On Estimating Rate of Change of Pitch

Edward P. Neuburg

National Security Agency, Fort Meade, Md. 20755

ABSTRACT

To find out how the pitch of speech is changing, how fast and in what direction, one could measure pitch and derive its slope. It is better, however, to measure slope directly. This can be done by comparing narrow-band spectra of adjacent speech segments; they will be very similar, but changing pitch will cause one to look like a stretched or compressed version of the other. We give an algorithm for estimating this stretch and converting it to rate of change of pitch, and present results showing the accuracy of the method.

1. Introduction

In a narrow-band spectrum of voiced speech, for example one obtained from an FFT with a 30-msec window, the salient feature is the peaks at intervals of the pitch frequency. The peaks are particularly clean in the region from about 300 Hz to about 2500 Hz, the first and second formant region; in what follows we deal only with that part of the spectrum.

To fix ideas, we represent each centisecond of speech by the logged spectrum obtained from a 512-point FFT with a 30-msec Hamming window. Of this 257-point spectrum we use only the 100 components numbered 18 to 117, the region of the first and second formants.

Spectra formed from adjacent speech segments are likely to be similar; if pitch is constant they will be very similar indeed. If pitch is going up, the peaks in the later spectrum will be farther apart; the second will be very like a stretched version of the first. If pitch is falling, the second will be like a compressed version of the first. Figure 1 shows spectra of two speech segments a centisecond apart, from a region of speech in which the pitch is changing. We can estimate the rate of change of pitch during this centisecond by estimating how much the first spectrum has to be stretched to make its pitch peaks line up with the peaks in the second.

Figure 1. Two log power spectra 10 msec apart, showing the effect of changing pitch.

2. A simple method

Call $S(J)$ the spectrum at time J . We will try to find out how much we have to stretch $S(J-1)$ to make it look like $S(J)$. The method is to stretch $S(J-1)$ by 21 test factors, ranging from .95 to 1.05 in steps of .005, and compare each stretched $S(J-1)$ with $S(J)$ to see which is the most similar. The stretch is done as follows: given a test factor T , we calculate, for each component of $S(J-1)$, what the amplitude of that component will be when $S(J-1)$ is stretched by T . To find the amplitude of the N th component after stretching, first form $R=N/T$; this is a real number. Let $M=[R]$ be the integer just below R . Using the M th and $(M+1)$ st components of $S(J-1)$ we can interpolate to find what the spectral amplitude should be at the "component" R . This amplitude is assigned to the N th component of the stretched spectrum.

Now that $S(J-1)$ is stretched, we want to know if it has the same shape as $S(J)$; we measure similarity with the correlation coefficient, calculated over the region defined above. Whichever of the 21 test factors causes the stretched $S(J-1)$ to correlate best with $S(J)$ is the estimated factor at time J ; let us call this estimated factor $E(J)$.

Figure 2. Spectrum of speech (solid) and inverse-filtered speech (dashed), showing how whitening makes peaks more uniform.

3. Modifications

Successive spectra are not identical except for pitch. Even if the only thing changing is pitch, the second spectrum will not be an exact stretch of the first; this would happen only if all pitch peaks were the same height. This is not the case; there is formant structure and spectral slope, so some peaks are higher than others. Further, formants and slope may be moving. To lessen these effects we "whiten" the speech to flatten the spectrum, by doing Linear Predictive Coding (LPC) on each segment of speech, and then inverse filtering that segment through the LPC polynomial. Figure 2 shows a typical spectrum of voiced speech, and a spectrum of that same speech after inverse filtering. Pitch peaks are still in the same place, but are more uniform in height.

We have another problem. A little noise can distort one of the spectra, causing a poor correlation score. To make the procedure more robust, we do a little time-averaging. First, we examine more than two successive spectra to find out if they are expanding or contracting; along with the current spectrum $S(J)$ we keep three earlier spectra, and assume that the stretching factor is constant during the indicated time. Thus, given a test factor T , we stretch $S(J-1)$ by T , $S(J-2)$ by T^2 , and $S(J-3)$ by T^3 . The score for test factor T is the sum of the correlations between $S(J)$ and the three stretched spectra.

There is one more averaging step; if the estimated factor is $E(J)$, the final factor $F(J)$ assigned to time J is the average of $E(J)$ and $E(J-1)$. This smoothing is done because the factor wanders (perhaps because of crude quantization), when pitch is not changing. The final factor should properly be assigned to the time at the center of the averaging process.

The factor $F(J)$ is our estimate of the stretch that occurs during the J th time interval; it can be expressed as pitch slope, if desired. For example, if the interval between spectra is $1/100$ of a second, as in the present case, then pitch slope in octaves/second is $100 \cdot \log_2(F(J))$.

4. Making "pitch tracks"

To test the output of a pitch tracker, one can plot (some multiple of) estimated pitch on a narrow-band spectrogram, and see if the pitch track follows a pitch bar. Figure 3 is a spectrogram of speech with the (unsmoothed) output of a pitch tracker (times 10) superimposed. There is no such familiar display on which to plot pitch slope to check its accuracy. One can, however, create a track of pitch (actually some unknown multiple of pitch) from pitch slope, by integrating; errors in pitch slope will cause jumps in the track. This device seems to indicate that pitch slope is estimated very accurately. Figures 4 and 5 are spectrograms on which pitch tracks of this sort are shown. At the left of the spectrogram, time 0, the "pitch" $P(0)$ is arbitrarily set to 1500 Hz. At time (centisecond) 1, $P(1)$ is estimated by multiplying $P(0)$ by $F(1)$, the factor at time 1. A dot is plotted at time 1, frequency $P(1)$. At time 2 a dot is plotted at time 2, frequency $P(2)=F(2)*P(1)$, etc. At each time J a dot is plotted at time J , frequency $P(J)=F(J)*P(J-1)$. The track of dots should follow the nearest pitch bar. Figure 4 shows a typical track for two seconds of speech. (The track is suppressed while speech is unvoiced, and restarted at 1500 Hz at onset of voicing.) Figure 5 shows the track for the same speech with noise (15db s/n) added (shown on the clean spectrogram, because the noisy one is messy).

5. Research issues

Estimation of pitch change could be done in other ways, and some others were tried. For example, the autocorrelation function "stretches" in the same way as the spectrum, when pitch changes. As in other speech processes, the practical way seems to be to transform the speech into a domain in which a familiar statistic, in this case the correlation coefficient, is appropriate. It is the "spikiness" of the spectra that makes it easy to evaluate a stretch in that domain.

Estimating stretch from several successive spectra, and/or smoothing successive estimated factors, gives more stable results at the expense of sensitivity. The parameters that seem to work best are a compromise, arrived at by looking at pitch tracks like those in Figures 4 and 5.

Use of more than two spectra to estimate stretch raises a statistical issue. When dealing with just two curves, correlation is a reasonable measure of similarity. However if we stretch two earlier spectra, we now have to decide if three curves look alike; there seems to be no standard statistic for making this test.

For an interval of 10 msec, stretching factors from .95 to 1.05 appears to capture the range of actual pitch slopes; this observation is based on histograms of stretching factors for a large number and variety of speech segments.

6. Conclusion

Measuring pitch slope by comparing closely spaced narrow-band spectra appears to be an accurate and robust process.

Figure 3. The dotted curve shows the (unsmoothed) output of a typical pitch tracker. Much fine structure is missed.

Figure 4. The dotted curve shows the "integrated" output of the pitch slope algorithm. Fine structure is well captured.

Figure 5. The dotted curve shows the output of the algorithm on noisy speech (15db s/n), plotted on the spectrogram of clean speech.