AUTOMATIC AND ACCUARTE PITCH MARKING OF SPEECH SIGNAL USING AN EXPERT SYSTEM BASED ON LOGICAL COMBINATIONS OF DIFFERENT ALGORITHMS OUTPUTS

K. Ashouri and M.H. Savoji

Electrical and Computer Engineering Faculty, Shahid Beheshti University
Tehran – 1983963113, Iran.
Phone: +98-21-29902258, fax: +98-21-2417940, e-mail:mh_savoji@yahoo.com

ABSTRACT

An expert system comprising a new pitch marking algorithm based on the estimation of the ideal excitation signal, using energy equalization of harmonics of the fundamental frequency present in speech, and three other competent tools is devised and explained in this paper. This expert system uses simple logical combinations of these tools outputs. The behaviour of a human expert is taken into account in developing the post-processing that is necessary to complete each tool and to further improve the results of their combinations. It is noted that, in most cases, combining the results of the new tool and the Childers method, itself based on what goes on behind hand marking by a human expert, is satisfactory. However, accurate and complete pitch marking is best achieved with all four outputs at the expense of some higher processing time.

INTRODUCTION

Pitch marking i.e. the determination of the exact moments of excitation of the vocal tract during the utterance of vowels and semi-vowels is an important task for prosody modifications in speech synthesis using waveform concatenation. Although this has been an on-going research subject for more than the past two decades [1]-[4], a good and simple solution is still to be found as witnessed by the recent efforts dedicated to this subject [5]-[10], due to the inherent difficulties of the problem. Actually, fairly performing algorithms do exist and a new one, based on the estimation of the ideal excitation signal using energy equalization of the harmonics of the fundamental frequency present in speech, has been developed [10]. Nevertheless, all these algorithms suffer from common problems in pitch marking, albeit with different degrees, i.e. missing pitch marks and spurious detected points.

The research described here is based on the idea of developing a simple expert system that combines the outputs of the newly developed algorithm and variants of other three contenders. The developed system has been trained and tested using a data-base of Farsi vowels and semi-vowels.

1. THE DATA-BASE

The waveforms of Farsi (Persian) phrases and words uttered by two male speakers were recorded at 11 and 22 KHz sampling frequencies and digitized with 8 and 16 bits. Then words were segmented into syllables to be saved in separate files as items of our data-base. The phonetic description of the files' contents and other characteristics such as the speaker code and the code of microphone used were attached to each file. A search engine permits to extract all files with a specific phonetic content and other needed characteristics such as the sampling frequency or bit representation for different experiments.

2. PITCH MARKING ALGORITHMS

2.1 A new algorithm based on the estimation of an ideal excitation signal using energy equalization of the harmonics of the fundamental frequency.

This new algorithm estimates an ideal excitation impulse train using the energy equalization of different harmonics of the fundamental frequency present in the speech signal. The idea behind it is fairly simple and is based on the assumption that an ideal excitation signal resulting from an exact pitch marking would be an impulse train with a slowly varying periodicity corresponding to pitch or closed glottis intervals. Therefore if the pitch period can be calculated accurately, this information can be used in a Gaussian shaped filter given by: $H_n(f) = e^{-\left(\frac{f-n.f_0}{n.s_0}\right)^2}$ where f_0 is the fundamental frequency of the voiced signal, n is the harmonic number and $s_0 = (f_{\max} - f_{\min}) / f_0$ is the relative bandwidth of the pass-band filter centered at f_0 . This filter is used to extract the signals corresponding to different harmonics and is zero phased. It does not alter the phase relationship between different harmonics which is introduced by the vocal tract. The result of not correcting the phase relationship of the harmonics in the estimated excitation signal reconstructed by first equalizing their energies and then summing them up will be an enlarged impulse train looking like a comb signal. The parameters of the filter, i.e. f_0 and s_0 are estimated using the famous SIFT algorithm [11]. Note that f_0 is the average fundamental frequency while $f_{\max}$ and $f_{\min}$ are respectively the maximum and minimum of this frequency as calculated by SIFT. N the total number of harmonics to be added (n=1,…, N) is set experimentally to 5. The detection of the excitation points is carried out using a variable threshold applied to overlapping frames of the input signal. The threshold is set as a percentage (e.g. 80%) of the average between the maximum values of the previous and present frames. The frame length is set low in comparison to the average pitch period and the frame overlap is 35 to 40%.

2.1.1 Post-processing

As some spurious excitation points are always detected as in other algorithms, a post-processing is used to remove them. This is basically the same for all other algorithms and is based on the average pitch period calculated for the whole signal undergoing the process. One of two detected points that are less spaced than a percentage of this value, say 70%, is removed in the following manner. One of the two points is taken at choice and the deviation between the pitch period at this point and the points before and after it is calculated. The same calculation is done for the other point. The correct pitch mark (PM) is the one with less deviation and the other point is removed.

2.2 The Childers algorithm

This algorithm is implemented as described in [5] although its parameters such as the frame size, the pattern length etc. are adjusted for the sampling frequency of the input signal. For instance the frame size is set so as to include, on average, three pitch periods of the voiced segment of the input signal.

The Childers algorithm comprises two parts: Separation of voiced / unvoiced parts using the first reflection coefficient and the energy of the linear prediction error or the residual signal. This is followed by a simple correction avoiding separation patterns that include interruptions of voiced sounds by unvoiced parts and vice-versa. It also calculates the approximate pitch period in the cepstrum domain for the voiced part. The excitation points are detected by forming first a template including, in the original algorithm, some 15 points before and some 30 samples after the peak point of the linear prediction excitation or error signal in each frame and cross-correlating the chosen template with the whole frame of the excitation signal. The maxima of the cross-correlation signal are then selected as PM points. In fact this algorithm simulates what is found to be done by a human expert during hand pitch marking.

The detection of excitation points is, in fact, more involved. We first find the two samples with the highest positive and negative values and form the template around the sample with the highest magnitude. The length of the template is adjusted, like other parameters, using a sampling frequency correction factor. The first pitch mark is selected as the maximum positive value in the cross-correlation of the template and the excitation signal. Having detected one pitch mark, we move from this point in both directions towards the beginning and end of the frame and using the approximate pitch periods search for the maximum points around the indicated points and detect the other points.

2.2.1 Post-processing

The Childers' algorithm leaves some spurious PM points that are removed using the same post-processing as explained before. Undetected pitch marks are very rare in this algorithm and if they happen they are located at the edges of the frames. This problem can be remedied by using some overlap between adjacent frames. In our case these points are corrected, as with all the other algorithms, by pooling the results of other pitch mark tools.

2.3 Multi-resolution pitch marking using Wavelet Transform

The Wavelet Transform (WT) is used here as explained in [7] i.e. is based on multi-resolution analysis in a Dyadic WT space. In fact, the algorithm of Hanzo etal. comprises two parts. In the first part, the WT of the signal is calculated using the Fast Wavelet Transform (FWT) algorithm of Mallat [12] and the Spline Wavelet kernels whereby the lower half-band of the signal is successively halved and decimated. Starting with a speech signal of 4KHz bandwidth and increasing the scaling from 1 to 5, a multi-resolution space corresponding to bandwidths of 2000Hz down to 125Hz is made available. The signal decomposed in its lower frequency components $S_i(w)$ i=1,…,5 is used first to detect the periodicity corresponding to the harmonics of the fundamental frequency of the voiced segment. Therefore, in the first part of this algorithm maxima and minima of the signals $S_i(w)$ are detected and points of maximum value are replaced with positive impulses and those corresponding to minima with negative impulses. An impulse present in s_5 must be accompanied by other impulses in the vicinities of the same point in higher bands i.e. s_4, s_3, s_2, s_1. Then, for each point in s_5 other scales are searched and if the impulse is not repeated in all scales it is deleted. The vicinity of each point is found taking into account the decimation factor of 2 and some normalization is used to facilitate the search by compensating the decreasing level in different scale due to halving of the bandwidth. In this way candidate impulses are detected. They must be spaced at least 2.5 msec corresponding to the highest pitch of 400Hz. Therefore, between two neighbouring impulses less spaced than 2.5msec, the smaller is removed as spurious. Now in the second part of the original algorithm, each positive – negative pair of impulses with the highest amplitude is considered as two successive pitch marks and using Dynamic Programming the best set of PMs that give the most consistent periodicity is selected with good accuracy. However, in our scheme only the first part of the algorithm is used avoiding the costly second part. Nevertheless, the same post-processing as in 2.1.1 is used to reduce the number of candidate points. This scheme can be applied to the speech signal or to its linear prediction residual.

2.3.1 Multi-resolution pitch marking results

Our observations show that using speech signal as input is preferred to employing the excitation signal, due to displacement of the detected PMs, although less spurious points are detected in this case. Choosing the vicinity of points in the algorithm is important and so is the selection of the correct peak in this vicinity. But all in all, this algorithm can detect the correct PMs albeit with some spurious ones (note that the second part of the original algorithm is not implemented). Rare are cases where pitch marks are left undetected although in some cases the detected points are misplaced.

2.4 Hilbert Transform based pitch marking

In the original work presented in [1], the Hilbert Envelope of the linear prediction residual i.e. the excitation signal of speech is used. In our implementation we used as input the speech signal, its low-pass filtered version (as used in the SIFT algorithm, filtered below 900Hz) and, the excitation signal. The Hilbert Transform (HT) is calculated in time domain using an FIR filter of 201 coefficients to implement this transform accurately. The Hilbert Envelope has this important characteristic that it shows a peak at the excitation point or PM which can be detected using a threshold. But, due to variations in the signal energy and some non-uniformity in the envelope, it is practically very difficult to use a constant threshold and some pre-processing is normally needed. We used different filtering schemes such as moving average, median, removing minimum value etc. to reduce the time-varying DC off-set which causes the non-uniformity. The best solution to tackle the problem was found to be a variable threshold calculated and applied as follows: The signal is segmented into frames of 150 samples, for instance, having an overlap of half segment length. The threshold for the frame under-going the process is calculated as a percentage, say 95%, of the average between the maximum value of the frame and of the previous one. Nevertheless, the result of applying this threshold is far from desired and the post-processing explained in 2.1.1 is used to remove the spurious PMs.

2.4.1 HT based pitch marking results

Using the low-pass filtered speech and the linear prediction residual, both void of the formants' structure, give usually same and in some cases worse results as the original speech signal. In rare occasions the low-pass filtered signal, which performs very similarly as the excitation signal, results in better pitch marking.

3. OVERALL ASSESSMENT OF DIFFERENT PITCH MARKING TOOLS

The detailed comparison of these algorithms appears in [10]. It can be summarized as follows: Among the four algorithms as implemented here, pitch marking based on estimation of the ideal excitation using energy equalization of harmonics of the fundamental frequency in the speech signal performs the best. Childers algorithm ranks second and has the advantage of giving the pitch period and separating the voiced / unvoiced parts at the same time. The algorithm based on multi-resolution analysis using WT, as implemented here without its dynamic programming sequel, can only be used if completed by an effective post-processing suggested here for all four tools. This algorithm is fairly costly in computation and therefore slow. The algorithm based on HT ranks last although it results in less spurious detections as compared with the former. This algorithm also can not be used without a further post-processing and suffers from fairly heavy computation.

4. AN EXPERT SYSTEM BASED ON LOGICAL COMBINATION OF THE PREVIOUS ALGORITHMS RESULTS

4.1 Finding common PM points among previous results using Logical AND

Here, by applying a simple AND operation on previous PM results we try to find the common points in them. In this operation many detected points are deleted as they not appear in other results. The common points are taken as true PMs and a post-processing is used to fill-in the gaps left by removed points. This operation can be carried out on two or more previous outputs. Most of the time, combining the results of the first two ranking algorithms is satisfactory. However, there are cases where combining all preliminary results is better. Therefore, the final program combines all previous outputs.

4.1.1 Restoration of removed and undetected pitch marks

To restore the removed or undetected PMs we first analyze the remaining points after the AND operation. We calculate the pitch period between the detected points. Any pitch period higher than 1.5 times the average pitch as calculated by the SIFT algorithm is considered to include one or more removed or undetected excitation points. To find these points, the previous and following pitch periods that fulfill the condition and are therefore considered correct are averaged. In case the previous or following periods are also labeled incorrect the average SIFT value is used instead. In any event, the average pitch calculated using the previous and following periods must not exceed the average SIFT value by more than a certain amount (say 30%). Then, using the approximate pitch value for the segment under consideration, the vicinity of the suspected point is searched and two extreme values on each side of the point are considered. The correct excitation point is preceded and followed by points of lower magnitudes. This is based on the damping present usually in the tail of a period and in general in the voiced signal after the excitation point. As there are usually undetected points in the on-set and off-set parts of the voiced segments, beginning and ending segments of the voiced part, as determined by the first and last PMs, are searched separately by the same method to complete the process.

It is worth noting that instead of the AND operator the Majority Rule can be used in the sense that a PM is considered valid if it is present in the majority of the outputs. There is no difference when the common points detected either way are further combined with the result of a Logical OR as explained next. Then, the Logical AND is preferred to the Majority Rule for its simplicity and speed.

4.2 Finding union of all PM points using Logical OR

The union of the preliminary outputs is obtained using a simple OR operator. In almost all cases the valid excitation points are found among this union but, obviously some spurious points are also detected which must be removed by a post-processing and it can be said that the performance of this combination depends on the post-processor. Again any two combinations of the four preliminary results can be used. Using the outputs of the first two best tools gives, in most cases, results as good as using all four outputs. But, there are cases that both algorithms miss a valid point and therefore, using all four outputs is preferred.

4.2.1 Removing spurious excitation points

A variant of the previously explained post-processor is used. It is assumed that the least pitch period can be 70% of the average SIFT value and therefore, between two PMs that are less spaced than this amount one must be removed. If the spacing is less than 20% of the average SIFT value, the point on the right hand side (farther in time) is removed. This is because, in this case, the second extreme point is mistaken as a valid PM. If this condition is not fulfilled we proceed on the basis of maximum deviations as explained in 2.1.1. But, if the deviations are close, the selection is based, this time, on comparing the magnitudes of candidate PMs and the one with higher magnitude is selected as valid.
The results of this combination are satisfactory in the same manner as those of the AND operator. But, there are still some difficult cases where both combinations fail. This is why we use a further OR operation on the previous two outputs. This step is also completed by a post-processing of the same kind as explained here above. A typical example is given in Figure 1 below.

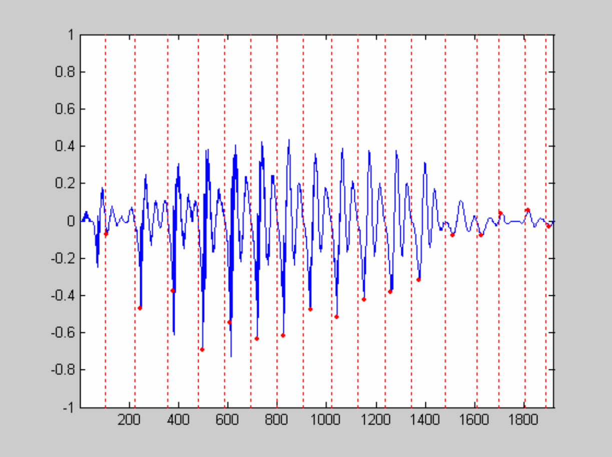

Figure 1: The final pitch marking result on a ‘wi’ sound (signal digitized at 11KHz with 8b/sample).

This way a completely satisfactory expert system is obtained which did not fail on any file in our data-base. Nevertheless, a facility is provided for hand correcting the final result using, among others, a zoom facility.

CONCLUSION

An expert system based on a new and three existing pitch marking tools has been devised. This expert system uses simple logical combinations of the tools’ outputs. The behaviour of a human expert system has been taken into account in developing the post-processing that is necessary to complete each tool and to further improve the results of their combinations. It is noted that, in most cases, combining the results of the new tool based on the estimation of the ideal excitation signal using energy equalization of harmonics of the fundamental frequency present in speech and the Childers method, itself based on what goes on behind hand marking by a human expert, is satisfactory. However, accurate and complete pitch marking is best achieved with all four outputs at the expense of some higher processing time.

REFERENCES

[1] T.V. Ananthapadmanabha and B. Yegnanarayana; “Epoch extraction from linear prediction residual for identification of closed glottis interval”, IEEE Trans. On ASSP, Vol. ASSP-27, No.4, August 1979.

[2] Y.M. Cheng and D. O’Shaughnessy; “Automatic and reliable estimation of glottal closure instant and period”, IEEE Trans. On ASSP, Vol.37, No.12, Dec. 1989.

[3] R. Hennig; “A fast expert program for pitch extraction”, EUROSPEECH Conference, 1989.

[4] F.M. Gimenez de los Galanes, M.H. Savoji, J.M. Pardo; “Marcador automatico de excitation glotal’; Proc. URSI 93; Valencia, Spain, 1993.

[5] D.G. Childers; “Speech processing and synthesis toolboxes”, John Wiley and Sons Inc., 2000.

[6] M. Sakomoto and T. Saitoh; “An automatic pitch-marking method using wavelet transform”, ICSLP-2000 Conference on Spoken Language Processing.

[7] L. Hanzo etal.; “wavelet and pitch detection”, Chapter 12 in “Voice compression and communication principles and applications for fixed and wireless channels”, IEEE Series on Digital and Mobile Applications, 2001.

[8] V. Colotte and Y. Laprie; “Higher precision pitch marking for TD-PSOLA”, EUSIPCO-2002 Conference.

[9] D. Tihelka and J. Matousek; “Comparison of various speech based signals and pitch marking detection methods for use in speech synthesis”, 16th International EURASIP-2002 Conference.

[10] K. Ashouri and M.H. Savoji ; ‘A new pitch marking algorithm based on harmonics energy equalization of the speech signal’; accepted for publication in SETIT-2004 Conf., Susa, Tunisia, March 2004.

[11] J.D. Markel and A.H. Gray; “Linear prediction of speech”, Springer-Verlag, 1976.

[12] S. Mallat; ‘A wavelet tour of signal processing’; Academic Press, 1999.