MAXIMUM A POSTERIORI PITCH TRACKING

James Droppo[1] *and Alex Acero*[2]

[1] University of Washington, Seattle, Washington 98195, USA

[2] Microsoft Research, Redmond, Washington 98052, USA

ABSTRACT

A Maximum *a posteriori* framework for computing pitch tracks as well as voicing decisions is presented. The proposed algorithm consists of creating a time-pitch energy distribution based on *predictable energy* that improves on the normalized cross-correlation. A large database is used to evaluate the algorithm's performance against two standard solutions, using glottal closure instants (GCI) obtained from electroglottogram (EGG) signals as a reference. The new MAP algorithm exhibits higher pitch accuracy and better voiced/unvoiced discrimination.

1. INTRODUCTION

Accurate pitch estimates of speech are necessary for several applications, including speech coding, speech recognition, and prosody extraction. With such a wide range of interest, many researchers have worked on constructing pitch determination algorithms that are ideal for their applications.

A comprehensive study of pitch tracking is presented in [3]. As mentioned in [9], most pitch determination algorithms have three phases: (a) Preprocessing or signal conditioning, (b) Generation of pitch candidates and (c) Post-processing. The preprocessing, or signal conditioning, stage is included to minimize signal properties that are not germane to pitch tracking, such as the formant structure. Typically this can include some band-pass filtering, though it can include whitening the signal with an auto-regressive model and nonlinear techniques [8]. Generally pitch candidates are generated through a two-dimensional function *f(t, p)* of the time *t* and pitch period *p*, that assigns higher values to more likely pitch candidates at each time. It can be formulated in several domains, including time [6][9], autocorrelation [8][10], cepstrum [4], and ACOLS [2]. All of these techniques extract short segments of speech with explicit or implicit windows. The post-processing stage consists of taking the likely pitch candidates and choosing one for each input frame. The most common approach is to use dynamic programming to find an optimal pitch track. This phase will also decide if a region is voiced or unvoiced. Many of these phases are *ad hoc*.

This paper will describe a unified statistical framework to integrate the generation of pitch candidates with the post-processing required to obtain a continuous pitch track. The concept of predictable energy will be introduced as a function of the normalized cross-correlation. Enhancements, such as forward-backward pitch prediction and positive cross-correlation, will be shown to improve pitch accuracy over state-of-the-art systems.

Section 2 presents the basic statistical framework. Enhancements will be presented in Section 3, and the calculation of the voicing decision in Section 4. Discussion and evaluation will then be presented in Section 5

2. BASIC FRAMEWORK

In this section we first introduce the concept of predictable energy distribution, then apply it to a set of frames to derive the pitch track using a MAP framework.

2.1 Predictable Energy Distribution

Given a signal $x[n]$, define $\mathbf{x}_t$ as a vector composed of N consecutive samples of $x[n]$ centered at time t:

$$\mathbf{x}_t = \left(x[t - N/2] \cdots x[t] \cdots x[t + N/2 - 1]\right) \quad (1)$$

If $x[n]$ is periodic with period P, then we can predict it from a past vector P samples in the past as:

$$\mathbf{x}_t = \rho \mathbf{x}_{t-P} + \mathbf{e}_t \quad (2)$$

where ρ is the prediction gain and $\mathbf{e}_t$ is a zero-mean Gaussian random vector with a standard deviation σ

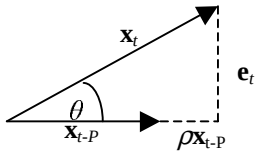

Figure 1. The prediction of $\mathbf{x}_t$ with $\mathbf{x}_{t-P}$ results in an error $\mathbf{e}_t$

We see from Fig. 1 that the minimum prediction error is

$$\left|\mathbf{e}_t\right|^2 = \left|\mathbf{x}_t\right|^2 - \left|\mathbf{x}_t\right|^2 \cos^2(\theta), \quad (3)$$

where

$$\alpha_t(P) = \cos(\theta) = \frac{< \mathbf{x}_t, \mathbf{x}_{t-P} >}{\left|\mathbf{x}_t\right|\left|\mathbf{x}_{t-P}\right|} \quad (4)$$

is the normalized cross-correlation between $\mathbf{x}_t$ and $\mathbf{x}_{t-P}$, having the property that $-1 < \alpha_t(P) < 1$. Therefore we can

model $f(\mathbf{x}_i \mid P)$ as a Gaussian density whose log-likelihood can be expressed as

$$\ln f(\mathbf{x}_t \mid P) = K + \frac{1}{2\sigma^2}\alpha_t^2(P)\left|\mathbf{x}_t\right|^2 \tag{5}$$

where K does not depend on the pitch period P. We define $E_t(P)$ as the predictable energy

$$E_t(P) = \alpha_t^2(P)\left|\mathbf{x}_t\right|^2 \tag{6}$$

because it tells us how much energy of the frame is predictable from the previous pitch period.

2.2 Maximum *a Posteriori* Pitch Tracking

Now let's define $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{M-1}\}$ as a sequence of input vectors for M consecutive frames centered at $t_i = iT$, a set of equally spaced time instants. Furthermore, if we assume that $\mathbf{x}_i$ are independent of each other, the joint distribution takes on the form:

$$f(\mathbf{X} \mid \mathbf{P}) = \prod_{i=0}^{M-1} f(\mathbf{x}_i \mid P_i) \tag{7}$$

where $\mathbf{P} = \{P_0, P_1, \ldots, P_{M-1}\}$ is the pitch track for the input.

The *maximum a posteriori* (MAP) estimate of the pitch track is:

$$\begin{aligned}\mathbf{P}_{MAP} &= \max f(\mathbf{P} \mid \mathbf{X}) = \max \frac{f(\mathbf{P}) f(\mathbf{X} \mid \mathbf{P})}{f(\mathbf{X})} \\ &= \max f(\mathbf{P}) f(\mathbf{X} \mid \mathbf{P})\end{aligned} \tag{8}$$

according to Bayes rule, with the term $f(\mathbf{X} \mid \mathbf{P})$ being given by (5) and (7).

2.3 *A Priori* Pitch Statistics

We can approximate $f(\mathbf{P})$ by

$$\begin{aligned}f(\mathbf{P}) &= f(P_0, P_1, \ldots, P_{T-1}) \\ &= f(P_{T-1} \mid P_{T-2}) f(P_{T-2} \mid P_{T-3}) \cdots f(P_1 \mid P_0) f(P_0)\end{aligned} \tag{9}$$

by assuming that the a priori probability of the pitch period at frame i depends only on the pitch period for the previous frame.

One possible choice for $f(P_t \mid P_{t-1})$ is a Gaussian distribution

$$\ln f(P_t \mid P_{t-1}) = K' - \frac{(P_t - P_{t-1})^2}{2\gamma^2} \tag{10}$$

so that when equations (5), (6) and (10) are combined, the log probability of transitioning from pitch P_{i-1} to P_i is

$$S_i(P_i, P_{i-1}) = E_i(P_i) - \lambda(P_i - P_{i-1})^2 . \tag{11}$$

with $\lambda = \sigma^2 / \gamma^2$, so that the log-likelihood in (8) can be expressed as

$$2\sigma^2 \ln f(\mathbf{P}) f(\mathbf{X} \mid \mathbf{P}) = E_0(P_0) + \sum_{i=1}^{M-1} S_i(P_i, P_{i-1}) \tag{12}$$

which can be maximized through dynamic programming. The term $\lambda(P_i - P_{i-1})^2$ acts as a penalty that keeps the pitch track from jumping around. Pruning can be done during the search without loss of accuracy. The value λ is empirically chosen.

3. ENHANCEMENTS

The normalized cross-correlation in (4) can be expressed as:

$$\alpha_t(P) = \frac{\sum_{n=-N/2}^{N/2-1} x[t+n]x[t+n-P]}{\sqrt{\sum_{n=-N/2}^{N/2-1} x^2[t+n] \sum_{n=-N/2}^{N/2-1} x^2[t+n+P]}} \tag{13}$$

There are several easy modifications that can be made to (13) so it will better suit our purpose. These modifications are outlined in the following sections.

3.1 Band-Pass Filtering

To improve the voiced/unvoiced discrimination, our signal preprocessing consisted in band-pass filtering between 100-2000 Hz. This eliminates low frequency noise that adversely affects the pitch estimate, as well as diminishes the energy of unvoiced fricatives, which do not carry relevant pitch information, and helps in the voicing decision

3.2 Variable Frame Length

The most obvious drawback to using (13) to compute the normalized cross-correlation is that one must choose the constant window length N before computation begins. Since we are interested in only predicting a single pitch period, we can replace the constant N with the period P. Making the window size variable, and as short as possible, gives us maximum time resolution.

In practice, it is useful to define a minimum useful frame length. With a window length of less than about 5ms, short signal segments can be similar merely because they are short, not because they are similar.

The idea of using a variable frame length in a pitch estimator is not new. One was formulated in [6] that required the use of three periods' data to achieve the published results. Our predictable energy distribution uses only two periods' data, and is a close relative of the cone-kernel time-frequency representation [5][11].

3.3 Positive Cross-Correlation

Using (13), it is possible that $\alpha_t(P) < 0$. In that case, there is negative correlation between $\mathbf{x}_t$ and $\mathbf{x}_{t-P}$, and it is unlikely

that P is a good choice for pitch. To account for this, we modify (13) as follows

$$\alpha_t'(P) = \max(\alpha_t(P), 0) \tag{14}$$

3.4 Forward-Backward Prediction

It can be troublesome to predict a pitch period across a spectral discontinuity. That is, although the glottis is vibrating at a regular interval, the vocal tract makes a sharp transition, so that the period to the left of the transition bears only a slight resemblance to the period on the right.

To overcome this problem, the distribution is modified to become symmetrical in time. At a spectral discontinuity, both the forward and backward predictable energies are computed. The distribution takes on the value of the better fit:

$$\alpha_t'(P) = \max(\alpha_t(P), \alpha_t(-P), 0) \tag{15}$$

3.5 Sub-harmonic Suppression

If the speech signal is periodic with pitch period P around time t, it is likely that the distribution will also have a large value at (t, 2P) and other sub-harmonics of P. To combat this problem, we transform the distribution in such a way as to decrease peaks that occur at sub-harmonics of other peaks.

$$\alpha_t''(P) = \alpha_t'(P) - \beta\alpha_t'(P/2) \tag{16}$$

The constant, β, is chosen such that $0 < \beta < 1$. A value of 0.2 is large enough to suppress the sub-harmonics, but small enough so that strong formants to not eliminate the true pitch period.

3.6 Logarithmic Sampling

It is possible to calculate the time-pitch predictable energy distribution for every discrete value of P. Fortunately, we do not need this precision for longer pitch periods. Pitch contours, like musical tones, are not perceived in the time domain, but in terms of relative pitches. As a result, we calculate the time-pitch predictable energy along contours logarithmically spaced in pitch. When the algorithm is modified to sample the energy distribution at a resolution of ¼ semitone, there is a significant processing savings, with no significant decrease in accuracy.

4. VOICING DECISION

A second DP pass takes the pitch track $\mathbf{P}_{MAP}$ in (8) and determines whether or not each frame is voiced or unvoiced and its corresponding probability.

To do this, we define a two-state HMM, one for voiced frames and the other for unvoiced frames. Each state is modeled as a single Gaussian density function of a two-dimensional vector composed by the frame energy E and its modified cross-correlation α"(P_{MAP}) in (16). Voiced frames tend to have high E and high α". Unvoiced frames have medium or low E, but usually have low α". Means, variances and mixture weights are estimated on a sentence-by-sentence basis using the EM algorithm on the unlabeled data.

We empirically determined the transition probability, which acts as a penalty to ensure that both voiced and unvoiced regions are not too short, and to impose continuity.

5. DISCUSSION AND EVALUATION

A popular approach to compute the pitch candidates for a given time is based on the local autocorrelation of the signal. Figure 2a is a plot of the positive half of the autocorrelation function of a short segment of speech. Although, theoretically, fine pitch resolution is possible, the autocorrelation function suffers the same drawbacks as other spectral techniques [2][4][8] — a long frame length can not yield fine time resolution. In addition, the use of long time windows results in a poor score in non-stationary regions.

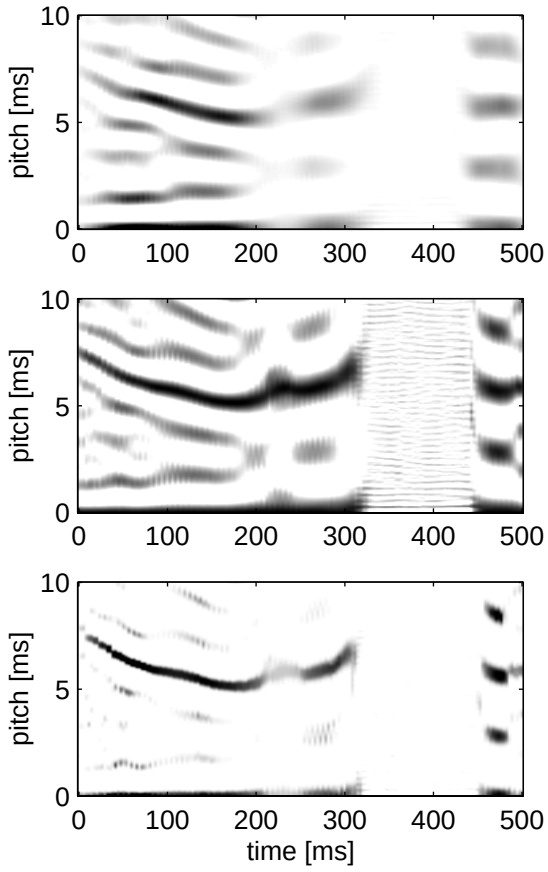

Figure 1. Energy distributions. (a) autocorrelation, (b) normalized cross-correlation, (c) predictable energy.

The normalized cross-correlation (Figure 2b) has some advantages over the correlation function. Although it still has a fixed window size, it gives better time and frequency resolution than the autocorrelation function. A variant of this function was used as part of a pitch estimator in [9]. A problem of this approach is that it cannot easily discriminate between voiced and unvoiced regions since noise regions contribute to the pitch track as much as strong voiced regions. The MAP approach (Figure 2c) weighs the normalized cross-correlation

by the energy of the current frame. This results in better voiced-unvoiced discrimination and focuses more on frames with high energy.

5.1 Evaluation

Rabiner *et. al*. [7] suggested several computable statistics for measuring the relative performance of pitch estimators. Such formal evaluation is, however, seldom used in the literature. In his paper, Rabiner defined two useful statistics: the number of "gross errors" and a mean and variance of the "fine errors". Rabiner also suggested that a count of both voiced to unvoiced and unvoiced to voiced errors would uncover the performance of the voicing decision.

The pitch estimation algorithm was tested on a database generated in-house by Microsoft Research. The database consisted of two hundred sentences each from one female and one male speaker. Each sentence consisted of an audio channel and an EGG channel. A gold standard of "true pitch" was extracted from the EGG channel of each sentence [1]. First, glottal closure instants (GCI) were extracted, giving a very accurate pitch-synchronous pitch estimate, which is easily converted to a time-synchronous estimate.

The new algorithm was compared to two optimized, standard time-domain algorithms. The first was the autocorrelation based pitch estimator based on a maximum-likelihood pitch estimate [10]. The second [9] has been implemented as a utility for extracting the fundamental frequency from speech included with Entropic Software's `xwaves`. We consider this solution to be "state of the art."

	Mark	Melanie
Maximum likelihood	0.46%	1.08%
`xwaves`	0.34%	0.74%
MAP Pitch	0.23%	0.27%

Table 1. Standard deviation of relative pitch errors.

Table 1 shows the standard deviation of relative pitch errors for the three algorithms, in which the proposed algorithm compares favorably. The algorithms were also compared in terms of the total percentage of voicing decision errors (see Table 2). Both `xwaves` and the MAP pitch estimation have lower error rates than the maximum likelihood algorithm.

	Mark	Melanie
Maximum likelihood	13.2%	20.8%
`xwaves`	8.0%	10.7%
MAP Pitch	7.2%	9.6%

Table 2. Percentage of voicing errors (voiced to unvoiced plus unvoiced to voiced).

Variable frame length, forward-backward correlation, and sub-harmonic suppression were all significant factors in reducing the pitch errors in the proposed approach. Still, it has been our observation that pitch period tracking errors do not come from missing the mark by a sample or two, but rather from choosing a sub-harmonic (e.g. pitch doubling) of the true pitch period of a formant. So pitch tracking remains a difficult challenge.

6. SUMMARY

This paper has introduced a MAP pitch estimation algorithm that integrates pitch candidate generation with dynamic programming for pitch tracking. Predictable energy is introduced as the current energy times the normalized cross-correlation. Improvements over the standard cross-correlation include forward-backward prediction, variable frame length and sub-harmonic suppression. The proposed algorithm has fewer voicing errors and higher pitch accuracy than state-of-the-art pitch trackers.

REFERENCES

[1] Acero A. "Source Filter Models for Time-Scale Pitch-Scale Modification of Speech". *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, Seattle, pp. 881-884, 1998.

[2] Kunieda N., Shimamura T., and Suzuki J. "Robust Method of Measurement of Fundamental Frequency by ACOLS-autocorrelation of log Spectrum". *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, pp. 232-235, May 1996.

[3] Hess W. *Pitch Determination of Speech Signals*. Springer-Verlag, New York, 1983.

[4] Noll A. M. "Cepstrum Pitch Determination". *Journal of the Acoustical Society of America*, vol. 14, pp. 293-309, 1967.

[5] Pitton J. W. and Atlas L. E., "Discrete-Time Implementation of the Cone-Kernel Time-Frequency Representation". *IEEE Transactions on Signal Processing*, vol. 43, no. 8, pp. 1996-8, Aug. 1995.

[6] Qian X. and Kimaresan R. "A Variable Frame Pitch Estimator and Test Results". *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, Atlanta, GA, pp. 228-231, May 1996.

[7] Rabiner L. R., Cheng M. J., Rosenberg A. E., and McGonegal C. A. "A Comparative Performance Study of Several Pitch Detection Algorithms". *IEEE Transactions on ASSP*, vol. 24, pp. 399-417, Oct. 1976.

[8] Rabiner L. R. "On the Use of Autocorrelation Analysis for Pitch Detection". *IEEE Transactions on ASSP*, vol. 25, pp. 24-33, 1977.

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

[10] Wise J. D., Caprio J. R. and Parks T. W. "Maximum-Likelihood Pitch Estimation". *IEEE Transactions of ASSP*, vol. 24, pp. 418-423, Oct. 1976.

[11] Zhao Y., Atlas L. E. and Marks R. J., "The Use of Cone-Shaped Kernels for Generalized Time-Frequency Representations of Non-stationary Signals". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 1084-91, July 1990.