Detecting voice pitch in adverse conditions

Ibon Saratxaga, Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sánchez, Iñaki Sainz, Imanol Madariaga

University of the Basque Country
ibon.saratxaga@ehu.es, ikerl@aholab.ehu.es, eva.navas@ehu.es, inma.hernaez@ehu.es, jon.sanchez@ehu.es, inaki@aholab.ehu.es, imanol@bips.bi.ehu.es

Abstract

The need for tools to detect the value of voice pitch (tone, F0 value) in noisy environments has increased in recent years due to new voice coding and voice recognition and conversion systems. This paper presents a detection algorithm which is sufficiently robust to operate in noisy environments, using cepstrum coefficients in combination with the Viterbi algorithm. The algorithm has therefore been evaluated with a specific database, and its performance compared to other algorithms.

Keywords: intonation, pitch, pitch detection

1. Introduction

F0 detection and marking have been extremely important from the beginnings of research into the voice. Calculating F0 has always been an essential factor in research into intonation, and new applications which have emerged - recognition of emotions (Navas et al., 2005), tone recognition of languages (Huang and Seide, 2000), voice conversion (Ney et al., 2004), speaker recognition (Kim et al., 2004) or confirmation (Luengo et al., 2006) – have revived the need for robust pitch detection and marking systems, and there is a particular need for systems which can operate with signals picked up in noisier environments outside recording labs.

With this in mind, a work team at the ECESS (European Center of Excellence on Speech Synthesis, www.ecess.eu) spearheaded a campaign to evaluate F0 detection algorithms. To this end a voice database was set up, hand-marked and visually checked (Kotnik et al., 2006). Our team took part in the evaluation with its improved version of the pitch detection algorithm. The results were presented at an ECESS meeting in Maribor, Slovenia, on 5 July 2006.

It is this pitch detection module which is presented here. The next section will provide a detailed explanation of the various blocks which make up the module. There follows a description of the characteristics of the evaluation along with the results obtained, compared to the published results of a number of other algorithms. The paper finishes with a summary of the results and some possible improvements.

2. Pitch detection module

The pitch detection module ascertains the voice signal tone values at each instant, plotting the points on the tone trend curve. An algorithm based on the values of cepstrum coefficients is used to obtain this curve, and one such device is the Viterbi algorithm.

The curve thus obtained is post-processed in a smoothing block together with information on the voiced or unvoiced nature of each signal interval.

Figure 1 shows the structure of the tone detection module. The only input it requires is the voice signal. This constitutes a considerable advantage, since it enables curves to be obtained for any unknown signal, without the pre-processing required by a number of other systems (to obtain phonetic segmentation, for instance).

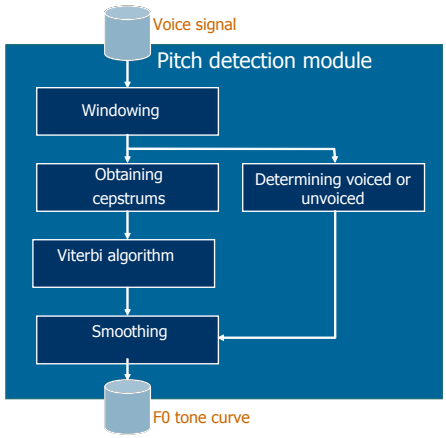

Figure 1: Expression of the pitch detection module block

The result of the algorithm is a PCM-format file containing pitch values in the sampling frequency determined by the user. The specific steps followed to obtain these values are set out below.

2.1. Windowing

It is necessary to conduct an examination of the periodicity or non-periodicity of the signal at each moment over a significant interval. In order to detect signal periodicity, windowing must be carried out using an interval which is sufficiently long but also short enough to allow no time distinction to be lost.

The window must include at least two pitch periods for good detection of periodicity. The lowest pitch expected is set, and in accordance with this the window will have the length of the two periods during the entire analysis.

This is a Hamming window, which reduces the edges, and so the effective length is shorter than two pitch periods.

2.2. Determining voiced or unvoiced

To determine whether a voice interval is voiced or unvoiced, we calculate the ratio between its strength P and Zero Crossing Rate, ZCR. The formula below is used to calculate the strength of the voice interval:

$$P_f = \frac{1}{N}\sum_{i=0}^{N-1}\left|x[i]\right|^2$$

where x[i] are N samples of the windowed signal interval.

Standardised ZCR is obtained by counting up the number of times the signal crosses zero and dividing this by the number of signal interval samples.

The division of P and ZCR reinforces the two independent measurements of the voiced/unvoiced feature. On the one hand, unvoiced signals have more high frequency than voiced signals, and thus the ZCR is greater. On the other, strength tends to be lower in unvoiced sounds. As a result, the P/ZCR ratio will be low in unvoiced signal intervals, and much greater in voiced signal intervals.

2.3. Calculation of cepstrum coefficients

The first step in calculation of the F0 value of the signal interval consists of calculating its cepstrum coefficients. The cepstrums take up the logarithms in the signal's Fourier transform, and these are obtained through the inverse transform

Where pitch values are physiologically defined, there is no need to calculate all the cepstrum coefficients. Thus at the outset the fmin minimum frequency / fmax maximum frequency interval is calculated, and only the maximums of the coefficients will be sought within that interval.

$$x[n] = TF\text{-}1\ \{\log(X(\Omega))\}$$

$$\text{where } X(\Omega) = TF\ \{x[n]\}$$

The next phase consists of finding the maximums for this group of coefficients, and to this end all the coefficients are standardised with a mean value.

The largest M coefficients found are taken to the next block, the Viterbi algorithm block, for selection. Another coefficient is added,

$$c'_i = \frac{c_i}{\bar{c}} \quad / \quad \bar{c} = \frac{1}{i_{\max} - i_{\min} + 1}\sum_{i=i_{\min}}^{i_{\max}} c_i$$

"non-frequency", in order to show that the signal interval is unvoiced.

2.4. Viterbi algorithm

The possible M+1 values which provide the best pitch curve following calculation of all intervals making up the signal must be chosen. This is done using the Viterbi algorithm, choosing values which minimise the sum of the two cost functions: local cost (the cost of simply choosing a value) and transition cost (the cost of choosing a value, taking due account of the value chosen in the previous signal interval).

Local cost is calculated using two components. The first component takes account of the fact that usually the value of F0 is linked to the highest cepstrum coefficient.

The second component of local cost, if the signal interval is voiced, takes account of the fact that the cepstrum value must be to a certain extent greater than the mean cepstrum value.

Transition cost also has two components. The first holds the criteria that the pitch curve, in voiced intervals, is that which continues with no sudden drops.

Excessively rapid changes could be due to selection of the wrong frequency, F0 harmonic or subharmonic.

Lastly, the second component of transition cost causes excessively rapid changes from voiced to unvoiced and vice-versa, since voiced or unvoiced situations persist over a considerable period.

With these cost functions, the Viterbi algorithm finds the curve which obtains the lowest accumulated cost. There may, however, still be occasional erroneous pitch values on this curve, and the result of the voiced/unvoiced detector is not used. It is the last block which does this.

2.5. Smoothing

The curves obtained from the Viterbi algorithm may be erroneous values, particularly when periodicity is unclear, and the information provided by the block specifying the voiced or unvoiced feature can be extremely useful.

Moreover, it is known that a speaker's tone follows the lognormal statistical distribution (Sonmez et al.,

1997). We may feel that values which deviate too much from the mean of this distribution are wrong, and so typical deviation and the mean of all pitch curve values are taken into account.

If the value of a pitch is too far from the mean it is eliminated, and the signal interval is marked as unvoiced. Subsequently, however, if the voiced/unvoiced detector says the signal is voiced, then the new pitch value is interpolated – intermediate of the surrounding instants' pitch values.

Even if, on the other hand, the voiced/unvoiced detector says the signal is unvoiced, if the Viterbi algorithm has suggested an appropriate pitch value, this course is followed.Irudiak ere sar ditzakezu, baina beti ere zenbatuta. Mesedez, testuan bertan irudia aipatu, batzutan mugitzen dira eta. Ikus 1. Irudia, esate baterako.

3. Evaluation

To evaluate the effectiveness of pitch measurement tools, their results must be compared with a group of signals whose pitch is known. The ECESS consortium made arrangements for the evaluation of pitch marking and detection tools, and provided a database of hand-marked signal and pitch curves to this end. The next section contains an explanation of the characteristics of the database in order to specify the results of evaluation.

3.1. Evaluation database

The database used was a subgroup of the Spanish SPEECON database. SPEECON databases were recorded in accordance with the European Commission's SPEECON project conditions (Iskra et al., 2002), and voice signals were recorded in various acoustic environments for the purposes of voice recognition (cars, offices, streets etc.).

The signals were recorded on four channels with simultaneous pick-up through different microphones. The first channel (C0) was recorded with a microphone hooked up to earphones, channel C1 with a Lavalier lapel microphone, channel C2 with a direction microphone at a distance of one metre from the speaker, and channel C3 with a multi-direction microphone positioned 2 or 3 metres from the speaker.

The signal-to-noise ratio (SNR) varies considerably between the different channels under these conditions. Through the clearest channel, C0, SNR is 30 dB, and at the other end of the scale it is 0 dB on channel C3.

60 speaker sentences were chosen to create the reference database (Kotnik et al., 2006), spoken by 30 men and 30 women between 19 and 79 years old. A one-minute recording was made for each speaker, and thus a total of 60 minutes for each channel. From the point of view of semantics, sentences from a variety of corpuses were used.

The channel C0 signals carrying least noise were automatically marked for pitch period, and all of them were then checked manually and corrected. The tone values were taken from these marks at millisecond intervals.

Since the recordings were made simultaneously for all four channels, the pitch values will be the same for all, and will only be more or less delayed because sound has travelled a greater or shorter distance from the speaker to the microphone. To allow the same references to be used on all channels, these delays were equalled out by realigning the recordings, and cross-correlation measuring with channel C0.

3.2. Evaluation criteria

Several common measures found in literature on this topic (Sun, 2002) were specified to gauge the effectiveness of the pitch detection module.

- High or low error values: these error rates measure which percentage is 20% above or below the correct value of the pitch - Gross Error High (GEH) and Gross Error Low (GEL), which are represented as accumulations to provide some idea of the total error. Silences and unvoiced intervals have no effect on these errors.
- Percentage error of voiced and unvoiced signal intervals: the percentage error of voiced signal intervals (voiced error, VE) measures how many intervals are classified as unvoiced error. In similar fashion, the percentage error of unvoiced signal intervals (unvoiced error, UE) measures how many signals are classified as voiced error.
- Differences in mean and standard deviation: these measure the difference between the mean and standard deviation of the estimated and real pitch curve.

3.3. Results

The results set out below were obtained by choosing the parameters which produced acceptable results for all categories and channels. When tests were carried out, the parameter structures which improved the results of each criterion or channel were observed, and so better results may be obtained than those appearing here for each specific application.

Table 1 below shows the results thus obtained:

	C0	C1	C2	C3
VE(%)	9,94	22,62	28,41	35,62
UE(%)	7,44	7,35	6,93	7,35
GEH(%)	0,65	0,31	0,57	0,97
GEL(%)	1,99	2,38	2,45	2,40
AbsMeanDiff(Hz)	0,54	1,87	7,57	11,92
AbsStdDiff(Hz)	1,45	4,46	4,86	6,03

Table 1: Final results

We will divide the results into three groups for the purposes of examination. On the one hand we have the two measurements for voiced/unvoiced errors, UE and VE (Graph 2). The block using Pot/ZCR is extremely noise-sensitive, and so it was not used on the noisy channels. On channels C1, C2 and C3, therefore, the voiced/unvoiced distinction is only deduced from the cepstrum coefficients. The other parameters remain unchanged for all measurements and channels.

In any case, the precision of classification rapidly deteriorates as channel noise increases, and the results are unimpressive. However, as we will observe in the comparison, the classification obtained is one of the best among the algorithms examined.

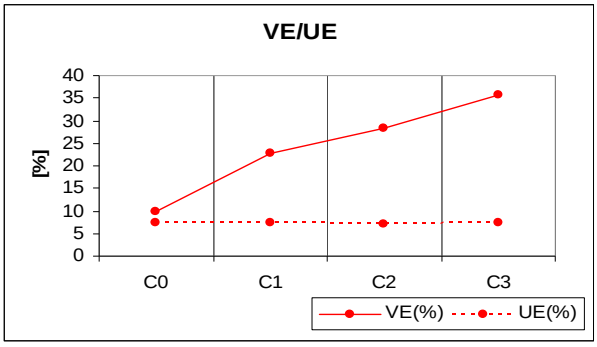

Figure 2: VE/UE trend in the channels

Moreover, obtaining the pitch value every millisecond (4 or 10 values during each pitch period) could create doubts with regard to the utility of this error measurement. At the beginning and end of a voiced signal section, at what points on the initial and end thresholds of the pitch periods should the voiced/unvoiced change be marked?

The next error pair is Gross Error High GEH and Gross Error Low GEL. The results are good in both these measurements for all channels – they rise only slightly even when noise increases, and the sum of both remains around 3% (see Figure 3).

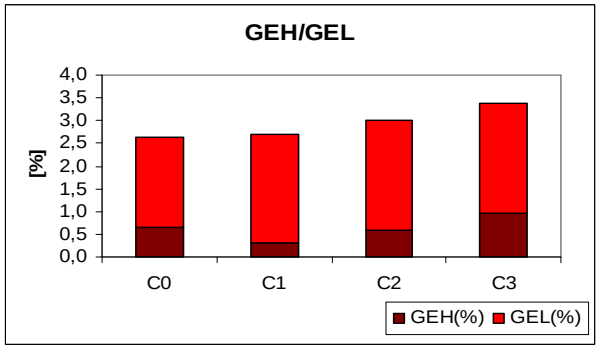

Figure 3: GEH/GEL trend in the channels

These two errors, the difference in mean and standard deviation of the real pitch curve and the estimated pitch curve, show a greater increase as the result of channel noise, as shown in Figure 4 below.

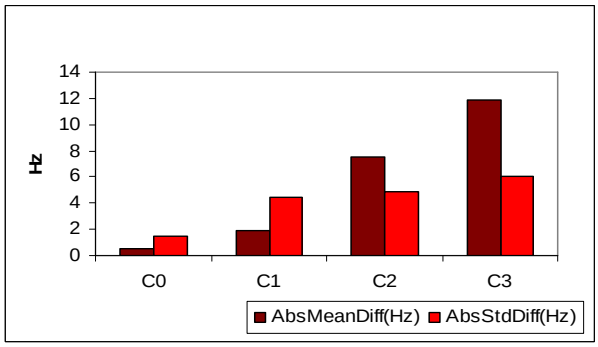

Figure 4: Difference in mean and standard deviation for each channel

3.4. Comparison of results

In order to compare the quality of the results we obtained in our measurements with other algorithms, we performed the same measurements with the Praat programme autocorrelation algorithm (Boersma et al., web). This performs autocorrelation of the windowed signal interval over more than one pitch period, dividing up window autocorrelation to eliminate the window's harmful effects (Boersma, 1993).

Moreover, the results obtained with another two algorithms have also been published (Kotnik et al., 2006) - the results of the authors' algorithm (KOT), those based on the Hilbert transform of LPC residues, and the results obtained by Goncharoff (Goncharoff and Gries, 1998), based on detection of the periodicity of short-term signal energy, with suitable parameters selected dynamically.

The results of our module and the other three sets of results are compared below.

3.4.1. Accumulated voiced and unvoiced interval errors

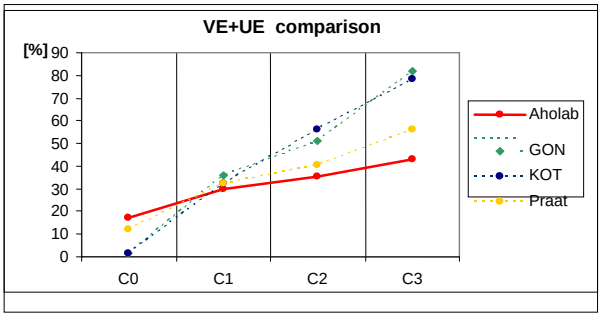

Figure 5: VE+UE comparison for each channel

We have mentioned the relative weakness of these measurements in that time algorithms produce better results in the absence of noise in channel C0. Time algorithms using ZCR, however, quickly fail, although cepstrum detection is much better (Figure 5).

3.4.2. Accumulated higher and lower pitch value errors

In all cases the results of the pitch as measured are ideal (Figure 6). When all channels are good, improvement in the others increases in proportion to the increase in sound from the channel.

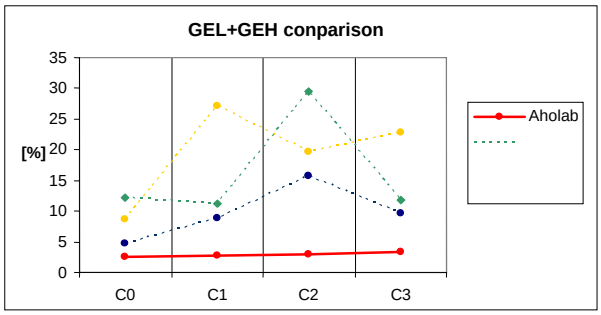

Figure 6. GEL+GEH in all channels

3.4.3. Variations in statistical values

Variations in mean values are almost negligible at the start for channel 0, and increase along with noise (Figure 7), although the increase is smaller than the increase in the other algorithms.

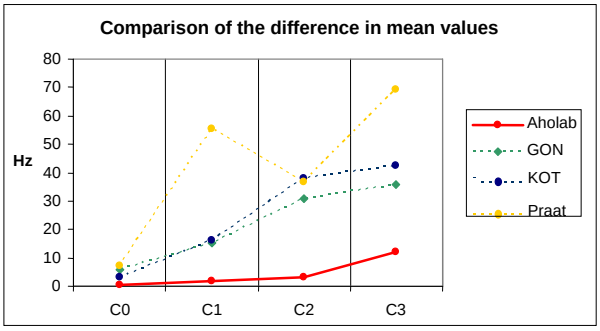

Figure 7. Difference in mean values, all channels

From the point of view of standard deviation (Figure 8), values increase with noise in all cases, and in this case as they increase slowly, our results are better than others in the noisy channels.

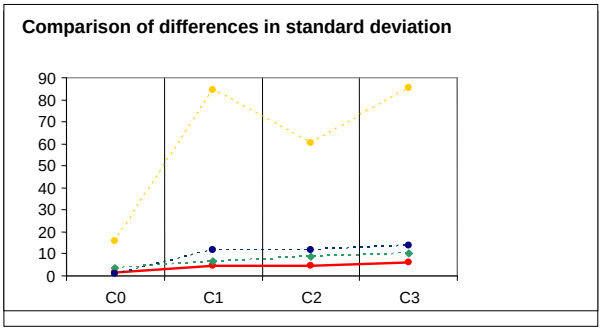

Figure 8. Differences in standard deviation

4. Conclusions

The conclusions of the evaluation of the algorithm proposed to detect tone are extremely encouraging. The system produces good results in low-noise conditions, and the trend is strong in noisy conditions, as the best result was obtained among the algorithms studied.

The algorithm, moreover, took part in the evaluation arranged by the ECESS and obtained some fine results. The pitch detection system obtained the best results in 19 of 32 sections evaluated. In terms of the accuracy of pitch values, it was the best of all in all channels, and obtained the best results over almost all measurements in noisy channels.

By way of a follow-up to our research work, we wish to develop a number of ideas concerning certain improvements which could be made:

- Examination of detection techniques to improve the weakest side of the algorithm: autocorrelation, filtering lower sound thresholds …
- Entering the definition module information in the Viterbi algorithm cost functions.
- Comparison with other algorithms and evaluation with other standard databases.

5. Acknowledgements

E. Navas, I. Hernáez, I. Luengo, J. Sánchez, I. Saratxaga. (2005). *Analysis of the Suitability of Common Corpora for Emotional Speech Modeling in Standard Basque*. Lecture Notes in Artificial Intelligence, LNAI 3658, pp. 265-272.

H.C.H. Huang, F. Seide. (2000). *Pitch tracking and tone features for Mandarin speech recognition*. Procs. ICASSP 2000. Estambul. pp. 1523 - 1526 vol.3.

H. Ney, D. Suendermann, A. Bonafonte, H. Hoege. (2004). *A first step towards text-independent voice conversion*. Procs. INTERSPEECH 2004, Jeju, Corea. pp. 1173-1176.

S. Kim, T. Eriksson, H.G. Kang, D.H. Youn. (2004). *Pitch Synchronous Feature Extraction Method for Speaker Recognition*. Procs. ICASSP 2004. Montreal. pp. 405-408.

I. Luengo, E. Navas, I. Hernáez. (2006). *Effectiveness of Short-Term Prosodic Features for Speaker Verification*. Procs. The Fundamentals of Verbal and Non-verbal Communication and the Biometrical Issue. Vietri sul Mare, Italia.

B. Kotnik, H. Höge, Z. Kacic. (2006). *Evaluation of Pitch Detection Algorithms in Adverse Conditions*. Procs. 3rd International Conference on Speech Prosody, Dresden, Alemania, pp. 149-152.

M. K. Sonmez, L. Heck, M. Weintraub, E. Shriberg. (1997). *A lognormal tied mixture model of pitch for prosody-based speaker recognition*. Procs. EUROSPEECH '97. Rodas, Grecia. vol. 3, pp. 1391-1394.

D.J. Iskra et al. (2002). *SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation*. Procs. LREC'2002. Las Palmas de Gran Canaria. pp. 329-333.

X. Sun. (2002). *Pitch Determination and Voice Quality Analysis Using Subharmonic-To-Harmonic Ratio*. Procs. ICASSP 2002. Orlando, EEUU. pp. 333-336.

P. Boersma, D. Weenink. *Praat: doing phonetics by computer (Version 4.3)* [Computer program]. Retrieved from http://www.praat.org/

P. Boersma. (1993). *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.

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