Noise Robust Pitch Tracking by Subband Autocorrelation Classification

Byung Suk Lee[1] *and Daniel P. W. Ellis*[1,2]

[1]LabROSA, Columbia University, New York, NY 10027, USA
[2]International Computer Science Institute, Berkeley, CA 94704, USA
{bsl, dpwe}@ee.columbia.edu

Abstract

Pitch tracking algorithms have a long history in various applications such as speech coding and extracting information, as well as other domains such as bioacoustics and music signal processing. While autocorrelation is a useful technique for detecting periodicity, autocorrelation peaks suffer ambiguity, leading to the classic "octave error" in pitch tracking. Moreover, additive noise can affect autocorrelation in ways that are difficult to model. Instead of explicitly using the most obvious features of autocorrelation, we present a trained classifier-based approach which we call Subband Autocorrelation Classification (SAcC). A multi-layer perceptron classifier is trained on the principal components of the autocorrelations of subbands from an auditory filterbank. Training on bandlimited and noisy speech (processed to simulate a low-quality radio channel) leads to a great increase in performance over state-of-the-art algorithms, according to both the traditional GPE measure, and a proposed novel Pitch Tracking Error which more fully reflects the accuracy of both pitch extraction and voicing detection in a single measure.

Index Terms: speech, pitch tracking, machine learning, subband, autocorrelation, principal components

1. Introduction

Determining the fundamental period of voiced speech signals (hereafter, "pitch tracking") is important in a range of applications from speech coding through to speech and prosody recognition and speaker identification. However, high-accuracy pitch tracking is difficult because of the wide variability of periodic speech signals [1]. There are many speech phenomena that can make the true pitch hard to identify or even define.

When acoustic degradations such as frequency band limitation and additive noise are introduced, the problem becomes still more challenging. This work is motivated by the problem of identifying and recognizing speech signals in low-quality radio transmissions, which we simulate, based on measurements of a real narrow-FM radio channel.

Computational approaches to finding pitch of speech have been studied extensively. There are two basic approaches in finding the periodicity—time-domain methods which utilize autocorrelation-like operations [2, 3]; and frequency-domain methods that rely on Fourier transform-like operations [4]. While periodic signals have obvious features in these domains, they also exhibit some ambiguity, leading to the well-known "octave errors" and other phenomena. These can be ameliorated by post-processing methods, such as Hidden Markov models (HMMs) that impose sequential consistency [2].

This work was supported by the DARPA RATS program.

In this paper, we extend a pitch tracking system based on the autocorrelation of multiple subbands coming out of an auditory filterbank. However, rather than attempting to explicitly detect the peaks that indicate particular pitches, we train a classifier on the full autocorrelation pattern corresponding to a corpus of labeled training examples. Since these training examples can be processed to include noise and channel characteristics specific to particular conditions, it can be made much more accurate in difficult conditions than "generic" pitch tracking. We also propose a new metric that gives a balanced evaluation of both pitch estimation accuracy and voicing detection.

The next section describes an existing subband autocorrelation algorithm on which our approach, described in section 3, is based. The new pitch tracking performance measure is described in section 4, and the experimental setup and the results are described in section 5. Section 6 makes some observations about the new algorithm and concludes the paper.

2. Previous work

Autocorrelation has been a successful basis both for predicting human pitch perception [5, 6], and for machine pitch tracking. Wu, Wang, and Brown proposed a robust multi-pitch tracking algorithm (henceforth, the Wu algorithm) [2] that combines pitch peaks identified in per-subband autocorrelations, followed by HMM pitch tracking. Since this is the basis of our system, we now describe it in more detail.

The input audio signal $a[n]$ is expanded into $s = 48$ subband signals $x_l[n]$, $l = 1 \ldots 48$, using an auditory filterbank. The normalized autocorrelation A_l is calculated for each subband every 10 ms (where t indexes the analysis frame and τ is the autocorrelation lag):

$$A_l(t,\tau) = \frac{r_l(t,\tau)}{\sqrt{r_l(t,0)}\sqrt{r_l(t+\tau,0)}} \quad (1)$$

where

$$r_l(t,\tau) = \sum_{n=-N/2}^{N/2} x_l[t+n]x_l[t+n+\tau] \quad (2)$$

and the window length $N = 400$. The largest lag is also $\tau = 400$, i.e., down to 40 Hz fundamental at 16 kHz sampling rate.

As a first attempt to identify the lag corresponding to the pitch period, all lags with local maxima in the autocorrelation are inspected for each subband. A period is selected if the normalized autocorrelation maximum is greater than $\theta = 0.945$ [2]. (The original paper used a different criteria for high-frequency subbands, but we used this single criteria throughout.) Selected maxima from different subbands are combined into a single score by spreading each peak according to an empirical Laplacian fit, then averaging across all subbands. The

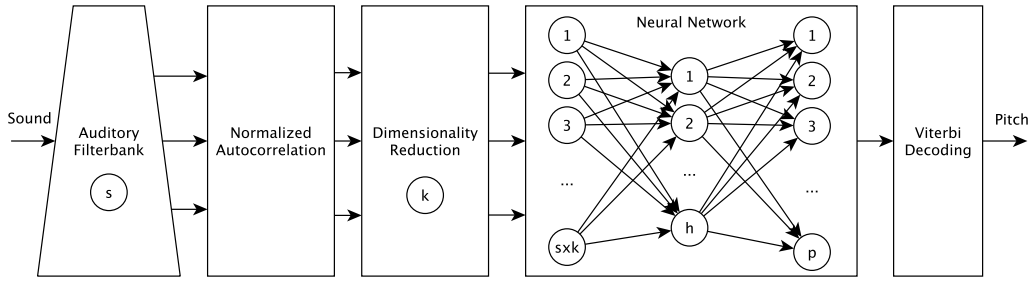

Figure 1: Diagram of the proposed Subband Autocorrelation Classification (SAcC) pitch tracking system.

result can be interpreted as the likelihood of the observations O_t at time t given a period hypothesis τ, i.e., $P(O_t|\tau)$

The Viterbi path through an HMM is used to smooth the pitch track, and to differentiate no-pitch and one-pitch states. The HMM finds the period sequence that maximizes the likelihood of the autocorrelation observations O_t by optimizing the sum across time of

$$P(O_t|\tau_t, \tau_{t-1}) = P(O_t|\tau_t)P(\tau_t|\tau_{t-1}) \quad (3)$$

where τ_t and τ_{t-1} are the pitches at frames t and $t-1$, and the transition probabilities $P(\tau_t|\tau_{t-1})$ are optimized empirically. $\tau_t = 0$ is a special case meaning no-pitch, whose probability is set to a fixed percentile of the real pitch probabilities.

Although our implementation of [2] differs from the original, it has performance essentially equivalent to the c-code released by the original authors for single-pitch conditions.

3. The SAcC Pitch Tracker

The diagram of the proposed Subband Autocorrelation Classsification (SAcC) pitch tracking system is shown in Fig. 1. The key change from the Wu algorithm is that the pitch period posterior is calculated by a single classifier working on the autocorrelations from all subbands, rather than explicit peak picking and cross-band integration. The modified stages are now described in more detail:

3.1. Subband PCA Dimensionality Reduction

Each subband autocorrelation $A_l(t, \cdot)$ is 400 points long; combining these across the $s = 48$ subbands would give an extremely large feature space. In fact, the normalized autocorrelation of each band-pass filtered signal $x_l[n]$ is highly constrained, leading to large redundancy. To simplify the classification problem, we reduce the dimensionality within each subband by applying Principal Component Analysis (PCA).

The principal components corresponding to the k largest eigenvalues were used to produce the subband k-dim PCA features $F_l(t, m)$ for each subband where $l = 1, \cdots, s$ is the subband index, and $m = 1, \cdots, k$ is the principal component index. We tried values for k in the range 5 to 20. The sorted eigenvalues of the PCA components decreased very fast, reflecting the redundancy in the autocorrelations.

3.2. MLP Classifier

The classifier for pitch candidates shown in Fig. 1 is a multilayer perceptron (MLP) trained using QuickNet[1]. The number of inputs to the MLP is $s \times k$. We used a single hidden layer with h hidden units, where h was varied between 50 and 800.

[1] http://www.icsi.berkeley.edu/Speech/qn.html

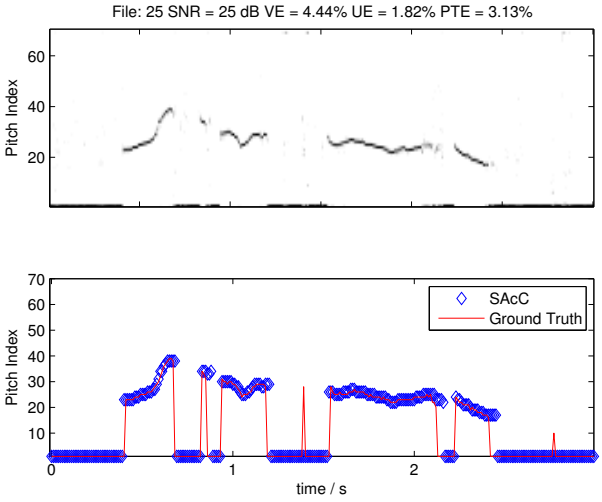

Figure 2: The MLP outputs $P(\tau|O_t)$ (top panel); and Viterbi tracking output of SAcC (blue diamond) and the ground truth (red line) on a sample speech corrupted with RBF and pink noise at 25dB SNR. (bottom panel)

The MLP had separate outputs for different pitch (period) values over a range which quantized 60 to 404 Hz using 24 bins per octave (in a logarithmic scale), a total of 67 bins. Each ground-truth pitch value in the training data was mapped to the nearest quantized pitch target. Any pitches outside this range were mapped to special "too low" and "too high" bins. Finally, an additional "no-pitch" target output accounted for unvoiced frames, giving $p = 70$ output units in total. To increase the range and volume of training data, each example was resampled at 8 rates from 0.6 to 1.6 and added to the training pool with a correspondingly-shifted ground truth pitch label.

The output of the MLP estimates the posterior probability of a pitch period given the observations, $P(\tau|O_t)$. Dividing by the pitch prior $P(\tau)$ gives a value proportional to $P(O_t|\tau)$ which can then be HMM (Viterbi) smoothed as in (3).

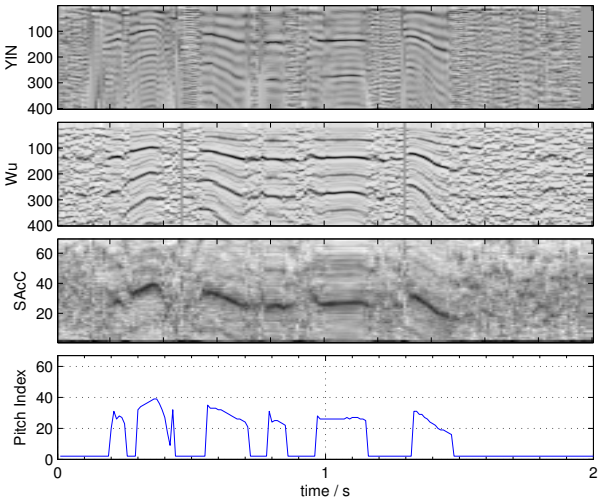

Figure 3: The observation pitch likelihood of YIN, Wu, and SAcC on a speech sample corrupted with RBF and pink noise at 25dB SNR. Note that the vertical axis is lag in samples (increasing downwards) for YIN and Wu, but quantized (log-frequency) pitch for SAcC. Also, for SAcC, we show $\log(P(\tau|O_t))$ to reveal detail in the incorrect pitch candidates.

4. Performance Metrics

The standard error measures for pitch tracking are Gross Pitch Error (GPE) and Voicing Decision Error (VDE) [7]:

$$\text{GPE} = \frac{E_{f_0}}{N_{vv}} \qquad \text{VDE} = \frac{E_{v\to u} + E_{u\to v}}{N} \tag{4}$$

where N is the total number of frames, N_{vv} is the count of frames in which both the pitch tracker and the ground truth reported a pitch, E_{f_0} counts the frames in which these pitches differ by some factor (typically 20%), $E_{v\to u}$ is the count of voiced frames misclassified as unvoiced, and $E_{u\to v}$ is the number of misclassified unvoiced frames. The problem with this measure is that GPE can be improved by labeling voiced frames whose period is ambiguous as unvoiced, thereby reducing the N_{vv} denominator. This will increase VDE, but it is difficult to compare overall performance with this pair of numbers.

We therefore propose a modified metric to evaluate pitch trackers which we call the Pitch Tracking Error (PTE). It is a simple average of Voiced Error (VE) and Unvoiced Error (UE):

$$\text{PTE} = \frac{\text{VE} + \text{UE}}{2} \tag{5}$$

$$\text{VE} = \frac{E_{f_0} + E_{v\to u}}{N_v} \qquad \text{UE} = \frac{E_{u\to v}}{N_u} \tag{6}$$

where N_v is the number of frames for which a pitch is reported in the ground truth, and $N_u = N - N_v$ is the remaining (unvoiced) frame count.

It is more transparent to compare VE, UE, and PTE between different pitch trackers because the denominators N_v and N_u do not change with the system. If we consider pitch tracking as detection, VE resembles miss rate, and UE is similar to false alarm rate. Another advantage of PTE is that it can balance the contribution of errors on voiced and unvoiced frames regardless of their proportion in the actual evaluation material. (Different weights for VE and UE could be considered for tasks where one kind of error was more important.)

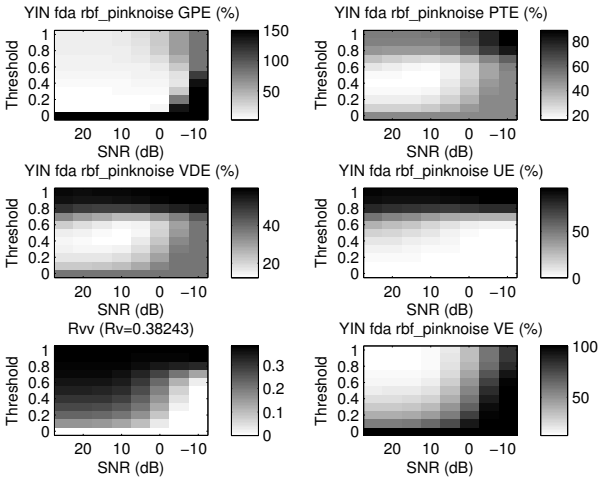

Figure 4: The GPE, PTE, VDE, UE, R_{vv}, and VE for YIN at various threshold and SNR points on FDA under RBF and pink noise condition.

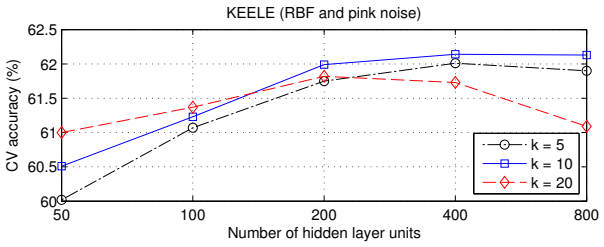

Figure 5: The Cross Validation (CV) accuracy of the MLP using the k-dimensional PCA feature.

5. Experiments

5.1. Data

We used the KEELE [8] and FDA [9] corpora for evaluation. The lengths of the datasets are 337 s and 332 s respectively. KEELE consists of 10 speakers each reading the same story for about 30 s; FDA has two speakers reading the same 50 short sentences of around 3 s each. Since KEELE includes greater variation, and to illustrate generalization, we chose to train on KEELE and report results on FDA. Since our interest is in pitch tracking that can be used on low-quality radio transmissions, our main experiment applied to both training and test material a simulated radio-band filter (RBF) modeled from a real recording made across a narrow-FM channel[2], which amounted to a bandpass spanning around 500 Hz to 2 kHz along with additive pink noise at various levels.

5.2. Experiment Setup

YIN [3], Wu [2], and SWIPE′ [10] algorithms are used for performance comparison. Both the ground truth and the pitch trackers gave pitch values for every 10 ms.

To use YIN and SWIPE′ as pitch trackers, the pitch strength outputs (aperiodicity for YIN and pitch strength for SWIPE′) are thresholded to provide voiced/unvoiced decisions. Fig. 4

[2] http://labrosa.ee.columbia.edu/projects/renoiser/

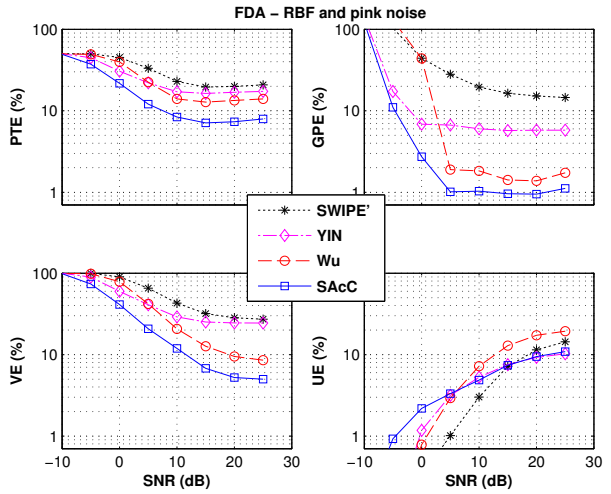

Figure 6: The PTE, GPE, VE, and UE for SAcC, Wu, YIN, and SWIPE′ on FDA under RBF and pink noise condition.

shows the GPE, PTE, VDE, UE, R_{vv}, and VE of YIN versus SNR for various thresholds for speech with RBF and pink noise, where $R_{vv} = N_{vv}/N$ and $R_v = N_v/N$. The threshold giving the best PTE was used in evaluation. For the Wu algorithm, the probability of no-pitch is searched over the 1^{st} to 90^{th} percentiles of the remaining pitch likelihoods to find the value that optimized PTE.

For the SAcC MLP, 66.7% of the data was used for training, with the rest used for cross validation (CV). Fig. 5 shows the CV accuracy as a function of k, the number of principal components retained, and h, the hidden layer size on the most challenging RBF case. From these results, we chose $k = 10$ and $h = 800$.

5.3. Results

Looking at the right column of Fig. 4, we see that lowering the threshold and thus reducing the proportion of voiced frames lowers UE (as all frames, including the unvoiced ones, are labeled unvoiced) while increasing VE. As the sum of these competing trends, PTE shows a clear optimum for a threshold around 0.4. In the left column, GPE appears to improve as the threshold decreases, but this hides the disappearing proportion of frames, N_{vv} (bottom pane), over which this measure is calculated. When $N_{vv} = 0$, an arbitrary high value (150%) is assigned to GPE to reflect that it is based on zero frames. VDE reveals an optimal threshold similar to PTE, but ignores actual pitch estimation errors.

The performance comparison of SAcC, YIN, Wu, and SWIPE′ on FDA dataset under the RBF plus pink noise condition is shown in Fig. 6. For SAcC, PTE is dominated by UE in the high SNR and VE in the low SNR. The result on FDA dataset under pink noise only (not shown) has similar trend: In low SNRs, pitch tracker outputs are mostly no-pitch, lowing UE and increasing VE. Note that PTE gives higher absolute values than GPE since it reflects both difficult voiced frames and voicing errors; we consider performance in these areas to be critical.

6. Discussion and Conclusion

The output of the SAcC MLP $P(\tau|O_t)$ on a sample speech is shown on the top pane of Fig. 2. The most likely pitch candidate for each frame has a significantly stronger value than the others. The HMM tracking result of SAcC on the same example is shown on the bottom panel in Fig. 2 along with the ground truth pitch. HMM tracking promotes continuous pitch tracks and discourages voicing transitions, which sometimes causes the extension of pitch tracks into unvoiced regions.

The observed pitch likelihood of YIN, Wu, and SAcC on another speech sample corrupted with RBF and pink noise at 25dB SNR is shown in Fig. 3. For SAcC, the log of the MLP output is shown to reveal details in the non-favorite candidates. Both YIN and Wu are based on autocorrelation operations, and have harmonic and subharmonic structures. Since SAcC is trained to discriminate between these otherwise ambiguous cases, it has one strong peak in most frames, reducing the likelihood of octave errors.

We have proposed a noise robust pitch tracking system, SAcC, based on subband autocorrelation classification. The proposed algorithm incorporates the learning power of an MLP classifier, the smooth tracking of an HMM, and the low dimensional representation of k-dimensional subband PCA. We have also proposed a performance metric, PTE, to give a balanced measure of performance in both voiced and unvoiced regions.

To simulate the target noise condition, a radioband filter was learned from real recorded samples and used in combination with additive pink noise to make a useful simulation of poor quality radio reception, the particular focus of our study. We believe, however, that the subband classification structure should be advantageous in many challenging acoustic conditions, particularly when matched training data is available.

The performance evaluation on KEELE and FDA datasets showed that SAcC improves the state-of-the-art for pitch tracking on this kind of data, particularly as measured by our PTE metric.

7. References

[1] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds., chapter 14. Elsevier Science B.V., 1995.

[2] M. Wu, D.L. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Tr. Speech and Audio Proc.*, vol. 11, no. 3, pp. 229–241, May 2003.

[3] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[4] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Tr. Audio, Speech, and Lang. Proc*, vol. 16, no. 2, pp. 255–266, February 2008.

[5] M. Slaney and R.F. Lyon, "A perceptual pitch detector," in *IEEE ICASSP*. IEEE, 1990, pp. 357–360.

[6] R. Meddis and M.J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Am.*, vol. 89, no. 6, pp. 2866–2882, 1991.

[7] W. Chu and A. Alwan, "Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *IEEE ICASSP*, 2009, pp. 3969–3972.

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

[9] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," in *EUROSPEECH*, September 1993, pp. 1003–1006.

[10] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, September 2008.