# ROBUST PITCH TRACKING
# FOR PROSODIC MODELING IN TELEPHONE SPEECH

*Chao Wang and Stephanie Seneff*

Spoken Language Systems Group, Laboratory for Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139 USA
http://www.sls.lcs.mit.edu, email:{wangc,seneff}@sls.lcs.mit.edu

## ABSTRACT

In this paper, we introduce a pitch detection algorithm that is particularly robust for telephone speech and prosodic modeling. The algorithm uses a *logarithmically* sampled spectral representation of speech, similar to that in the subharmonic summation approach [2]. Constraints for $\log F_0$ and $\Delta \log F_0$ are combined in a dynamic programming search to find an optimum pitch track. The search algorithm is able to find a *continuous* pitch contour regardless of the voicing status, while a separate voicing decision module computes a probability of voicing per frame. We evaluated the algorithm using the Keele pitch extraction reference database [4] under both studio and telephone conditions. Our algorithm is very robust to channel degradation, and compares favorably to XWAVES under telephone conditions. It also significantly outperforms XWAVES when used for tone classification on a telephone quality Mandarin digit corpus.

## 1. INTRODUCTION

Reliable pitch detection is very crucial to the analysis and modeling of speech prosody. The fundamental frequency ($F_0$) is found to be highly correlated with prosodic features such as lexical stress, tone, and sentence intonation, which provide important perceptual cues to human speech communication. However, most current automatic speech recognition and understanding (ASRU) systems under-utilize prosodic features, especially those related to $F_0$. Besides the fact that speech prosody is a highly complex phenomenon, this is also partially due to the lack of a robust parameter space for statistical modeling. More specifically, errors in $F0$ contours, both in terms of pitch accuracy and voicing decision, can affect feature measurements dramatically.

Various pitch detection algorithms (PDAs) have been developed in the past [3]. While some have very high accuracy for *voiced* pitch hypotheses, the error rate considering voicing decision is still quite high; and the performance degrades significantly as the signal condition deteriorates. We are interested in developing a PDA that is particularly robust for telephone quality speech and prosodic modeling in ASRU applications. Pitch extraction for telephone speech is an especially difficult problem, due to the fact that the

fundamental is often weak or missing, and the signal to noise ratio is usually low. To deal with discontinuity of the $F0$ space for prosodic modeling, we believe that it is more advantageous to emit an $F_0$ value for each frame, even in unvoiced regions, and to provide separately a parameter that reflects probability of voicing. This is based on the considerations that, first, voicing decision errors will not be manifested as absent pitch values; second, features such as those describing the shape of the pitch contour are more robust to segmental misalignments; and third, a voicing probability is more appropriate than a "hard" decision of 0 and 1, when used in statistical models.

Our PDA is based on the frequency-domain analysis of the speech signal, namely, the *discrete logarithmic Fourier transform* (DLFT) as introduced in [7]. To address the problem of "missing fundamental" for telephone speech, we try to rely on the overall harmonic structure to make a decision. Our approach is similar to the *subharmonic summation* algorithm [2] in that the spectrum is also *logarithmically* spaced. However, both the signal processing and the tracking algorithms are quite different.

When the DLFT based PDA was first introduced in [7], it required an external source of voicing decision for pitch tracking. Since then, we have implemented a dynamic programming (DP) search module for continuous pitch tracking, and have added a voicing probability estimation module. In this paper, we give a detailed description of the new developments, and provide some formal evaluation results.

## 2. PITCH TRACKING

On a logarithmic frequency scale, harmonic peaks appear at $\log F_0$, $\log F_0 + \log 2$, $\log F_0 + \log 3$, ..., etc. To find the $F_0$ value, one can sum the spectral energy spaced by $\log 2, \log 3, ...,$ etc., from the pitch candidate and choose the maximum, as in [2]. This is equivalent to correlating the spectrum with an $n$-pulse template, where $n$ is the number of included harmonics. As described in [7], we adopt a similar method to find $\log F_0$ for each frame, using a carefully constructed template in place of the pulse sequence. More importantly, we also utilize a reliable estimate of $\Delta \log F_0$ across two *adjacent* voiced frames of the speech by simple correlation. We now combine the two constraints in a DP search to find an overall optimum solution. In the following, we give a detailed description of the algorithm and some enhancements.

## 2.1. Two Correlation Functions

The constraints for $\log F_0$ and $\Delta \log F_0$ estimations are captured by two correlation functions.

The "template-frame" correlation function provides constraints for $\log F_0$ estimation by aligning the speech DLFT spectrum with a template, as shown in Equation 1. $T(n)$, the template, is the weighted DLFT spectrum of a Hamming-windowed impulse train of $200Hz$, and $X_t(n)$ is the $\mu law$ converted DLFT spectrum at the $t^{th}$ frame. The template is normalized to have unit energy, and the correlation is normalized by the signal energy. The bounds for the correlation, $[N_L, N_H]$, are determined by the $F_0$ range.

$$R_{TX_t}(n) = \frac{\sum_i T(i)X_t(i-n)}{\sqrt{\sum_i X_t(i)^2}} \quad (N_L < n < N_H) \qquad (1)$$

The position of the correlation maximum should correspond to the difference of $\log F_0$ between the signal and the template. However, as in all PDAs, frame based peak picking can not totally avoid the problem of pitch doubling and halving. The correlation function has a relatively high peak when the harmonic lobes of the template align with $2F_0, 4F_0, 6F_0, ..., etc.$, of the signal spectrum, especially when the fundamental is missing. To reduce the tendency for pitch doubling, we added negative lobes between the positive lobes in the template, so that such an alignment will be penalized by the negative contributions from the $3F_0, 5F_0, ...$ peaks. The weighting of negative lobes was optimized empirically to be 0.35.

The "cross-frame" correlation function provides constraints for $\Delta \log F_0$ by aligning two adjacent frames of the signal DLFT spectra, as shown in Equation 2. The correlation is normalized by the energy of both signal frames. Because $F_0$ should not change dramatically across two frames, the correlation bound $N$ is set to be about 10% of the number of samples in the DLFT spectrum. The maximum of the correlation gives a robust estimation of the $\log F_0$ difference across two voiced frames.

$$R_{X_t X_{t-1}}(n) = \frac{\sum_i X_t(i)X_{t-1}(i-n)}{\sqrt{\sum_i X_t(i)^2}\sqrt{\sum_i X_{t-1}(i)^2}} \quad (|n| < N) \quad (2)$$

Figure 1 shows examples of the "template-frame" and "cross-frame" correlation functions in the voiced and unvoiced regions of a speech signal. For unvoiced regions, it is observed that the "template-frame" correlation is more or less random, and the "cross-frame" correlation stays fairly flat both within an unvoiced region and upon transition of voicing status. This has important implications for our DP based continuous pitch tracking algorithm.

## 2.2. Dynamic Programming Search

Given the two constraints as described previously, we can easily formulate the problem of pitch tracking as a DP search. We define the target function in an iterative manner as in Equation 3, where $i$ is the index in the "template-frame" correlation function. The pitch value can be converted from the index by inverting the logarithmic quantization. The search is extended by inheriting the best
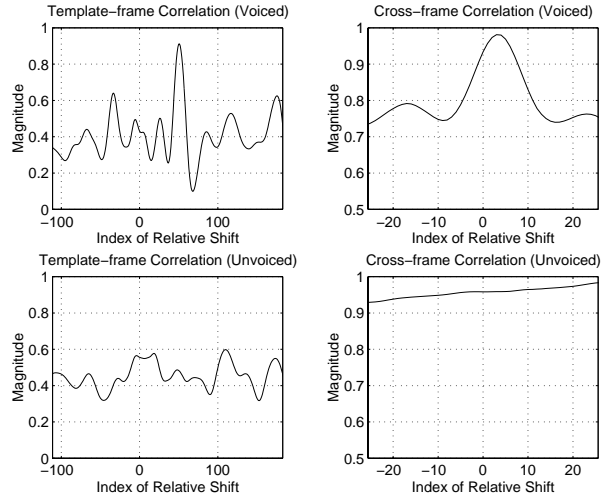


Figure 1: Examples of "template-frame" and "cross-frame" correlations for voiced and unvoiced DLFT spectra.

past score as *weighted* by the cross-frame correlation plus the template-frame correlation for the current node. The pointer to the best past node is saved for back tracking upon arriving at the last frame. Due to the logarithmic sampling of the DLFT, the search space for pitch value is naturally quantized logarithmically, with constant $\Delta F_0/F_0$.

$$score_t(i) = \begin{cases} \max_j\{score_{t-1}(j) \cdot R_{X_t X_{t-1}}(i-j)\} \\ \quad +R_{TX_t}(i) \quad (t > 0) \\ R_{TX_0}(i) \quad\quad\quad (t = 0) \end{cases} \quad (3)$$

The target function ensures a very smooth pitch contour. An expansion of Equation 3 reveals that the internal score of a particular node on the path is weighted by a *series* of cross-frame weights from that node to the current node before contributing to the cumulative score. We also tried replacing the multiplication in Equation 3 with addition. This score function imposes constraints only across the neighboring frames. We obtained slight performance improvement in pitch accuracy, because the search is more flexible to follow abrupt changes in the pitch contour, such as those caused by glottalization. However, we think such sensitivity is less robust for prosodic modeling, and thus did not pursue it further.

The DP search is forced to find a pitch value for every frame, even in unvoiced regions. We experimented with adding a node for unvoiced state in the search and incorporating the voicing probability into the target function. We found that this increased the number of pitch errors propagated from the voicing decision errors. It is observed that the cross-frame correlation stays relatively flat when at least one frame is unvoiced. Thus, upon transition into unvoiced regions, the best past score will be inherited by all nodes; and the scores become somewhat random. However, once in voiced regions, the sequence of nodes corresponding to the true pitch values will emerge because of high internal scores enhanced by high cross-frame correlation coefficients.

Figure 2 shows the waveform, DLFT spectrogram, and phonetic and word transcriptions for a telephone utterance.
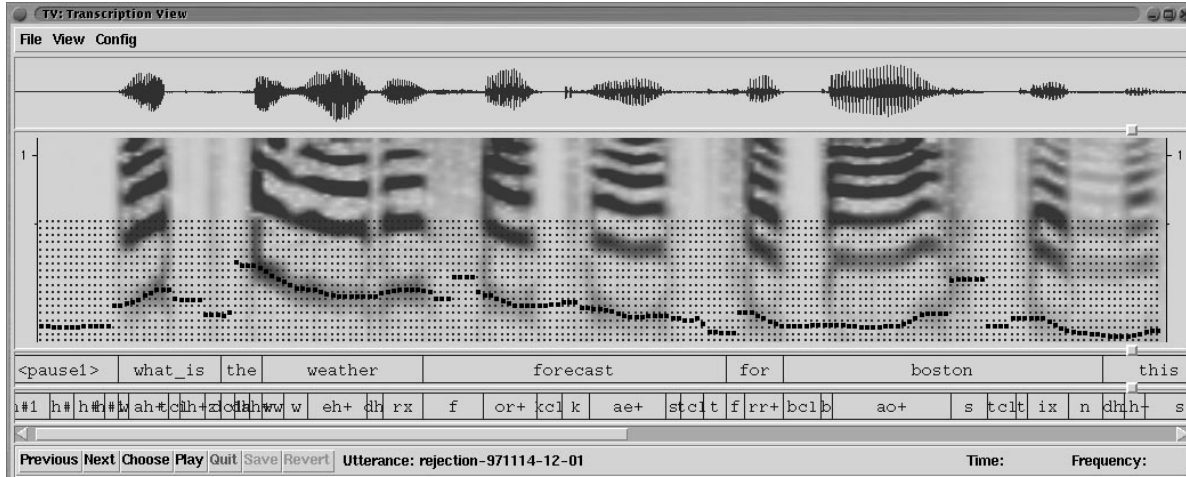
Figure 2: Waveform, DLFT spectrogram and transcriptions for the utterance "What is the weather forecast for Boston this ...". Part of the quantized search space for $F_0$ and the chosen path are overlayed with the DLFT spectrogram.

The DLFT spectrum is computed in the $[150, 1200]\, Hz$ range. The search space for $F_0$ is from $50Hz$ to $550Hz$, part of which is overlayed with the DLFT spectrogram. As shown in the figure, the first harmonic of the spectrum is fairly weak; nevertheless, the DP search is able to track $F_0$ whenever there is clear harmonic structure. The pitch track in unvoiced regions is arbitrarily chosen by the search and probably does not have a meaningful interpretation.

## 3. VOICING PROBABILITY ESTIMATION

To increase robustness for statistical modeling, the voicing decision module computes a voicing probability for each frame instead of making a hard decision.

Given an observation $\vec{O}$, we can obtain the posterior probabilities by applying Bayesian Rules as shown in Equation 4, where $V$ stands for voiced, and $U$ for unvoiced. $P(V)$, $P(U)$, $P(\vec{O}|V)$ and $P(\vec{O}|U)$ can be obtained *a priori* from training data.

$$\begin{cases} P_V = P(V|\vec{O}) & = P(\vec{O}|V)P(V)/P(\vec{O}) \\ P_U = P(U|\vec{O}) & = P(\vec{O}|U)P(U)/P(\vec{O}) \\ \quad P(\vec{O}) & = P(\vec{O}|U)P(U) + P(\vec{O}|V)P(V) \end{cases} \quad (4)$$

The observation vector has two elements. One is the maximum of the unnormalized template-frame correlation, which can be interpreted as the "periodic energy" of the signal. The second element is the minimum of the cross-frame correlation. It is small for voiced frames and close to 1 for unvoiced frames. We use the minimum of the forward and the backward cross-frame correlations to improve the prediction for the first and last frames of voiced regions, following the example in [1]. Mixtures of diagonal Gaussian models were used to model the prior distribution.

## 4. EVALUATION

A PDA is usually evaluated on two aspects: pitch estimation and voicing decision [5]. Accuracy for voiced pitch estimation can be evaluated in terms of "gross error" rate (GER), which is the percentage of voiced hypotheses that deviate from the reference by a certain amount (often 10% or 20%), and the mean and variance of the absolute value of the error. The voicing decision can be evaluated by the sum of voiced to unvoiced and unvoiced to voiced errors. Since our PDA does not make a hard decision for voicing, we will focus the evaluation on voiced frames. Our final goal is to apply our PDA in prosodic modeling. In this regard, we also evaluated telephone quality Mandarin tone classification performance using the PDA for pitch tracking.

We compared the performance of the DLFT based PDA with an optimized PDA provided by XWAVES [6] in these aspects. To ensure similarity, both PDAs are set to have an $F_0$ range of $50Hz - 550Hz$, and a frame rate of $100Hz$. The default is used for all internal parameters of XWAVES.

### 4.1. Voiced Pitch Accuracy

We use the Keele pitch extraction reference database [4] for this evaluation, because it provides reference pitch obtained from a simultaneously recorded laryngograph trace as "ground truth". Pitch values are provided at a $100Hz$ frame rate, with zero used for clearly unvoiced frames, and negative values used for uncertain frames (refer to [4] for a detailed description). There are five male and five female speakers, each speaking a short story of about 35 seconds. The Keele database is studio quality, sampled at $20KHz$. In order to evaluate the PDAs under telephone conditions, we transmitted the waveforms through a noisy telephone channel and recorded at a sampling rate of $8KHz$. We carefully calibrated the transmitted waveforms with the original, so that the pitch reference is still valid. Since we do not have other verified data to optimize the parameters of our PDA, we set aside two speakers (f1 and m1) as our development data, and tested on the remaining eight speakers. After optimization, the same parameters are used for all experiments including Mandarin tone classification.

We used only the "clearly voiced" frames in the Keele

| Configuration | | XWAVES:V | | | XWAVES:UV | | Overall(%) |
|---|---|---|---|---|---|---|---|
| | | GER(%) | Mean(Hz) | Std.(Hz) | V→UV(%) | GER(%) | |
| Studio | XWAVES | 1.74 | 3.81 | 15.52 | 6.63 | - | 8.37 |
| | DLFT | 3.24 | 4.61 | 15.58 | - | 1.01 | 4.25 |
| Telephone | XWAVES | 2.56 | 6.12 | 25.10 | 20.84 | - | 23.41 |
| | DLFT | 2.10 | 4.49 | 14.35 | - | 2.24 | 4.34 |

Table 1: Summary of performance on "clearly voiced" reference frames. Under each signal condition, the *voiced* reference data are divided into two subsets according to whether XWAVES determines them to be voiced, i.e., XWAVES:V and XWAVES:UV. All percentages are with reference to the total number of "clearly voiced" frames.

database for evaluation. Since XWAVES makes both gross errors and voicing errors, we divide the data into two subsets based on the outcome of XWAVES' V/UV decision, and summarize the performance for each subset in Table 1. The table gives both 20% GER and mean and standard deviation on absolute errors. The overall performance counts a voicing error as equivalent to a 20% GER. While XWAVES performs well on studio speech, the performance degrades severely for telephone speech, particularly with regard to voicing decisions. As expected, our PDA is less accurate for the "XWAVES:V" subset under studio quality, because it ignores spectral information below $150Hz$, and favors a smooth contour. However, it performs substantially better than XWAVES on both subsets for telephone speech. Even if we adjust the parameters of XWAVES to bias for voiced decisions, such that it mislabels frames as unvoiced less than 1% of the time, the 20% GER on the "XWAVES:UV" subset is still three times as high as that obtained by our PDA.

### 4.2. Tone Classification Accuracy

We compared the tone classification performance using $F_0$ derived by our system and XWAVES on a telephone-quality, Mandarin digit corpus [7]. The $F_0$ contour for each utterance was normalized by its average and adjusted for a sentence level downshift. The same tone features as described in [7] were used. We conducted two sets of experiments with and without an additional probability of voicing feature.

As summarized in Table 2, the result using the DLFT system (d) is significantly better than that using XWAVES (a). We tried two approaches to dealing with the unvoiced frames when using XWAVES: (b) interpolate $F_0$ from the surrounding voiced frames, and (c) bias the V/UV decision threshold to greatly favor "voiced", followed by interpolation. As seen in the table, neither method was particularly successful.

### 5. SUMMARY AND DISCUSSION

In this paper, we have demonstrated that the DLFT based PDA is robust to signal degradation inherent in telephone speech. In fact, the overall GER for studio and telephone speech is nearly the same (4.25% vs. 4.34%). The benefit of using a logarithmically sampled spectrum is that signals with different $F_0$ can be aligned by simple *linear shifting*. By correlating the DLFT spectrum with a template, we can obtain a robust estimation of the pitch, even when the fundamental is missing. By correlating two adjacent frames of the DLFT spectra, we can obtain a very reliable estimation of $F_0$ change instead of an arbitrary smoothing function.

| Configuration | Error Rate w/o $P_v$ (%) | Error Rate w/ $P_v$ (%) |
|---|---|---|
| (a) XWAVES | 25.4 | 25.6 |
| (b) XWAVES (intp'd) | 24.1 | 23.6 |
| (c) XWAVES (biased) | 24.9 | 25.4 |
| (d) DLFT | 19.2 | 18.2 |

Table 2: Summary of tone classification performance.

We have observed that the DLFT based PDA generally performs better for female speech than for male speech. When $F_0$ is low, the template-frame correlation suffers from missing low harmonics, and the cross-frame correlation suffers from compact spacing of higher order harmonics. This can potentially be improved by using gender-dependent parameters, or by adaptive signal processing, such as a variable frequency range for the DLFT.

### 6. ACKNOWLEDGEMENT

### REFERENCES

[1] J. Droppo and A. Acero, "Maximum *a posteriori* pitch tracking," in *Proc. ICSLP'98*, pp. 943-946, 1998.

[2] D. Hermes, "Measurement of pitch by subharmonic summation," *The Journal of the Acoustic Society of America*, vol. 83, no. 1, pp. 257–273, 1988.

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

[4] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH'95*, Madrid, pp. 837–840, 1995.

[5] L. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on ASSP*, vol. 24, pp. 399–417, 1976.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and synthesis* (Elsevier, ed.), pp. 495–518, 1995.

[7] C. Wang and S. Seneff, "A study of tones and tempo in continuous mandarin digit strings and their application in telephone quality speech recognition," in *Proc. ICSLP'98*, Sydney, pp. 635–638, 1998.