ENHANCED SPEECH CLASSIFICATION AND PITCH DETECTION

Peter Vepřek† and Michael S. Scordilis‡

†Electrical and Electronic Engineering Department
The University of Melbourne, Australia
‡Wire Communication Laboratory
The University of Patras, Greece

Abstract — Speech classification into voiced and unvoiced or silent portions is important in many speech processing applications. In addition, segmentation of voiced speech into individual pitch epochs is necessary in several high quality speech synthesis and coding techniques. In this paper, two different pitch detection methods are evaluated and a set of enhancements is presented which substantially enhance performance.

INTRODUCTION

Pitch detection algorithms (PDAs) aim at determining the pitch period defined as the time between the beginnings of two successive glottal excitations. They also achieve classification of the speech signal into voiced or unvoiced regions. There exist a large variety of algorithms addressing the pitch detection problem. Most of them are frame-based and as a result, they characterise the signal according to the state which happens to dominate the selected signal segment (see Hess (1983) and (1992) for a detailed overview). For voiced speech these techniques provide only an averaged pitch profile for the given time and are unable to resolve the inter-frame temporal differences and to yield higher structural accuracy.

While such approaches work well and are useful in many speech processing tasks, their performance is inadequate for many current speech coding and synthesis applications, where the pitch period or pitch epoch is an essential unit. In speech synthesis in particular, new techniques based on the pitch-synchronous analysis of the signal are increasingly used (Moulines et al., 1990; Dutoit, 1994) and they require marking of individual pitch epochs. A number of techniques with good performance have been developed for segmenting speech into pitch epochs (Cheng & O'Shaughnessy, 1989; Medan et al., 1991; R. Di Fransesco, E. Moulines, 1989). However, manual inspection and correction of errors are still necessary (Moulines et al., 1990).

In this paper, objective performance criteria for pitch segmentation algorithms are introduced and used in the evaluation of two basic PDAs. The sources of errors are analysed and a set of enhancements is applied to each algorithm. Pitch marking performance is measured against manually segmented speech data and it is used to rate its success in the segmentation of speech into voiced or unvoiced segments and in the correct marking of the pitch epochs. In this work the maximum signal value in a pitch period was taken to coincide with the center point of the pitch epoch. The sampling period represented the temporal resolution unit. In order to interpret the results in a meaningful way, voiced classification performance was measured as the percentage of signal correctly classified during voice activity. Similarly, unvoiced or silence classification was measured against the hand-labelled marking. Pitch marking success was measured as the total number of insertions and deletions in the speech material used for analysis.

TWO BASIC PITCH DETERMINATION ALGORITHMS

Two basic PDAs were selected for evaluation: the simplified inverse filter tracking (SIFT) and the optimal temporal similarity (OTS) algorithms. They represent two completely different approaches to the problem since the first is a well-tested frame-based method while the second is an event-based method. They were evaluated on a common speech database and their performance was analysed.

The SIFT algorithm

This pitch determination algorithm was an implementation of the SIFT algorithm proposed by Markel (1972). The speech signal was first low pass filtered using a FIR filter with sharp cut-off frequency at 800 Hz and then downsampled by a factor of 5. A fourth order asynchronous linear prediction (LP) analysis was performed on the resulting signal using a 32 ms frame length and 12 ms frame shift. Every frame was Hamming windowed and optimally pre-emphasised. The prediction residual signal was obtained by inverse filtering the input signal. The residual signal was subsequently windowed and its normalised autocorrelation was calculated. The autocorrelation function was searched for peaks within a range of allowed values which corresponded to F0 from 40 Hz to 500 Hz, and which also exceeded a certain threshold value. If this search succeeded, the frame was classified as voiced. Otherwise, if the two previous frames were both voiced, the current maximum value was compared to a second, lower threshold. When that threshold was exceeded, the frame was also classified as voiced, otherwise it was classified as unvoiced. When a single unvoiced frame between two voiced frames was detected, the decision for that frame was changed from unvoiced to voiced and a pitch value corresponding to the average pitch of the two adjacent voiced frames was assigned. To increase the resolution of the estimated pitch period, the autocorrelation function of every voiced frame was interpolated in the neighbourhood of the calculated preliminary pitch using parabolic interpolation based on the peak and its two adjacent values.

The OTS algorithm

This is the integer pitch determination method proposed by Medan et al (1991). It is an event-based PDA. As in the SIFT, speech was first low-pass filtered using an FIR filter with sharp cut-off frequency at 800 Hz and downsampled by a factor of 5. A search pointer was set at the beginning of the utterance and a block of the first 2n samples of the utterance was selected and divided into two consecutive, non-overlapping frames of equal length. The value of n represented the pitch period (in samples) and corresponded to a fundamental frequency range between 40 Hz and 500 Hz. The two signal frames were then used to compute their normalised cross-correlation. This computation was repeated for all values of n in the allowable range. The resulting cross-correlation function was searched for its maximum value, which if it was lower than an experimentally set global voicing threshold, Th_G, caused the first smallest signal block to be considered as unvoiced. Otherwise, the beginning of the utterance was considered voiced. For voiced speech, the cross-correlation function exhibits regular local maxima at multiples and possibly sub-multiples of the true pitch period, T0. Those intervals, n_i for i =1, 2, ..., I, where the function exceeded Th_G, form the initial set of legitimate pitch period candidates, $N_p = \{n_1, n_2, ..., n_I\}$. Setting the pitch period equal to the value of n which corresponds to the maximum value of the computed cross-correlation function may result in the wrong pitch. As an alternative, all N_p values were used to calculate the normalized cross-correlation for all such pairs of signal segments equal to the longest candidate pitch period, n_I, and separated by all candidate pitch periods, n_i. The first candidate to exceed a preset threshold was chosen as the pitch period. Finally, a new search pointer was set to the first signal sample after the smallest block if the signal was classified as unvoiced or to the first sample after the first detected fundamental period in the case of voiced classification. The process was repeated until the end of the utterance. While the voicing threshold was preset for all speech material used, the pitch threshold was dynamically adjusted throughout every utterance to minimise the chances of wrong pitch assignment.

Evaluation of the Basic Algorithms

Performance of the two basic pitch detection algorithms was tested on speech which was hand-marked into voiced/unvoiced regions, and individual pitch epochs. Five measures of performance were used:

1) unvoiced to voiced speech classification,
2) voiced to unvoiced speech classification,
3) pitch period deletions,
4) pitch period insertions, and
5) pitch period inaccuracies.

The first two types of error occurred during the initial speech classification into voiced and unvoiced or

silent regions. The next three types of error occurred in voiced speech regions only. The assignment of multiple of the true pitch period resulted in pitch period deletions, while halving the true pitch period errors caused pitch insertions. The last error measure concerns all other errors which caused deviation from the true pitch. In this work, resolution was equal to one sampling period.

Typical adult speech was used for the evaluation. The speech material was collected in a "quiet" office environment at a sampling rate of 11025 Hz, with 16-bit amplitude resolution. Speech underwent extensive manual labelling before it could be used. The spoken material consisted of a set of seven phonetically-rich sentences, all read by two female and two male non-professional speakers. Additional variability was included by having the sentences produced at varying speaking rates and in different orders for each speaker. Overall, 58% of the speech data were labelled as voiced, containing over 10,000 individual pitch periods in total.

The performance of the two algorithms is summarised in Table 1. The OTS PDA performed better with overall error of 13.94%. The SIFT PDA had the overall error of 16.93%. In terms of computational requirements, the OTS required more computations than SIFT.

Algorithm	Speakers	Error rates [%]					
		UV to V	V to UV	Deletions	Insertions	Inaccuracy	Total
SIFT	Male	8.97	7.27	0.00	0.27	0.86	17.25
	Female	7.38	4.11	4.07	0.27	2.72	16.81
	Both	**8.13**	**5.59**	**2.15**	**0.27**	**1.81**	**16.93**
OTS	Male	8.09	2.11	1.82	0.50	1.35	12.82
	Female	8.29	3.67	1.68	0.09	1.91	14.94
	Both	**8.20**	**2.93**	**1.75**	**0.29**	**1.64**	**13.94**

Table 1. Performance of the two basic pitch period determination algorithms.

PROBLEM AREAS

Examination of the classification results of the basic algorithms revealed that a number of specific cases existed which caused degraded performance. These problems are summarised as follows.

Unvoiced speech classified as voiced — The algorithms occasionally misclassified regions of the signal with low energy as being voiced and proceeded to assign certain pitch periods. These regions were usually non-speech leading or trailing the spoken utterance, left over by imperfect end-point detection, as well as pauses between words and closures preceding stop consonants.

Voiced speech classified as unvoiced — Rapid coarticulation resulting from fast speech production rates caused the signal characteristics to change rapidly, resulting in voiced speech having low levels of similarity between adjacent segments. When that occurred, both algorithms tended to classify voiced regions as being unvoiced.

Multiple and sub-multiple pitch period detection — Both algorithms were quite robust against double and half pitch period detection within most voiced regions, especially vowels. However, in the presence of initial nasalisation or when the first formant was prominent there was a tendency to favor half of the pitch period as the correct. Multiple pitch period errors were sometimes made as a result of relatively high similarity between two adjacent pitch epoch pairs.

Gross pitch period inaccuracy — Faster speech rate and coarticulation effects resulting in lower levels of similarity between adjacent pitch epochs caused the algorithms to detect wrong pitch other than multiple or sub-multiple of the true pitch.

Entrapment in wrong pitch range — When the multiple and sub-multiple pitch period or the gross pitch period inaccuracy errors occurred several times in succession then this kind of error could be propagated further into the immediately following voiced region as a result of tracking. In this situation, the pitch period range would be wrongly narrowed down and the true pitch period would be outside this range. As a result, the correct pitch period would not be detected until the end of this voiced region, and only then the pitch period range would be widened.

End of utterances in rough or diplophonic speech — At the trailing end of an utterance or in rapidly varying or irregular speech, pitch assignment was often problematic. This was particularly the case with one male speaker who often produced diplophonic speech and in some cases where speakers switched to creaky voice.

BASIC ALGORITHMIC ENHANCEMENTS AND THEIR EVALUATION

Pitch detection techniques use principles which aims to reveal the periodicity in speech, and they are designed to provide best performance within their inherent assumptions and constraints. Besides the possible fine-tuning of a PDA with the purpose of making it more effective for given speech material, there are additional methods that could be used to improve performance. These methods fall into two categories, namely, the processing of the speech signal prior to the presentation to the PDA, and the post-processing of the obtained information after the detection process. Pre-processing is often dependent on a particular pitch detection technique, and includes clipping the input signal in different ways. Although this approach might work well with some methods, it is a non-linear operation that emphasises the high-frequency energy of the speech and that could be detrimental to others. Post-processing usually involves smoothing the derived pitch contours. Smoothing aims to locate isolated pitch estimates and bring them back to line with the general contour. This is best achieved with median filtering, and it best suits methods that yield pitch contours rather than methods that mark individual periods (Hess, 1992).

In this work, the error sources were identified and a set of post-processing methods applicable to both pitch detectors was sought to address the most serious problem areas, and improve performance. These techniques were introduced after analysis of the identified problem areas, and testing a set of heuristics. These enhancement procedures are listed in Table 2. The first procedure modifies the basic algorithm of the event-based technique, while the remaining are applicable to both approaches.

Enhanced pitch period range tracking for the event-based PD algorithm (OTS)

After detecting three consecutive unvoiced frames, each 5 ms long, the basic OTS algorithm widened the pitch period range to its maximum which in this work corresponded to pitch frequencies between 40 Hz and 500 Hz. This caused big changes in pitch period over short unvoiced or non-speech segments. In other words, when the algorithm misclassified voiced speech as being unvoiced, it consequently lost track of pitch period due to the pitch period range widening. This in turn allowed significantly different pitch in the next portion of the signal to be detected. The new approach was to widen the pitch period range by 10% for the second, third, up to sixth unvoiced frame and set it to the maximum range afterwards. This prevented big changes in pitch while allowing for smaller changes to be detected. Also, when an unvoiced frame after a voiced one was detected, the pitch period range was immediately widened by 5%. This allowed for bigger-than-usual jumps in the value of pitch period as compared to the basic algorithms. Finally, the basic algorithm was modified to adaptively narrow down the pitch period range after detecting three consecutive voiced frames. The modified algorithm widened the pitch period range by 5% within the first three voiced frames and then performed adaptive tracking thus allowing for the first estimate of pitch period to be possibly wrong and not to influence further values.

Initial half period correction

The presence of nasals at the start of phrases often caused half the pitch period to be selected as correct. Also, the transition into a following vowel with clear periodicity caused the algorithms to classify the short transition between the two voiced sounds as unvoiced. When the initial voiced sound was shorter than 65 ms, the unvoiced gap was shorter than 25 ms and the following voiced segment was at least 200 ms long, then

1. Enhanced pitch period range tracking
2. Correcting initial half pitch period
3. Correcting multiple and sub-multiple pitch periods
4. Rejecting short voiced segments
5. Rejecting short unvoiced segments
6. Rejecting low energy voiced segments
7. Rejecting high energy unvoiced segments
8. Histogram computation and correction of multiple and sub-multiple pitch periods

Table 2. List of pitch algorithm enhancements.

the initial voiced segment would be re-examined. If the initial pitch estimate was half of that in the long voiced segment, after allowing for a 15% tolerance, then it would be doubled, otherwise it would be left unaltered.

Multiple and sub-multiple period correction

Problems caused by nasals or any other semi-vowel were not restricted to phrase-initial positions but could occur anywhere. In these instances, the location of the first formant caused the second or third pitch harmonic to dominate in the waveform resulting in the algorithm selecting multiples of the period. Another problem was caused by vocal registers in diplophonic speech, where a dominant pitch epoch separates a group of one or more epochs of lower energy. This situation resulted in sub-multiples of the pitch being selected. Detecting most such errors and correcting them was achieved with a procedure that examined all voiced segments of 40 ms or longer where pitch multiples or sub-multiples appeared to exist in a maximum total of 25% of the segment. If such segments existed, then all corresponding estimates outside a band of 20% tolerance were brought within the right range by appropriate upscaling or downscaling.

Rejection of short voiced segments

Noise contributions, electronic or acoustic, or pseudo-periodicity occasionally caused short unvoiced segments to appear as voiced, and to be assigned a pitch period by the detection algorithm. Such segments were usually well separated from adjacent relatively long voiced areas, and were rejected as unlikely to occur in normal speech production. Whenever such segments were less than 35 ms long and separated by at least 25 ms gaps on either side, they were rejected and classified as unvoiced.

Rejection of short unvoiced segments

Similarly with the previous enhancement, short unvoiced speech segments may appear in relatively long voiced segments, as the result of rapidly changing pitch and articulation or non-speech clicks and acoustic noise. Such short segments were declared voiced when: (a) they were less than 10 ms long, and (b) the adjacent voiced segments on their left and right were longer than 20 ms each, or if one of them was longer than 30 ms, and (c) the absolute difference between the pitch periods to the left and right of the segment was less than 1 ms. The pitch period assigned to such segments was the average of the adjacent periods.

Rejection of low energy voiced segments

Within a phrase, voiced speech usually has more energy than unvoiced segments or non-speech. This was used to check the validity of the classification of speech as voiced. For that purpose the smoothed energy contour of a sentence was computed and low and high energy thresholds were tracked. Whenever the energy level of a voiced region was below the threshold for voiced speech, that region was classified as unvoiced.

Rejection of high energy unvoiced segments

Short unvoiced regions with energy levels greater than a threshold were re-examined. If the energy exceeded a certain threshold, then such segments were declared voiced and were assigned pitch period value equal to the average of the periods to its immediate left and right.

Corrections using pitch period histogram

Pitch period histograms of several utterances were obtained and it was observed that while a properly classified utterance peaked only in one location in the histogram, an utterance with pitch errors produced a histogram with secondary peaks, which corresponded to multiple and sub-multiple pitch values. These peaks were usually located in areas distinct from the dominant peak which was in the region of the correct pitch. This method located the secondary peaks and eliminated the lower ones by multiplication and the higher ones by division of their periods with the appropriate integer value. This procedure proved quite effective in eliminating this class of errors.

Evaluation of the enhancements

Results of evaluation of the enhanced algorithms are summarised in Table 3. It shows that overall performance of both pitch determination algorithm included in this study was improved. The overall error of the enhanced SIFT algorithm was 9.95% and that of the OTS method was 6.46%.

For SIFT the enhancements significantly reduced unvoiced-to-voiced errors, practically eliminated deletions (multiple pitch periods), and reduced all other errors with the exception of voiced-to-unvoiced errors. In the case of OTS the enhancements reduced all types of errors. In particular, unvoiced-to-voiced errors were greatly reduced, while deletions and voiced-to-unvoiced errors were reduced considerably.

It was observed that the enhancements were particularly effective in correcting deletions and insertions as result of corrections based on the pitch period histogram. This class of errors was practically eliminated with all remaining errors being less than 1%. In case of voicing errors, the more problematic unvoiced-to-voiced errors were also largely reduced.

Algorithm	Speakers	Error rates [%]					
		UV to V	V to UV	Deletions	Insertions	Inaccuracy	Total
SIFT	Male	2.64	8.11	0.00	0.05	0.72	11.49
	Female	1.39	4.51	0.44	0.22	2.30	8.60
	Both	**1.98**	**6.20**	**0.23**	**0.14**	**1.56**	**9.95**
OTS	Male	3.28	2.32	0.00	0.08	1.14	6.67
	Female	3.03	1.48	0.30	0.28	1.34	6.20
	Both	**3.15**	**1.87**	**0.16**	**0.18**	**1.24**	**6.46**

Table 3. Performance of the two enhanced pitch period determination algorithms.

CONCLUSION

A set of enhancements for automatic segmentation of speech into voiced and unvoiced or silent intervals and for segmentation of voiced speech into individual pitch epochs was developed. Objective performance measures were introduced and used to compare two markedly different pitch detection approaches, first in their basic form and then in enhanced form, against the markings of hand-labelled sentences spoken by female and male speakers. The evaluation showed that the developed set of enhancements can be successfully applied to other pitch detection algorithms as well. The resulting enhanced algorithms, particularly the OTS method, can be effectively used for pitch synchronous analysis currently used in many speech synthesis and coding applications.

REFERENCES

Cheng, Y. M., O'Shaughnessy, D. (1989) *Automatic and reliable estimation of glottal closure instant and period*, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, pp. 1805-1814.

Di Francesco, R., Moulines, E. (1989) *Detection of the glottal closure by jumps in the statistical properties of the signal*, Proceedings of the Eurospeech, pp. 39-42.

Dutoit, T. (1994) *High quality text-to-speech synthesis: a comparison of four candidate algorithms*, Proceedings of ICASSP, vol. 1, pp. 565-568.

Hess, W. J. (1983) *Pitch determination of speech signals*, (New York: Springer-Verlag).

Hess, W. J. (1992) *Pitch and voicing determination*, in S. Furui, M. M. Sondhi, eds., Advances in speech signal processing, (New York: Marcel Dekker).

Markel, J. D. (1972) *The SIFT algorithm for fundamental frequency estimation*, IEEE Transactions on Audio and Electroacoustics, vol. 20, pp. 367-377.

Medan, Y., Yair, E., Chazan, D. (1991) *Super resolution pitch determination of speech signals*, IEEE Transactions on Signal Processing, vol. 39, pp. 40-48.

Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J. L., Le Faucheur, L., Marty, F., Charpentier, F., Sorin (1990), C. *A real-time French text-to-speech system generating high-quality synthetic speech*, Proceedings of ICASSP, vol. 1, pp. 309-312.