A Fine Pitch Model for Speech

Jasha Droppo, Alex Acero

Microsoft Research, Redmond, USA

Abstract

An accurate model for the structure of speech is essential to many speech processing applications, including speech enhancement, synthesis, recognition, and coding. This paper explores some deficiencies of standard harmonic methods of modeling voiced speech. In particular, they ignore the effect of fundamental frequency changing within an analysis frame, and the fact that the fundamental frequency is not a continuously varying parameter, but a side effect of a series of discrete events.

We present an alternative, time-series based framework for modeling the voicing structure of speech called the *fine pitch model*. By precisely modeling the voicing structure, it can more accurately account for the content in a voiced speech segment.
Index Terms: speech analysis, pitch estimation, fundamental frequency

1. Introduction

An accurate model for the structure of speech is essential to many speech processing applications, including speech enhancement, synthesis, recognition, and coding.

The premise underlying modern models for speech is that it can be decomposed into an excitation signal and a linear model of the vocal tract. Figure 1 presents this standard *mixed-excitation* model for speech. Under this model, speech signals are composed of a linear combination of a colored noise signal and a filtered sequence of impulses.

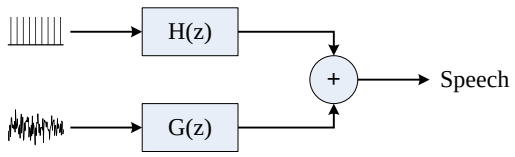

Figure 1: *The mixed excitation model for speech.*

Most of the energy in a voiced speech signal is generated by the repeated closing of the glottal folds within the larynx. This typically occurs at a rate as low as 60 Hz for men, and as high as 300 Hz for women[1], and is what imparts voiced speech with its fundamental frequency. It is this sequence of discrete glottal closure events that is modeled by the sequence of impulses in Figure 1.

Engineers have used this fundamental frequency to improve speech representations and processing for many years in both coding [2, 3], enhancement [4, 5, 6], and noise-robust automatic speech recognition [7, 8, 9].

Each of these techniques is built using frame-based spectral analysis. The speech signal is broken into overlapping segments, each between 25ms and 45ms long. Then, a single pitch is inferred for this segment based on either autocorrelation or spectral analysis.

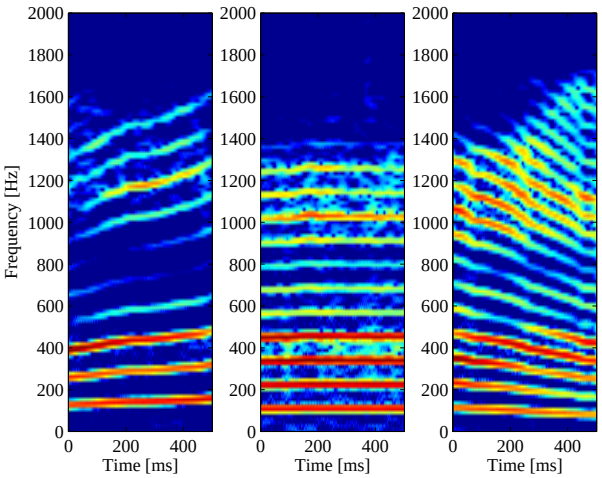

Figure 2: *Three narrowband spectrograms of a man's voice with rising, constant, and falling fundamental frequency.*

In this paper, we propose an alternative pitch analysis framework called the *fine pitch model* for speech. It operates time-synchronously with the digitized samples of the incoming speech signal, assigning a unique pitch for every sample. It is able to precisely track the changing periodic structure of speech and operates well even in low signal-to-noise environments.

Section 2 examines the core of the new algorithm, how it defines instantaneous pitch, and how it can be superior to previous frame-based harmonic methods. Section 3 describes an efficient algorithm for estimating a *fine pitch track*, as well practical considerations such as reasonable bounds on the range and precision of the estimate.

2. The Structure of Voiced Speech

Figure 2 shows three different narrowband spectrograms of a man uttering the same speech sound, `/r/`, with different fundamental frequency patterns. The most striking features in these spectrograms are the horizontal bands of energy that appear as harmonics of a fundamental frequency f_0 near 100 Hz.

2.1. The Harmonic Model of Voiced Speech

Under the assumptions that f_0 is perfectly constant, and that the vocal tract transfer function $H(z)$ doesn't change, and that the signal exists for all time, voiced speech is a truly periodic signal and can be represented as a trigonometric series:

$$x(t) = \sum_{k=1}^{K} a_k \cos(2\pi k f_0 t) + b_k \sin(2\pi k f_0 t) \qquad (1)$$

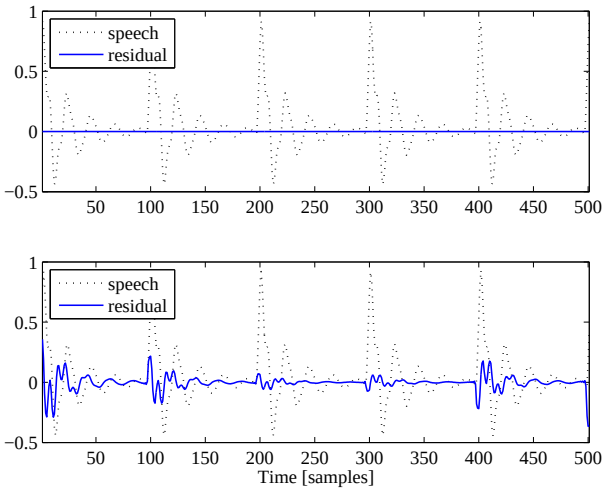

Figure 3: *Perfectly periodic synthetic speech has a zero residual (above), but if the pitch changes by as little 0.5 samples per pitch period (below), the residual is no longer perfect.*

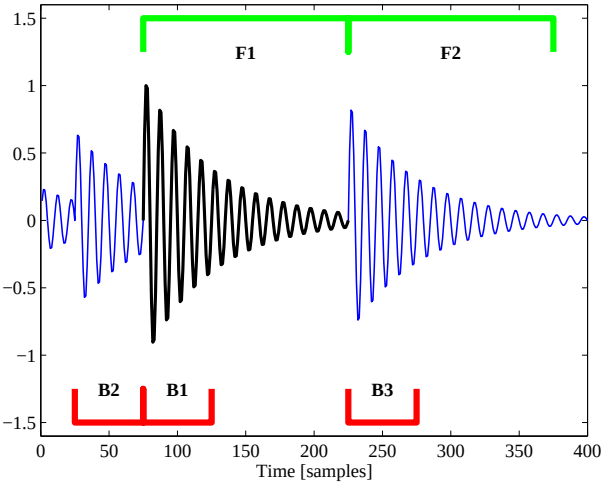

Figure 4: *A simulation of voiced speech with pitch period increasing over time. The instantaneous pitch value depends on whether the prediction is forward or backward in time, and is sometimes undefined.*

The ideal signal $x(t)$ has all of its energy concentrated at f_0 and its harmonics. It is very appealing to use this harmonic model for voiced speech, because once the fundamental frequency is known, the speech can be represented with only a handful of parameters.

Some algorithms use the harmonic model directly, such as in Laroche *et al.*[10] and Seltzer *et al.*[8]. Both estimate the voiced speech spectrum by projecting the measured speech spectrum onto the subspace of all possible harmonic spectra with a given fundamental frequency.

Another common use of the harmonic model is to select frequency bins from the measured spectrum, and label them as either voiced speech or noise. These techniques are described as using "peak picking" or "harmonic tunneling", and several good examples exist in the recent literature[7, 9, 5, 6, 4].

2.2. Problems with the Harmonic Model

Unfortunately for the harmonic model, voiced speech does not have a constant fundamental frequency.

Figure 3 shows how even slight variations in pitch can cause the model to fail. These speech signals were synthesized with an all-pole vocal tract model estiamted from the same data as in the middle section of Figure 2 and exciting it with a series of discrete pulses.

The speech in the top of Figure 3 has a constant fundamental period of 100 samples. When it is passed through a comb filter with a period of 100 samples, the harmonic model of Eq. 1 indicates that the residual signal should be zero. Observe that the plot of the residual has almost no energy. In other words, the harmonic model can account for all the energy in the signal.

The bottom half of Figure 3 is similar to to the top half, except the pitch period is increasing at a rate of 0.5 samples per pitch period. The length-100 comb filter is too long for the first periods, and too short for the last. The harmonic model is not able to account for all of the energy in the signal.

One way to solve this problem is to use a time-warping function on the incoming signal to enforce a constant pitch. This can produce quite accurate pitch estimates[11], but the constant pitch signal will contain some FM distortion as a result of the processing.

This paper solves the problem by constructing a model capable of describing the periodic nature of speech without a constant pitch requirement.

2.3. The Fine Pitch Model

Since the rate of glottal closures can change over time, it is useful to dismiss the notion of fundamental frequency in favor of *instantaneous pitch period*. Whereas the harmonic model estimates a constant fundamental frequency for each analysis frame, the fine pitch model assigns a sequence of instantaneous pitch periods that can be different for every sample of the speech signal.

Because of the nature of voiced speech generation, namely that it is generated from a sequence of discrete events, a sequence of instantaneous pitch periods should not have a slope. It should be piecewise constant, exhibiting a stair step pattern. According to the mixed excitation model, the instantaneous pitch should not change in between glottal closure events. In this paper, we refer to these regions as *epochs*.

The instantaneous pitch period for a segment of voiced speech is the shortest delay, either in the future or the past, between it and another substantially similar, non-overlapping segment of the signal. This leaves open the questions of how the segments are defined, and what "substantially similar" means. Depending on how these questions are answered, particular implementations of the fine pitch model will differ in accuracy and efficiency.

Figure 4 contains an exaggerated view of what happens as the pitch period changes during voiced speech.

The signal in region B1 is similar to a segment from the previous epoch (B2) and a segment from the future epoch (B3). The backward pitch period during B1 is the relative lag between B1 and B2, and exists for the entire current epoch. The backward pitch period during B3 is the relative lag between B1 and B3. The segment of signal between B1 and B3 does not have a backward pitch defined. This is always the case when pitch period is increasing: signal segments near the end of each epoch do not have a backward pitch defined.

The forward pitch period in the segment B2 is the relative lag between B1 and B2. The forward pitch period in the segment F1 is the relative lag between F1 and F2. Because the pitch period is increasing, the forward pitch is well defined everywhere.

Notice that the speech segment B1 overlaps with the speech segment F1. In this region, both a forward pitch period and a backward pitch period are defined, and that they are quite different.

3. Estimating the Fine Pitch Track

An ideal solution to the fine pitch model would segment the voiced speech into non-overlapping epochs that obey the forward and backward similarity measures discussed above.

A practical solution is to estimate a set of pitch lags, one for every sample of speech, that minimizes an objective function that approximates the true model. For this paper, we chose an objective function that balances how well the received signal sample $y[n]$ is predicted by the sample $y[n-\tau_n]$, with a measure of the smoothness of the sequence of pitch period estimates.

$$y_{\mathrm{r}}[n] = y[n] - y[n-\tau_n] \quad (2)$$

$$\mathcal{F} = \sum_n \left(y_{\mathrm{r}}[n]\right)^2 + \gamma \sum_n \left(\tau_n - \tau_{n-1}\right)^2 \quad (3)$$

If the instantaneous pitch period estimates τ_n match the true pitch periods, the time-varying comb filter of Eq. 2 will eliminate much of the voiced speech energy, leaving only a small residual $y_{\mathrm{r}}[n]$. The first term in Eq. 3 measures the energy of this residual. Since the time-varying comb filter is only good at eliminating signals that have a coherent pitch, minimizing this term is equivalent to finding and eliminating the voiced speech components.

The second term in Eq. 3 forces the algorithm to choose a pitch sequence that is mostly smooth over time. Ideally, we would choose a constraint that would force a piecewise-constant pitch sequence, but this first-order Markov assumption is much more efficient to compute. The parameter γ controls the relative importance of residual signal energy and smooth pitch sequence, and can be set empirically on a small set of data.

Because $\mathcal{F}$ is first-order Markov in τ_n, the optimum instantaneous pitch sequence can be found using standard dynamic programming search techniques.

First, a forward recursion is performed. Assume the cost associated with the best possible instantaneous pitch assignments up to time $t-1$ and ending at τ_{t-1} is available, and call that cost $q(\tau_{t-1}, t-1)$. The set of new costs $q(\tau_{t-1}, \tau_t, t)$ is found by adding the cost of starting in the previous pitch period value ($q(\tau_{t-1}, t-1)$), the cost of transitioning pitch period values ($(\tau_t - \tau_{t-1})^2$), and the cost of the new pitch period value ($(y[t] - y[t-\tau_t])^2$). Before repeating the recursion, only the best route to τ_t at time t is retained, as well as the identity of the transition corresponding to this route.

Second, a backward recursion is performed. At the end of the utterance, choose the instantaneous pitch value that has the lowest cost. Then, follow the retained "best transition" to find the previous instantaneous pitch. Repeat following the best transitions backward in time until eventually the entire file has been processed.

Figure 5 shows how the fine pitch model is much better at predicting the energy of voiced speech. The test signal is the same as the bottom half of Figure 3, with pitch increasing

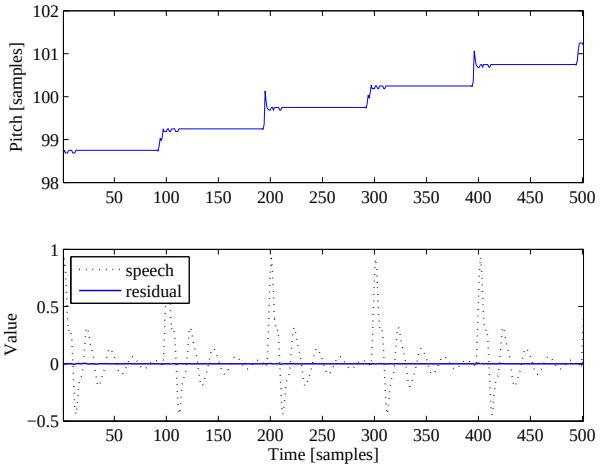

Figure 5: *The same signal from the bottom of Figure 3, modeled by the fine pitch model. Because the pitch is increasing (top), the backward looking pitch has problems near the epoch boundaries. The residual from the fine pitch model (bottom) is near zero.*

at the rate of about 0.5 samples per sample. In Figure 3, the model assumed a constant pitch and was unable to account for the observed speech. Figure 5 uses the fine pitch model, and leaves very little residual energy. Also note how the estimated pitch has the predicted stair step pattern.

Figure 6 demonstrates the result of applying the fine pitch model to real speech. The test signal is the same as used in the spectrogram with falling fundamental frequency from Figure 2. It is apparent from the small residual that the fine pitch model is able to account for most of the energy in the speech signal. The estimated pitch exhibits the same piecewise constant pattern as the simulated data, with an enticing twist: sometimes the measured pitch will be different between the first and second half of each epoch.

3.1. Practical Considerations

The precision of the instantaneous pitch estimate is very important, because even small rounding errors in the fundamental frequency estimate can become large errors in estimating the position of the highest harmonic. Luckily, although the original sampling rate limits the Nyquist bandwith, it is still possible to get sub-sample resolution on the relative phase of the signal components.

A deep voice may have a fundamental frequency near 50 Hz, and a 40th harmonic at 2 kHz. To get that harmonic accurately placed to within 5 Hz, the fundamental frequency should be accurate to within $\frac{1}{8}$ Hz. At an 8 kHz sampling rate, this corresponds to measuring a 160 sample pitch period to within 0.4 samples.

At the other end of typical fundamental frequencies, a high pitched voice can have a fundamental frequency of 400 Hz, with a 5th harmonic at 2 kHz. To place that harmonic to within 5 Hz, the fundamental frequency should be accurate to within 1.0 Hz. At an 8 kHz sampling rate, this corresponds to measuring a 20 sample period to within 0.05 samples.

In general, to get harmonics near 2 kHz accurate to within 5 Hz, a pitch period at n samples must be accurate to within approximately $n/400$ samples.

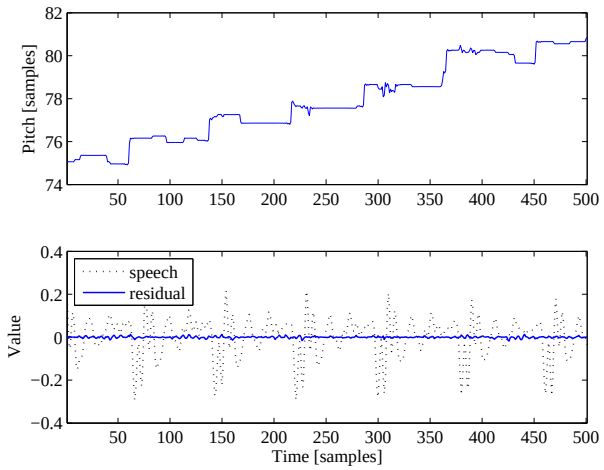

Figure 6: *The real speech signal from Figure 2 with a falling fundamental frequency. Even though the pitch is changing at about one sample per epoch, the fine pitch model is able to capture most of the signal's energy.*

With this variable-precision need in mind, the current system is implemented in two passes over the data. In the first pass, pitch is estimated to within one sample. This covers the first few harmonics accurately, but is not yet precise enough. In the second pass, a higher resolution pitch estimate is formed with the restriction that it can't vary by more than two samples from the original estimate. To accomplish this, the signal is up-sampled by a rate consistent with the extra precision needed to refine the first-pass pitch estimate.

4. Conclusions

This paper advocates replacing the popular harmonic models for speech with a time-synchronous fine pitch model. Preliminary experiments indicate that the FPM is able to precisely track the time-varying nature of voiced speech, which allows it to accurately represent more of the energy present in such signals.

5. References

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Prentice Hall PTR, 2001, ch. 2.

[2] M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (celp): High quality speech at very low bit rates," in *Proc. ICASSP*, Tampa, FL, March 1985.

[3] R. Ramachandran and P. Kabal, "Pitch predicition filters in speech coding," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 4, pp. 467–478, April 1989.

[4] A.-T. Yu and H.-C. Wang, "New speech harmonic structure measure and it application to post speech enhancement," in *Proc. ICASSP*, vol. I. IEEE, 2004, pp. 729–732.

[5] H.-G. Kim, M. Schwab, N. Moreau, and T. Sikora, "Speech enhancement of noisy speech using log-spectral amplitude estimator and harmonic tunneling," in *Proc. International Workshop on Acoustic Echo and Noise Control*, Kyoto, Japan, September 2003, pp. 119–122.

[6] J. Beh and H. Ko, "A novel spectral subtraction scheme for robust speech recognition: Spectral subtraction using spectral harmonics of speech," in *Proc. ICASSP*, vol. I. IEEE, 2003, pp. 648–651.

[7] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech," in *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[8] M. Seltzer, J. Droppo, and A. Acero, "A harmonic-model-based front end for robust speech recognition," in *Proc. Eurospeech*, 2003, pp. 1277–1280.

[9] T. Kristjansson and J. Hershey, "High resolution signal reconstruction," in *Proc. ASRU*. IEEE, 2003, pp. 291–296.

[10] J. Laroche, Y. Stylianou, and E. Moulines, "HNM: A simple efficient harmonic + noise model for speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY, October 1993.

[11] B. Resch, M. Nilsson, A. Ekman, and W. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 813–822, March 2007.