, using a reduced set of signals and speakers. Others use ad-hoc small to medium size databases (some minutes) with very few speakers (2 to 5). In the last years, two speech databases have been used as reference for evaluation, mainly due to their public availability: the CSTR database and the Keele Pitch Reference database [3][5][9][10][11][12]. The first is about 5 minutes long and contains data from two speakers [9]. The second is about 10 minutes long with speech from five males, five females and five children [13].

This paper presents a novel pitch detection algorithm based on a classic representation (the cepstrum coefficients) followed by dynamic programming. We also present its evaluation comparing its performance with four other popular algorithms, using the CSTR database and a 60 minutes long database recorded by 60 speakers in four different recording channels [14].

Section 2 of this paper is dedicated to the description of the cepstrum-based detection algorithm and the conditions used for the selection of the best path using dynamic programming. Section 3 presents the performed experiments and the results. Conclusions are drawn in Section 4.

2. CEPSTRUM BASED PITCH DETECTION ALGORITHM

The proposed algorithm, called CDP, is based on cepstrum calculation followed by a dynamic programming module. After windowing the input signal, a set of pitch candidates is generated. This set is used by the dynamic programming algorithm to select the best pitch curve. As final step this curve is smoothed.

2.1. Selection of F0 candidates

The input signal is windowed by means of a Hamming window 58ms long. The length of the window has been chosen as to account for at least two periods in the minimum pitch case. Pitch values in the range of [35Hz-500Hz] have been considered. This range also applies to the selection of the cepstrum coefficients, in such a way that only coefficients with indexes included in [i_{max}, i_{min}] range will be considered:

$$i_{\max} = \left\lfloor \frac{F_s}{f_{\max}} \right\rfloor \qquad i_{\min} = \left\lfloor \frac{F_s}{f_{\min}} \right\rfloor \tag{1}$$

with F_s being the sampling frequency. The indexes are calculated as the closest integer to the bracketed expression.

Before proceeding to the search of the maximum coefficient whose index is supposed to give the pitch value, the coefficients c_i are normalized to the mean value inside the considered frame, giving the normalized coefficient c'_i:

$$c'_i = \frac{c_i}{\bar{c}} \tag{2}$$

$$\bar{c} = \frac{1}{i_{\max} - i_{\min} + 1} \sum_{i=i_{\min}}^{i_{\max}} c_i \quad (3)$$

This normalization offers a more coherent scale throughout the signal than the coefficients themselves, making the values more independent to changes in the signal, such as its energy.

A set of M+1 candidates for the final pitch value is generated selecting the biggest M normalized cepstrum coefficients. The last candidate is a "No-Pitch" value used to allow the Viterbi algorithm to decide for an unvoiced frame. In this way, no decision upon voicing is made in this module, letting the final decision to the Viterbi algorithm. The M+1st candidate will be called the *unvoiced candidate*.

2.2. Viterbi algorithm

Once the set of M+1 candidates is available for each frame the values that will build the pitch curve must be selected. This is done by using dynamic programming, selecting those that belong to the minimum cost curve. As usual, the selection cost has two components: a local cost C_l, associated to the selected candidate and independent of the neighbouring candidates; and a transition cost C_t, that considers the previous candidate. Four classical criteria are used to define the cost values:

1. The most probable candidate is the index of the maximum normalized cepstrum coefficient.
2. If the maximum value is small (smaller than a certain threshold), the frame is probably unvoiced.
3. There should not be sudden changes in the pitch curve. If there is one such sudden change, it is probably because it is an harmonic or a sub-harmonic of the real F0 value.
4. It is very unlikely that there is a fast voiced-unvoiced-voiced transition. An isolated unvoiced frame is probably due to noise and can be considered a detection error.

The defined local cost for the *m*-th coefficient in the *j*-th frame is calculated as the sum of two terms:

$$C_{l,m}(j) = -w_V C_{V,m}(j) + w_{thr} C_{thr,m}(j) \quad (4)$$

where

$$C_{V,m}(j) = \begin{cases} \log\left(c'_m(j)\right) & m = 1 \ldots M \\ 0 & m = M+1 \end{cases} \quad (5)$$

$$C_{thr,m}(j) = \underset{m=1\ldots M}{\begin{cases} 0 & c'_m(j) > T \\ 1 & else \end{cases}} \quad (6a)$$

$$C_{thr,M+1}(j) = \begin{cases} 0 & c'_m(j) < T \quad m = 1 \ldots M \\ 1 & else \end{cases} \quad (6b)$$

The first term in (4) is related to voiced candidates, assigning a cost reversely proportional to the logarithm of the normalized cepstrum coefficient. We assume here that the most probable candidate is the index of the maximum normalized cepstrum coefficient (criterion 1). This term will be zero for unvoiced candidates.

The second term in (4) is related to the cost of exceeding the voiced-unvoiced decision threshold *T*. For voiced candidates, it will take the value "1" only if the coefficient is below the pre-established threshold (see eq. (6a)). For unvoiced candidates, this cost will take the value "1" if there is at least one cepstrum coefficient over the threshold and "0" otherwise (see eq. (6b)).

The transition cost is calculated according to the following expression:

$$C_{t,m}(j, j-1) = \begin{cases} w_{cont} \left| \log\left(\frac{f_j}{f_{j-1}} \right) \right| & V - V \\ w_{VUV} & V\text{-}U \text{ or } U\text{-}V \\ 0 & U\text{-}U \end{cases} \quad (7)$$

where f_j is the fundamental frequency candidate for frame j.

As can be seen in (7), the frame transition between voiced candidates is favoured (low cost) if their corresponding frequencies are close. Transition cost from voiced candidates to unvoiced candidates and vice versa is constant. Finally, unvoiced-unvoiced transitions are assigned cost zero.

All the weights (w_V, w_{thr}, w_{cont}, w_{VUV}) and thresholds (*T*) have been set empirically and remain constant for all the experiments presented in section 3. Specific tuning of these values for different speakers or noise conditions can improve the results.

The result of this module is an intermediate pitch curve that will be post-processed in the next module by nonlinear smoothing.

2.3. Post-processing

The fundamental frequency of a speaker can be modelled by a log-normal distribution [15]. Detected pitch values too far away from the mean of that distribution can be considered as errors, and the farther away from that mean they are, the bigger the probability that an error has been made is.

A threshold T_L has been defined to decide for out-of-range F0 values. This threshold is calculated as L times the standard deviation of the log-normal distribution for the whole signal. L can be non-integer and values from 2.5 to 3.5 produce good results. A more conservative value of 2.5 has been chosen in the experiments presented in this paper. For other speech databases (such as expressive speech databases) a higher value would be more convenient.

The smoothing module checks each F0 value of the output curve from the Viterbi module to know whether that value is out-of-range. However, being out of range might not always be an error: if that out-of-range value belongs to a voiced segment, it will be declared unvoiced only if the whole segment lies out of range.

3. EXPERIMENTS

3.1. Evaluation databases

To evaluate the behaviour of the algorithm under noisy conditions, a subset of SPEECON Spanish database was used. SPEECON databases [14] sought to get speech signals in different acoustic environments mainly for speech recognition. Signals were acquired by four channels simultaneously, using different microphones and in different locations like cars, offices, public places, etc. The first channel, C0, was recorded with a close-talk microphone. C1 was recorded with a Lavalier microphone; C2 with a directional microphone 1 meter away from the speaker, and C3 with an omnidirectional microphone 2-3 meters away from the speaker.

Obviously, the signal to noise ratios of each channel are very different. Recordings of the C0 channel have SNR of about 30 dB, while, at the other end, the C3 channel has lower ratios, around 0 dB.

The subset of the SPEECON [16] used in the experiments contains sentences from 60 speakers, 30 male and 30 female. Each speaker recorded 1 minute of speech, adding up 60 minutes per channel. Reference pitch values have been obtained automatically and thoroughly revised manually, in the way described in [16].

The algorithm has also been evaluated against the CSTR database [9], a classical PDA evaluation reference database. This database contains five minutes of speech from one male and one female speaker. The sentences were specifically chosen to test pitch detection algorithms, thus including utterances containing voiced fricatives, nasals, liquids and glides, since PDAs generally find these difficult to analyse. The reference pitch values estimated from simultaneously recorded EGG are also provided.

For both databases the evaluation experiment has been the same: reference pitch values have been obtained at 1 ms rate and every value has been compared with its counterpart in the estimated pitch curve. The high sampling rate of the pitch curves to be compared poses demanding requirements in the accuracy of the algorithms.

3.2. Pitch estimation methods and evaluation criteria

Comparative data is necessary to assess the actual performance of the pitch detection module; therefore another four renowned algorithms were chosen to be evaluated along with our CDP algorithm. These algorithms were:

1. AM: Praat[1] PDA based on accurate autocorrelation method [4].
2. SRPD: Edinburgh Speech Tool Library's[2] implementation of super resolution pitch determinator [9].
3. SHR: PDA based on subharmonic to harmonic ratio, in an implementation described in [3].
4. RAPT: The KTH's WaveSurfer/ESPS implementation[3] of a robust algorithm for pitch tracking [6], a method based on normalized cross-correlation and dynamic programming.

In all cases, default values were used for the parameters, except for the F0 range. The accuracy of the different pitch estimation methods was measured according to the following criteria:

1. Classification Error (CE): it is the percentage of unvoiced frames classified as voiced and voiced frames classified as unvoiced.
2. Gross Error (GE): percentage of voiced frames with an estimated F0 value that deviates from the reference value more than 20%.
3. Mean (M): mean of the absolute value of the difference between the estimated and the reference pitch curve (Gross Errors are not considered for this calculation).
4. Standard Deviation (SD): standard deviation of the absolute value of the difference between the estimated and the reference pitch curve (Gross errors are not considered).

3.4. Results

Results of the evaluation with the SPEECON database are shown in tables 1 to 4, grouping the results by channel. With regard to the classification error (CE) for the channel 0 with low noise levels, correlation based algorithms get the best results. Our frequency domain based classification lies in the middle of the range. Predictably, performance of correlation based classification drops substantially as noise increases, as occurs in C1, C2, C3. This fall is not so sharp for our algorithm which gets the best results for these three channels.

With respect to the Gross Error measure, CDP clearly outperforms the rest of the algorithms, remaining robust even in the strong noise conditions of C3.

Finally, regarding the statistical features (M and SD) of the committed error, the SRPD algorithm gets the best results for almost every channel, although RAPT and CDP produce also close values. It is worth noticing that these magnitudes are little affected by noise due to the fact that gross errors are excluded from their calculation.

Method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	18.83	**0.83**	2.86	4.36
AM	**12.20**	5.88	8.74	7.93
RAPT	16.47	2.60	2.46	3.53
SHR	24.25	4.70	4.64	5.24
SRPD	18.13	3.11	**2.26**	**3.47**

Table 1. *Comparison of errors for C0 channel (SPEECON db.)*

method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	**28.64**	**0.72**	2.66	4.13
AM	32.56	15.57	9.43	8.70
RAPT	36.22	2.68	2.41	**3.48**
SHR	42.06	4.47	4.55	5.20
SRPD	45.18	4.04	**2.29**	**3.48**

Table 2. *Comparison of errors for C1 channel (SPEECON db.)*

method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	**35.32**	**0.83**	2.98	4.60
AM	37.12	13.93	8.72	8.54
RAPT	36.15	3.50	2.32	**3.49**
SHR	41.44	7.13	4.23	5.04
SRPD	53.62	4.42	**2.29**	3.58

Table 3. *Comparison of errors for C2 channel (SPEECON db.)*

method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	**43.20**	**1.19**	3.47	5.46
AM	52.93	17.48	10.04	9.70
RAPT	54.65	5.06	**3.09**	**4.59**
SHR	63.24	6.53	4.60	5.47
SRPD	72.57	5.60	3.36	5.21

Table 4. *Comparison of errors for C3 channel (SPEECON db.)*

The second set of evaluation experiments was carried out employing the CSTR database, aiming to test our algorithm's performance against a recognized benchmark. The results - presented in tables 5 to 7 - are coherent with the ones obtained in the tests with the equivalent cleanest channel C0 of the SPEECON database. The AM algorithm gets the best results in voiced/unvoiced classification; CDP is the best in the overall gross errors although it is closely followed by RAPT and SRPD. This last algorithm gets also the best results in the statistical magnitudes.

[1] www.praat.org

[2] www.cstr.ed.ac.uk

[3] www.speech.kth.se/wavesurfer

Method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	20.53	0.85	2.22	3.27
AM	**12.98**	18.43	10.25	7.13
RAPT	16.83	0.90	2.24	3.01
SHR	29.28	2.80	3.96	4.04
SRPD	20.82	**0.78**	**1.22**	**1.60**

Table 5. *Comparison of errors for male speakers (CSTR db.)*

Method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	14.86	0.42	3.97	5.53
AM	**10.53**	3.96	16.10	12.14
RAPT	12.84	**0.39**	3.90	5.00
SHR	27.20	1.60	6.31	6.60
SRPD	13.30	0.56	**3.34**	**4.15**

Table 6. *Comparison of errors for female speakers (CSTR db.)*

Method	*CE(%)*	*GE(%)*	*M(Hz)*	*SD(Hz)*
CDP	17.04	**0.62**	3.14	4.68
AM	**11.71**	10.74	13.67	10.75
RAPT	14.45	0.63	3.12	4.27
SHR	27.71	2.16	5.17	5.64
SRPD	16.74	0.66	**2.39**	**3.42**

Table 7. *Comparison of errors in the whole CSTR database*

4. CONCLUSIONS

A novel pitch detection algorithm based on classical cepstrum calculation followed by dynamic programming has been presented. To evaluate its performance, a 60-minute database with 60 speakers and containing a great variety of recording conditions has been used. As second evaluation the CSTR reference database was used. Results prove that cepstrum coefficients are superior in detecting pitch even in noisy conditions, provided a careful tracking of the pitch value is performed. The described algorithm has produced the best results for every channel concerning Gross Errors. Further improvements can be achieved by using more information for voiced/unvoiced detection in the clean speech case. On the other hand, evaluation of the algorithms for noisy speech should consider realistic and extensive databases to allow reliable comparisons. This work is a first effort to provide such reference.

5. ACKNOWLEDGEMENTS

The authors want to thank the members of the ECESS (http://www.ecess.eu) Consortium for granting the use of the subset of the SPEECON database and the reference pitch marking information. We also want to acknowledge the free use of CSTR database.

This work was partly financed by the Spanish Ministry of Education (TIC 2003-08382-C05-03) and the University of the Basque Country through a post-graduate grant.

6. REFERENCES

[1] A.M. Noll, "Cepstrum Pitch Determination". *The Journal of the Acoustical Society of America*, vol. 41, Issue 2, pp. 293-309, 1967.

[2] W. Hess, *Pitch determination of speech signals: algorithms and devices*. Springer-Verlag, Berlin, 1983.

[3] X. Sun, "Pitch Determination and Voice Quality Analysis Using Subharmonic-To-Harmonic Ratio". *Proc. ICASSP 2002*, Orlando, USA, vol. 1, pp. 333-336, 2002.

[4] P. Boersma, "Accurate short term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound". *Procs. Institute of Phonetic Sciences 17*. Univ. Amsterdam. pp. 97-110. 1993.

[5] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Trans. ASSP*, vol 39, pp. 40-48, 1991.

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

[7] T. Abe, T. Kobayashi and S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency". *Proc. ICSLP'96*, Philadelphia, USA, vol.2. pp 1277-1280, 1996.

[8] H. Quast, O. Schreiner and M.R. Schroeder, "Robust Pitch Tracking in the Car Environment". *Proc. ICASSP 2002*, Orlando, USA, vol. 1. pp. 353-356, 2002.

[9] P.C. Bagshaw, S.M. Hiller and M.A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching". *Proc. Eurospeech'93*, Berlin, Germany, pp. 1003–1006, 1993.

[10] Ch. Wang and S. Seneff, "Robust Pitch Tracking For Prosodic Modelling In Telephone Speech". *Proc. ICASSP 2000*, Istanbul, Turkey, pp. 1143-1146, 2000.

[11] P. Dikshit, "An Algorithm For Locating Fundamental Frequency (F0) Markers In Speech. *Proc. ICASSP 2005*, Philadelphia, USA, vol. 1, pp. 233-236, 2005.

[12] G.S. Ying, L.H. Jamieson and C.D. Mitchell, "A Probabilistic Approach to AMDF Pitch Detection". *Proc. ICSLP'96*, Philadelphia, USA, pp. 1201-1204, 1996.

[13] F. Plante, G.F. Meyer and W.A Ainsworth, "A pitch extraction reference database", *Proc. Eurospeech'95*, Madrid, Spain, pp. 837-840, 1995.

[14] D. Iskra, B. Grosskopf, K. Marasek, H. van den Huevel, F. Diehl and A. Kiessling, "SPEECON-Speech Databases for Consumer Devices: Database Specification and Validation". *Proc. LREC'02*. Las Palmas de Gran Canaria, Spain, pp. 329-333, 2002.

[15] M.K. Sonmez, L. Heck, M. Weintraub, E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition". *Proc. EUROSPEECH '97*. Rhodes, Greece, vol. 3, pp. 1391-1394, 1997.

[16] B. Kotnik, H. Höge, and Z. Kacic, "Evaluation of Pitch Detection Algorithms in Adverse Conditions". *Proc. 3rd International Conference on Speech Prosody*, Dresden, Germany, pp. 149-152, 2006.