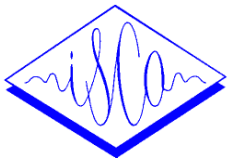

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

CONSISTENT PITCH MARKING

Raymond Veldhuis

IPO, Center for User-System Interaction, Eindhoven, The Netherlands
R.N.J.Veldhuis@tue.nl

ABSTRACT

The pitch-marking algorithm presented in this paper avoids inconsistency errors between pitch markers in subsequent fundamental periods by adding the requirements of waveform and pitch consistency. The approach is as follows. Candidate pitch markers satisfying user-defined properties for pitch marking are selected first. Dynamic programming is then used to find the sequence of candidate pitch markers which optimally satisfies the waveform- and pitch-consistency requirements. The algorithm is described in detail and results are presented.

1. INTRODUCTION

Pitch markers are used for various purposes in various applications. Examples are: defining concatenation points in unit-selection based speech synthesis [1], providing center points of analysis windows for pitch-synchronous signal analysis or modifications of pitch and duration [3], and providing segmentation boundaries [1]. The positioning of pitch markers is usually done based on some characteristic property, e.g. the position of a marker within a fundamental period is: at one of the higher maxima of the absolute value of the signal, at the first zero crossing before the maximum positive peak [1], at the instant of glottal excitation [7], at the moment of excitation according to the LPC model [6, 2], or at the center of gravity [4]. The characteristic property may be selected depending on the application in which pitch marking is applied. Of course, one would also require that the distance between adjacent pitch markers be near to one fundamental period of the speech signal.

In fact, using only the characteristic property as a requirement may lead to errors in two ways. First, several positions in a fundamental period may satisfy the characteristic property requirement to some extent and a wrong pitch marker may be selected. Second, there may be just one position in a fundamental period satisfying the characteristic property, but it may not be optimal in the sense that a small shift would produce a better marker. The pitch-marking algorithm presented here avoids both types of errors by allowing more than one candidate for pitch marking per fundamental period and by imposing additional requirements relating the positions of pitch markers to those in adjacent fundamental periods.

Three requirements are imposed. First, the *characteristic-property requirement* demands that the markers satisfy a predefined property, such as in the examples mentioned above. Second, in most cases adjacent pitch periods have similar waveforms. This observation leads to the formulation of the *waveform-consistency requirement* demanding that signal portions around adjacent pitch markers also be similar. Similarity of the waveforms of subsequent fundamental periods in voiced speech is also successfully employed in the pitch tracker RAPT that was published in [5]. Waveform consistency may not always be present across certain phoneme boundaries. Therefore, assuming that the fundamental frequency of the speech varies only slowly between adjacent fundamental periods, the *pitch-consistency requirement* demands that the distance between two consecutive pitch markers be close to a fundamental period.

In a nutshell, the algorithm works as follows. First, candidate pitch markers satisfying the characteristic-property requirement to a certain extent are computed. Local costs are computed, which quantify to what extent each candidate satisfies the characteristic-property requirement. A limited set of possible predecessors is determined for each candidate. Transition costs are computed expressing to what extent each predecessor in this set is waveform consistent with the current candidate. If the minimum transition cost is above a threshold, it is decided that the current candidate cannot not satisfy the waveform-consistency requirement and transition costs based on the local fundamental period, quantifying pitch consistency, are used instead. The optimal sequence of candidates, for which the sum of local and transition costs is minimum, is obtained by dynamic programming.

An attractive feature of this algorithm is that it is not dependent of the particular choice of candidate pitch markers, but that it can be used for any characteristic property for which a local cost function can be specified. Section 2 will describe the algorithm in detail.

The best, and possibly the only, way of demonstrating

that a new method for pitch marking works is to compare it with previous approaches in the application it was intended for. This is so, because a pitch marker is not a concept of the speech signal, but rather a concept of an application. The quality of the pitch-marking method should therefore be measured indirectly through the performance of the application. Such an extensive evaluation will be presented in a forthcoming journal paper. Here Section 3 will present some results for two characteristic properties. These results will demonstrate the benefits of this approach over approaches without the requirements of waveform or pitch consistency. Finally, Section 4 will present conclusions.

2. THE ALGORITHM

Speech signals are sampled data, with sampling frequency F_{s}. The algorithm computes pitch markers for a voiced subsequence $\{s_k|k = k_{\min}, \ldots, k_{\max}\}$ of a speech signal. A subsequence of estimates $\{n_{0,k}|k = k_{\min}, \ldots, k_{\max}\}$ of the fundamental period at each sample position must be available. Such a subsequence can be derived from the output of a pitch tracker.

Candidate pitch markers are the positions in a fundamental period which exhibit a certain user-defined characteristic property such as listed in the introduction. It is recommended to add a few neighboring candidates for each candidate. This will give the algorithm the possibility of finding pitch markers close to the original candidates, which will improve the results. Often the characteristic property prescribes that candidates be located at (or near) maxima of a positive function $F(k)$, $k = k_{\min}, \ldots, k_{\max}$, such as the power envelope or the absolute value of the signal. In that case a recipe for including neighboring candidates is to select all sample points c around the positions $\hat{c}$ of the local maxima of $F(k)$ for which $F(c) \geq \gamma |F(\hat{c})|$. The choice $\gamma = 0.9$ has produced good results. The final candidates are represented as an index set $\{c|k_{\min} \leq c \leq k_{\max}\}$.

The local cost function of a candidate c, expressing to what extent it is a suitable pitch marker according to the characteristic property, is denoted by $C_{\mathrm{l}}(c)$. The user can define $C_{\mathrm{l}}(c)$ in many ways. If the characteristic property requires candidates to be located at (or near) maxima of a positive function $F(k)$, $C_{\mathrm{l}}(c)$ can be computed in a straightforward manner by

$$C_{\mathrm{l}}(c) = 1 - \frac{F(c)}{F_{\max}(c)}, \tag{1}$$

with

$$F_{\max}(c) = \max_{c-0.5n_{0,c} \leq d \leq c+0.5n_{0,c}}. \tag{2}$$

Equation (1) compares c with the best candidate, according to the characteristic property, in the fundamental period of which c is the center. Of course,

$$0 \leq C_{\mathrm{l}}(c) \leq 1, \tag{3}$$

and equality only holds on the left-hand if $F(c)$ is the maximum in this fundamental period.

The normalized cross-correlation coefficient

$$\rho(c,d) = \max\left(0, \frac{\sum_{k=0}^{K-1} s_{c+k}\, s_{d+k}}{\sqrt{\sum_{k=0}^{K-1} s_{c+k}^2 \sum_{k=0}^{K-1} s_{d+k}^2}}\right) \tag{4}$$

of the

$$K = \mathrm{round}\frac{(n_{0,c})}{2} \tag{5}$$

samples following candidate pitch markers c and d, $d < c$, respectively is used to define the transition cost function

$$C_{\mathrm{w}}(c|d) = 1 - \rho(c,d), \tag{6}$$

which quantifies the waveform consistency between the candidates c and d. Because $0 \leq \rho(c,d) \leq 1$, it follows that

$$0 \leq C_{\mathrm{w}}(c|d) \leq 1. \tag{7}$$

Equality only holds on the left-hand when the K samples following c and d are identical; $C_{\mathrm{w}}(c|d)$ increases with decreasing similarity between the groups.

A weaker consistency requirement is pitch consistency, demanding that the interval between adjacent pitch markers be close to one fundamental period. The transition cost function used to quantify pitch consistency between the candidates c and d, $d < c$, is given by

$$C_{\mathrm{p}}(c|d) = \left(\frac{c - (d + n_{0,d})}{\alpha n_{0,d}}\right)^2. \tag{8}$$

This cost function is proportional to the square of the relative deviation of the expected location $d + n_{0,d}$ of c in the case of maximum pitch consistency. The constant α can be used to tune the acceptable range of this deviation. The choice $\alpha = 0.07$ has proven to give good results.

The consistent pitch marking algorithm presented here tries to find a candidate sequence $c_1, \ldots, c_M$, with c_1 in the first and c_M in the last fundamental period of the voiced subsequence $\{s_k|k = k_{\min}, \ldots, k_{\max}\}$, which minimizes the total cost

$$C(c_1, \ldots, c_M) = C_{\mathrm{l}}(c_1) + \sum_{j=2}^{M} C_{\mathrm{t}}(c_j|c_{j-1}) + C_{\mathrm{l}}(c_j), \tag{9}$$

with $C_{\mathrm{t}}(c|d)$ the transition cost function, defined by

$$C_{\mathrm{t}}(c|d) = \begin{cases} C_{\mathrm{w}}(c|d), & \text{if } \max\limits_{d \in D(c)} \rho(c,d) > \rho_{\min}, \\ C_{\mathrm{p}}(c|d), & \text{otherwise.} \end{cases} \tag{10}$$

Here $D(c)$ denotes the search space

$$D(c) = \{d | (1-\delta) n_{0,d} \leq c \leq (1+\delta) n_{0,d}\} \qquad (11)$$

in which likely predecessors of candidate c are sought. The choice $\delta = 0.5$, with which the search space is equal to the interval between 0.5 and 1.5 fundamental periods before c, has produced good results. Equation (10) expresses that the algorithm will initially look for transitions with maximum waveform consistency, but if the waveform consistency is low, pitch consistency will be optimized. The threshold $\rho_{\min}$ is set equal to 0.5.

Dynamic programming is used to recursively compute the candidate sequence minimizing (9). Let

$$C_{\min}(d) = \min_{e_1,\ldots,e_k<d} C(e_1,\ldots,e_k,d) \qquad (12)$$

with e_1 in the first fundamental period of the voiced subsequence. Then, for $c > d$

$$C_{\min}(c) = C_{\mathrm{l}}(c) + \min_{d \in D(c)} \left(C_{\min}(d) + C_{\mathrm{t}}(c|d) \right). \qquad (13)$$

The predecessor $P(c)$ of c is the candidate d for which the minimum in (13) is reached. It is stored for each candidate c. The optimal sequence of pitch markers is found by backtracking from the candidate c_M in the last fundamental period of the voiced subsequence with minimum total cost.

3. RESULTS

Results are presented for two characteristic properties: (1) the marker's position in a fundamental period is near one of the higher maxima of the absolute value of the signal and (2) it is near the moment of excitation according to the LPC model. The results are compared with those obtained with pitch marking without the requirements of waveform and pitch consistency.

For the first characteristic property (markers near maxima of the absolute signal), the cost function $C_{\mathrm{l}}(c)$ was defined as in (1), with $F(c) = |s_c|$. The candidates c for which $C_{\mathrm{l}}(c) = 0$, i.e. the candidates at the locations of the maxima within a fundamental period, were selected as the pitch markers without consistency requirements.

For the second characteristic property (markers near moments of LPC excitation), the approach for epoch detection described in [2] was adopted. In [2] it is argued that the maxima of the Frobenius norm $\|S(k)\|_{\mathrm{F}}$ of a sliding data matrix

$$S(k) = \begin{pmatrix} s_k & s_{k-1} & \cdots & s_{k-p} \\ s_{k+1} & s_k & \cdots & s_{k-p+1} \\ \vdots & \vdots & \vdots & \vdots \\ s_{k+p} & s_{k+p-1} & \cdots & s_k \end{pmatrix} \qquad (14)$$

as a function of k are suitable pitch markers. The number p in (14) is the order of the assumed underlying LPC model. Here the common choice $p = \mathrm{round}(F_{\mathrm{s}}/1000) + 4$ is adopted. The Frobenius norm of a matrix is equal to the sum of the squares of its elements. For the matrix $S(k)$ given in (14), however, it can be computed efficiently as a function of k by filtering the squared signal samples with a triangular window of length $2p-1$. The cost function $C_{\mathrm{l}}(c)$ was defined as in (1), with $F(c) = \|S(c)\|_{\mathrm{F}}$. Again, the candidates c for which $C_{\mathrm{l}}(c) = 0$ were selected as the pitch markers without consistency requirements.

For both characteristic properties, the parameters for the calculation of the transition costs and for the dynamic programming were as given in Section 2. Pitch tracking was performed by an implementation of RAPT, such as described in [5], which produces F_0 estimates every 10 ms. Their reciprocal values were used to compute the fundamental periods for each sample position by linear interpolation.

Figure 1 shows the pitch markers obtained with and without consistency requirements for both characteristic properties in the Dutch word 'file', uttered by a female voice, $F_{\mathrm{s}} = 11025$ Hz. Clear differences between the pitch

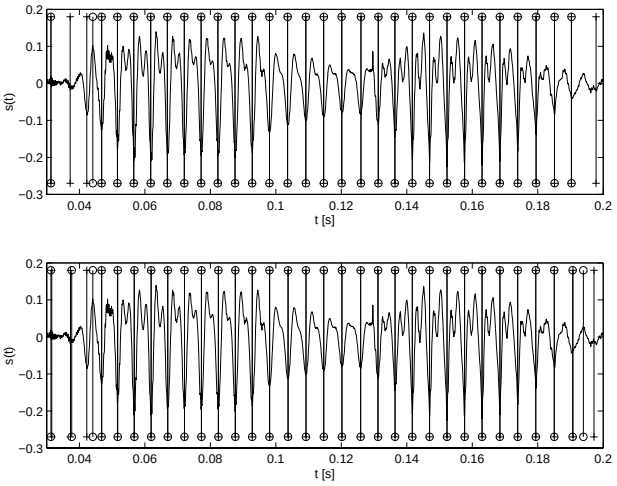

Figure 1: Pitch markers (vertical lines) in the Dutch word 'file' uttered by a female speaker. Results obtained with consistency requirements are marked with a '+', those obtained without with a 'o'. The upper panel shows pitch markers obtained from candidates near the maxima of the absolute sample values. The lower panel shows markers obtained from candidates near the maxima of the Frobenius norm.

markers obtained with and without the consistency requirements can be observed at the onset and the decay of the waveform, where the pitch markers obtained without consistency requirements are somewhat spuriously distributed and occasionally missing. The position of the pitch mark-

ers seems not to depend strongly on the characteristic property.

The thicker lines in Figure 1 indicate that two pitch markers are close but do not coincide. Figure 2 shows such a situation for markers obtained from candidates near the maxima of the absolute sample values. It can be seen that the waveform-consistency requirement has forced a marker away from the location of a local maximum of the absolute value to a position in which the local waveforms following adjacent pitch markers match better.

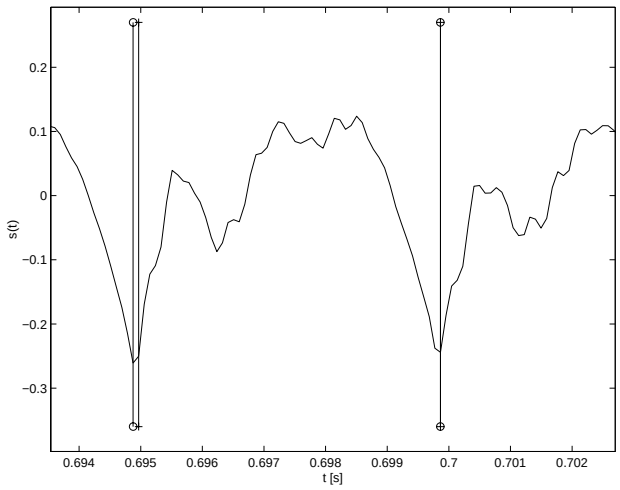

Figure 2: Detail showing pitch markers obtained from candidates selected near the maxima of the absolute sample values. Pitch markers computed with consistency requirements are marked with a '+', pitch markers computed without those requirements with a 'o'.

The pitch marking algorithm has been observed to be somewhat sensitive to larger errors, e.g. doubling or halving due to false octave jumps, in the fundamental periods $\{n_{0,k}|k = k_{\min}, \ldots, k_{\max}\}$. The occurrence of such errors depends on the pitch-tracking method that has been used. High-quality pitch tracking is therefore recommended.

4. CONCLUSION

An algorithm for pitch marking with requirements of waveform and pitch consistency has been introduced. It applies dynamic programming to find the optimal sequence of pitch markers given the consistency requirements. This algorithm thus avoids the typical errors that may occur with pitch marking without consistency requirements. It is flexible in the sense that it allows the user to specify the (application dependent) characteristic properties of the pitch markers.

The quality of pitch marking can only be assessed indirectly by assessing the performance of the application it is applied in. However, the results which have been presented indicate that the algorithm presented here indeed avoids certain errors that occur with unconstrained pitch marking, albeit that accurate pitch tracking is a prerequisite. Extensive results, evaluating the method in an application, will be presented in a forthcoming publication.

ACKNOWLEDGEMENT

The author would like to thank Maarten Holtrust for his preliminary investigations.

REFERENCES

[1] M. Balestri, A. Pacchiotti, S. Quazza, P. Salza, and S. Sandri. Choose the best to modify the least: a new generation of concatenative synthesis systems. In *Proceedings of Eurospeech 1999*, pages 2291–2294, Budapest, 1999.

[2] C. Ma, Y. Kamp, and L.F. Willems. A frobenius norm approach to glottal closure detection from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(2):258–265, April 1994.

[3] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467, 1990.

[4] Y. Stylianou. Synchronization of speech frames based on phase data with application to concatenative speech synthesis. In *Proceedings of Eurospeech 1999*, pages 2343–2346, Budapest, 1999.

[5] D. Talkin. A robust algorithm for pitch tracking. In W.B. Kleijn and K.K. Paliwal, editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier Science B.V., Amsterdam, 1995.

[6] D.Y. Wong, J.D. Markel, and A.H. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(4):350–355, August 1979.

[7] B. Yegnanarayana and R.N.J. Veldhuis. Extraction of vocal-tract system characteristics from speech signals. *IEEE Transactions on Speech and Audio Processing*, 6(4):313–327, July 1998.