Extracting Phase from Voiced Speech

Jamie Taylor, Donald Bitzer, Robert Rodman, David McAllister

Department of Computer Science, North Carolina State University, Raleigh, N.C. 27695-8206, U.S.A.

Abstract

A technique is presented to extract a useful phase signal from voiced speech. The speech is divided into glottal pulses and spectral phase is computed for each glottal pulse. The phase signal is then corrected for a variety of defects. Techniques are also presented to improve the reliability of the extracted signal. Finally, the signal is filtered to remove portions which are still unreliable.

1. Introduction

One division of speech sounds is the distinction between voiced and unvoiced (or voiceless). Voiced sounds are made by vibrating the vocal chords, whereas unvoiced sounds are made without vibrating the vocal chords. The vibration of the vocal chords produces glottal pulses, where one glottal pulse (GP) is the sound produced by a single vibration of the vocal chords. The spectrum of voiced speech is the product of the glottal pulse (the driving signal) and the mouth shape (the filter). To prevent aliasing in any processing of the spectrum of a voiced speech sample, the processing must be done over the period of only one glottal pulse at a time. To divide the sample into glottal pulses, the glottal pulse periods (GPPs) are computed by the algorithm described in [4].

Voiced speech is the result of applying a filter (supplied by the mouth and nasal cavities) to the driving signal produced by the vocal chords (glottal pulses). The filter is important in determining mouth shape parameters. The discrete Fourier transform is used to extract the filter parameters from input speech. When a discrete Fourier transform is applied to a segment of speech input of length n, the result is a list of $^n/_2$ complex numbers that define the amplitude and phase of $^n/_2$ harmonics. If the complex number $x + yi$ is treated as a vector in the complex plane, the amplitude is the magnitude of the vector, $\sqrt{x^2 + y^2}$. The phase angle of the vector is measured in radians as $\arctan(^y/_x)$. While the measurement of phase from the discrete Fourier transform is always in the range $-\pi .. \pi$ (due to the nature of the arctan function), the information of interest, the phase of the filter, is not restricted to that range. Because the measured phase is restricted in range, it may exhibit 2π jumps. The phase of the actual filter usually does not have any discontinuities.

Most previous work has used the amplitude characteristics for speech analysis (e.g. [1, 2, 3]). Prevailing wisdom among linguists and speech processing experts is that the phase component of speech does not carry useful information. Previous experiments in which the phase of speech signals was intentionally altered have shown that human listeners do not rely on phase information to determine the content of speech [9]. From this work, one can conclude that human speech perception is dominated by the amplitude portion of the input. This does not imply useful information is not carried by the phase component. On the contrary, synthesized speech which contains phase information most like human speech is more intelligible and natural sounding than synthesized speech which does not [5, 6]. Furthermore, speech synthesized by taking only the phase component of natural speech, while sounding strange, retains a high degree of intelligibility [7]. Finally, phase has proven effective in distinguishing highly similar phonemes in speech such as /m/ and /n/ [8].

2. Extracting phase

Extracting a useful phase signal from speech input is a nontrivial problem. Unlike amplitude, phase is extremely time-sensitive. That is, phase changes linearly with delay and proportionally with harmonic frequency. While amplitude has a range of $0 .. \infty$, measured phase has a range of $-\pi .. \pi$. Furthermore, since measured phase has a constrained range but the actual phase of the filter does not, the measured phase can exhibit 2π jumps even though the actual phase is continuous. Finally, phase is unreliable at low amplitudes because a small change in position in the complex plane can correspond to a large change in phase angle when amplitude is small. Each of these issues will be detailed, and solutions presented.

2.1. Subtracting phase changes caused by phase delay

Voiced speech is the result of applying a filter (supplied by the mouth and nasal cavities) to the driving signal produced by the vocal chords (glottal pulses). The filter is important in determining mouth shape parameters. Therefore, in order to eliminate the effect of the driving signal from the calculations, the speech signal is processed over a single glottal pulse period each time.[1]

In order to reduce the impact of noise in the calculations, an average value is computed across the glottal pulse. For a glottal pulse of length x samples starting at time t, amplitude and phase are computed over x different signal windows. Each window is of length x, but starts at time $t + i$ where i is between 0 and $x - 1$. Amplitude and phase are then averaged over the glottal pulse period. This does not present any problems for amplitude because it typically does not vary significantly across the glottal pulse. Phase, however, does vary across the glottal pulse, as shown in figure 1. To obtain meaningful information from the averaging process, the part of the phase variation caused by moving the window (that is, phase delay) must be determined and removed.

The phase delay function may be derived by beginning with time delay, a measure of the difference in starting time between two different signal windows. Time delay is related

[1] Actually, two contiguous glottal pulses are used so that an average value for the glottal pulse can be obtained by sliding a window of one GP length across the two glottal pulses and averaging the results, as described later in this section.

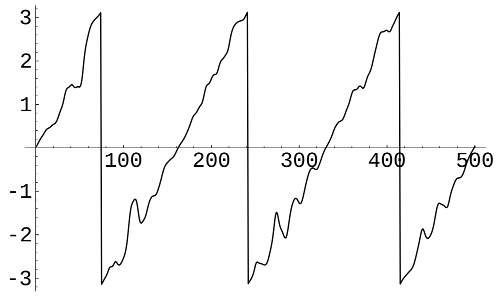


Figure 1. The phase (in radians) of the first harmonic from a piece of speech input vs window start time (in samples) before correcting for change in phase due to phase delay.

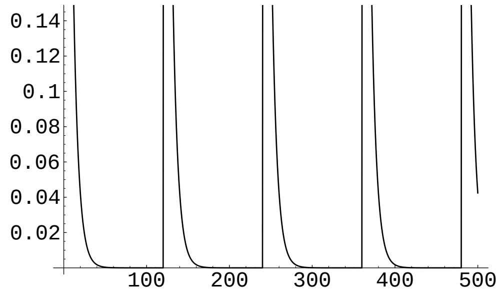


Figure 2. An artificially generated signal with an exponentially decaying pulse repeated at intervals of 120. (Shown as signal amplitude vs sample number.)

to phase by the equation $D(\omega) = -\frac{\partial\Phi}{\partial\omega}$, in which ω is equivalent to harmonic number when time is expressed in terms of the glottal pulse period (i.e. the window size in samples). From this equation it can be derived that $\Phi(\omega) = -\int D(\omega)\partial\omega$. The expected phase function can be computed, examining what happens to phase when time delay increases as the window advances. $D_{i+1}(\omega) = D_i(\omega) + \Delta t \Rightarrow \Phi_{i+1}(\omega) = -\int D_i(\omega)\partial\omega - \int \Delta t\partial\omega = \Phi_i(\omega) - \Delta t\omega$ The effects of the moving window are removed by subtracting the $\Delta t\omega$ term, remembering that Δt must be expressed in terms of the window size as $\frac{-2\pi\delta}{windowsize + 1}$ where δ is the number of samples the window has moved. The Δt term is negative because it represents advancing the signal, rather than delaying it, thus, it is a negative time delay. The phase at the first harmonic completes its revolution through an entire 2π in windowsize increments. This requires $windowsize + 1$ samples, thus accounting for the $\frac{2\pi}{windowsize + 1}$ contribution to Δt.

After removing the effect of the sliding window from the measured phase, the range of the result must be re-constrained. The measured phase has constrained range and experiences 2π jumps, but the “delay corrected” phase calculation does not. The resulting phase, therefore, will have a 2π jump every place the measured phase does, as shown in figure 5. Since there is no particular advantage to using a range of $-\pi..\ \pi$ rather than $0..\ 2\pi$, the later is used because it is more efficient to compute. Both ranges will exhibit some problems discussed in a later section. This gives the final formula for computing the corrected phase: $\mathrm{mod}(measuredPhase - \frac{2\pi\delta\omega}{windowsize + 1}, 2\pi)$. A more rigorous derivation is given in [8].

The formula for delay corrected phase can be verified by constructing an input signal using a known driving signal and filter, such as in figure 2. The phases for several harmonics for one pulse period of this function are shown in figure 3, along with the corresponding expected phases. The corrected phases corresponding to figure 3 are shown in figure 4.

2.2. Correcting for jumps

One of the first steps in the processing of phases, as in the processing of amplitudes, is to average the phases to obtain a single phase per harmonic per glottal pulse. Before this can be done, the phase signal must be corrected to remove any 2π jumps which may appear.

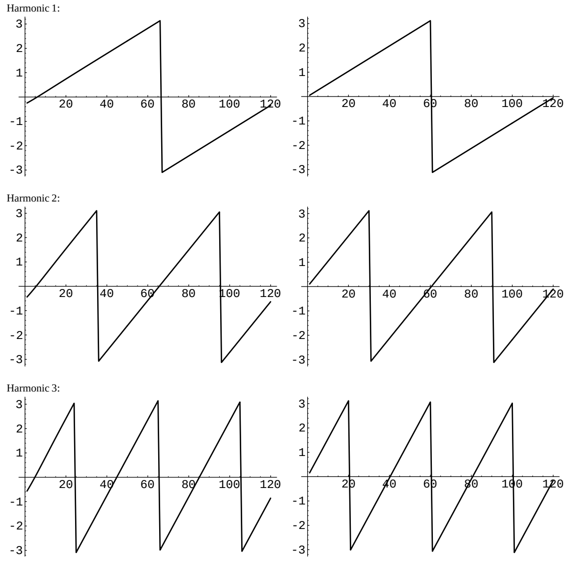


Figure 3. Some of the (uncorrected) phases of the first pulse period of the signal shown in figure 2 appear on the left. The corresponding phases expected due to phase delay are shown on the right. (Shown as phase vs sample number)

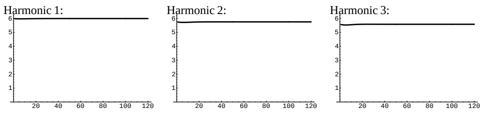


Figure 4. The corrected phases corresponding to the phases in figure 3. The results agree with the phases calculated for the constructed input signal. (Shown as phase vs sample number)

Since the phase represents an angle in polar coordinates, a change of ±2π in phase is not really a change. Values which are over π apart are more likely to have crossed a 2π boundary and should be less than π apart. For example, consecutive values of $(2\pi - 0.1)$ and 0.1 will be calculated as a change of $(0.2 - 2\pi)$, when the actual change in phase is more likely to have been $+0.2$.

To eliminate the 2π jumps, the phase signal is modified so that no sample is further than π away from the previous sample. This restriction means that the phase shift between to samples at the highest harmonic frequency considered is assumed to have

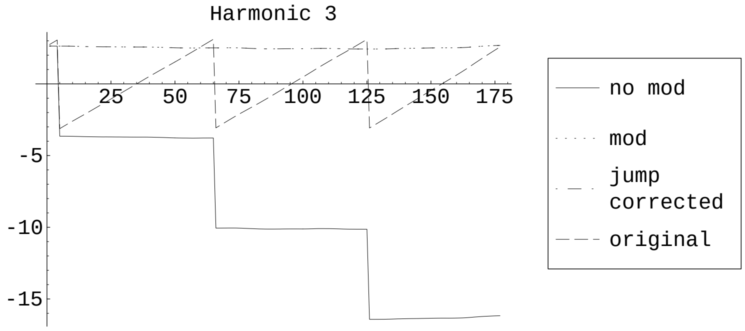


Figure 5. A comparison of the corrected phase vs input window start sample with and without correction for phase delay (as described in section 2.1). The figure shows uncorrected phase ("original"), phase corrected for phase delay without correcting for 2π jumps, both before and after having the range restricted to 0.. 2π ("no mod" and "mod", respectively), and phase corrected for phase delay and corrected for 2π jumps ("jump corrected"). In this example, the "mod" and "jump corrected" lines coincide because the "mod" line is not near a 2π boundary (and thus does not exhibit any jumps).

magnitude less than π. The results are shown in figure 6.

The correction for 2π jumps must be performed any time two angles are compared. The next step in the processing of phases is to compute time delay, the difference in phase between successive harmonics. Since the difference in phase is a difference between two angles, the result must be corrected for 2π jumps.

2.3. Starting the glottal pulse at minimum phase

As mentioned previously, speech input is separated into glottal pulses, and the phase calculations are done over one glottal pulse at a time. The glottal pulse tracker used to identify glottal pulse boundaries is accurate to within a few samples. However, the tracker only identifies the length of the glottal pulses. It would be ideal to also identify the start of the glottal pulse, defined as the window where the phase calculation yields minimal time delay, i.e. the slope of the phase vs harmonic is minimal. This will provide more consistent results when performing averaging of the derived phase characteristics.

Figure 7 shows the results of moving the input signal window by a few samples before computing phase. A line is fit through the phases using the amplitudes as weights. The line is weighted because phase is more likely to be correct when the amplitude is high. This is discussed in further detail in a later section. Figure 7 also shows the slope of this line in relation to the amount the input signal window is offset. The window is closest to the actual start of the glottal pulse when the slope of the line is closest to 0.

Finding the exact beginning of the glottal pulse is computationally intensive. It is not clear that it is necessary to be so precise. Adequate results have been obtained by using a combination of finding an approximation for the beginning of the glottal pulse and simply subtracting the slope of the best-fit line from the phases.

3. Enhancing the phase signal

By far, the most difficult part of extracting useful information from phase is dealing with areas of the input signal which have a low signal to noise ratio. The amplitude component of speech does not present much of a problem in this area, since the amplitude is miniscule in areas of low or no signal. Phase, however, is essentially random in these areas. More accurately, any phase information from a weak signal is easily overwhelmed by noise and thus appears to be random. Figure 8 shows why this is so.

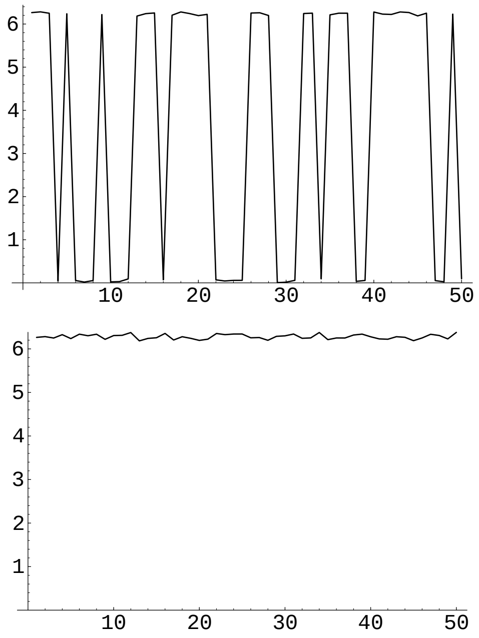
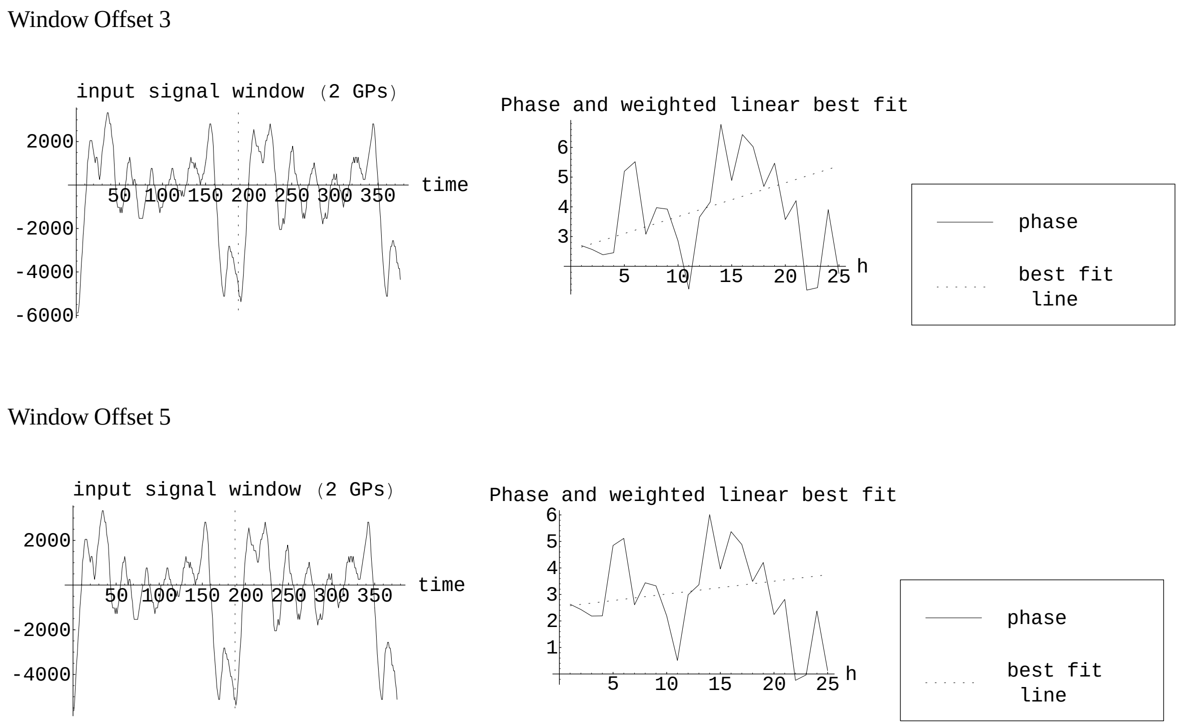


Figure 6. A signal with range 2π±.01 before and after correcting for 2π jumps. The *x* axis is in the time domain.

Several techniques are presented which attempt to reduce the amount of error in the extracted phase signal.

3.1. Strengthening the signal

One way of increasing the signal strength is to combine information from multiple glottal pulses. In order to do this, one may exploit a property of the Fourier transform when applied to periodic signals. For a signal consisting of a periodic function *f*



Window Offset 3

Window Offset 5

Figure 7. The results of moving the input signal window before computing phase. The window offset 5 is closer to the actual start of the glottal pulse because the slope of the line fit through the phase is closer to 0.

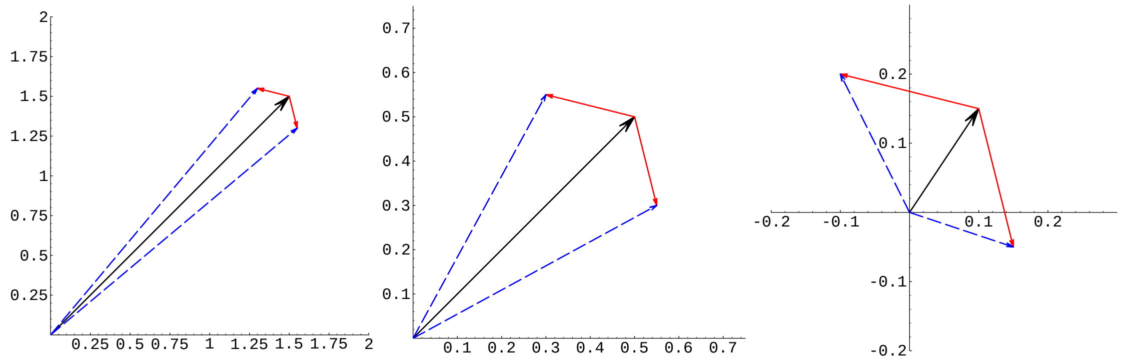


Figure 8. The effects of two different noise vectors (shown in red) with the same amplitude but different phase on various inputs (shown in black). When the amplitude of the input is low, the noise vectors have a disproportionately large effect on the phase of the combined vectors (shown in dashed blue). Note that the figures do not share the same scale. The noise vectors are identical in all three figures.

with period *p*, the sums of certain harmonics of the Fourier transform with window size $n * p$ display certain properties. The value for harmonic $h * n$ ("non-cancelling harmonics") is n times the value for harmonic *h* of the Fourier transform of a single cycle of f using window size *p*. It is also true that the other harmonics ("cancelling harmonics," i.e., ones which are not an integer multiple of *n*) are zero.[1] This property has already been used for $n = 2$ to find the glottal pulse lengths. Similarly, the glottal pulse length may be recomputed over a window of 5 GPs. In practice, this value never deviates from the original GP length estimate by more than 1, thus indicating the high quality of the glottal pulse tracker. One also can calculate phase information using a 5 GP window, processing the results as described in previous sections. This reduces variance in the phase and increases signal strength as illustrated in figure 9

3.2. Filtering weak signals

In some cases, the signal remains weak even after computing over 5 glottal pulses. In these cases it is necessary to filter unusable portions of the phase signal. Several filtering methods are presented below. The results of these methods are compared in figure 10.

[1] See [8] for further details.

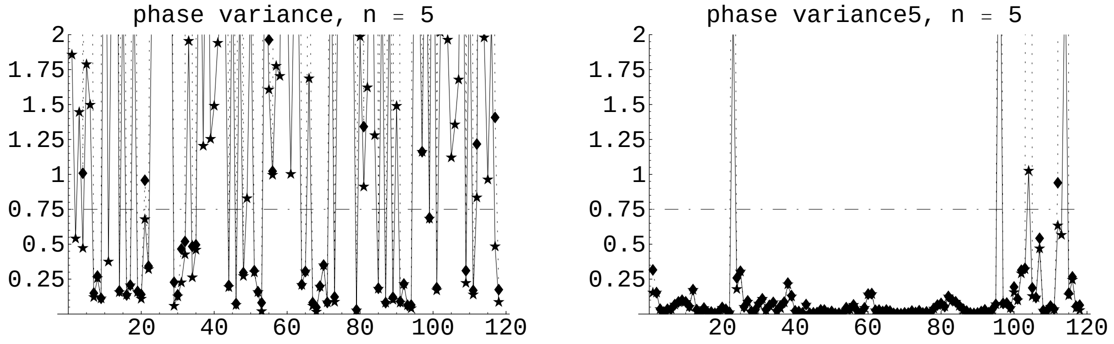


Figure 9. Variance in measured phase is greatly reduced by calculating over a 5 GP window ("phase variance5", right) rather than a 2 GP window ("phase variance", left) (shown as phase vs GP number). The fifth harmonic is shown (as signified by "n = 5") because this is where the nasal filter appears to have a zero, and thus has lower amplitude which is more susceptible to noise.

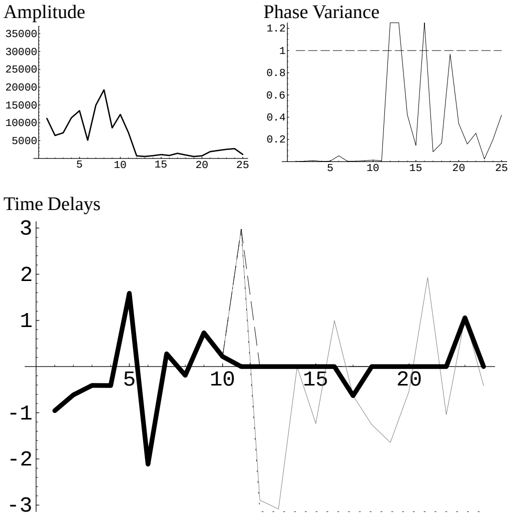


Figure 10. Examples of the filters discussed in section 3.2. The amplitude and phase variances are provided for reference. The results of applying the filters to time delays are overlayed in the chart. The original time delays are in gray. The amplitude filtered time delays are the dotted line. The phase variance filtered time delays are the heavy black line. The x-axis on all of the charts is harmonic number.

3.2.1. *Filter by amplitude*

Since the signal can be assumed to be good when the amplitude is high, the phase signal is filtered based on a simple amplitude threshold. Selecting a single threshold for all inputs is problematic, as samples can vary widely in input speech volume. This problem can be reduced by selecting the threshold as a percentage of the maximum amplitude of the input sample, but this requires *a priori* knowledge of the entire sample, and issues could remain if the amplitude varies widely within the sample.

This method has the advantage of being simple and fast. However, it makes the contribution of the amplitude to the resulting signal very large. The techniques used to extract shape parameters from the signal may pick out the shape of the filter over the shape of the phase information. In particular, the interesting region in the frequency spectrum for distinguishing /m/ and /n/ may have low amplitude. It is desirable to find a way to filter the signal which minimizes the impact of the amplitude on the result.

3.2.2. *Filter by phase variance*

This method removes the reliance on amplitude information completely. One of the first steps in the processing is to average the phases within each glottal pulse to produce a single number per harmonic per glottal pulse. In addition to the mean, the variance, which will be referred to as the phase variance, may also be computed. The phase variance is inversely correlated with amplitude in speech input. The phase signal (after averaging) can then be subjected to a filter based on a threshold of phase variance. Phase variance and the filtered phase signal are actually computed separately, with separate thresholds, for the phases computed using the normal 2 GP window and the extended 5 GP window.

Variance in the phase for a harmonic can come from two sources: noise or change in phase across the glottal pulse. A change in phase across the glottal pulse should result in a constant slope on the corrected phase. Noise is assumed to produce a more random phase signal, with a high variance. (These generalizations have held true in all of the author's experimental observations.) Since the phase variance is to be used as a measure of signal strength, any contribution from the change in phase across the glottal pulse should be minimized or removed, since such a change is a legitimate signal. This can be done by fitting a line to the phase and subtracting it before computing the variance, thus removing any slope from the phase.

The phase variance is easy to compute, and removes the reliance on amplitude information. Choosing the appropriate phase variance threshold appears to be relatively straightforward, and has yielded reasonable results on the input samples that have been considered thus far. Computing the variance of the sine and cosine components separately may yield a more precise descriminator, but it is not clear that this is necessary. Future work on this filter may include applying a moving average filter to the phase variances before applying the threshold.

A variation on the phase variance filtering method is to filter by variance in the change in phase between harmonics ("delta-phase"). If the features to be extracted from the phase signal are to be computed based on delta-phase (because, for example, the shape of the curve is deemed more interesting that the actual values), then it makes sense to filter on delta-phase. There are differences when filtering by phase vs. delta-phase, but they do not appear to be significant. As shown in figure 11, they appear highly correlated.

4. Conclusion

We have shown how to extract the phase signal from voiced segments of human speech. The method yields accurate (low variance) results for speech at a sufficiently high amplitude, but may have a large variance when the amplitude falls below a certain threshold. We noted actions that may be taken in such a contingency.

The phase signal proves useful in several venues of speech processing, including the discrimination of hard-to-distinguish phonemes such as /m/ and /n/ during such endeavors as speech recognition or lip synchronization, as discussed in much detail in [8].

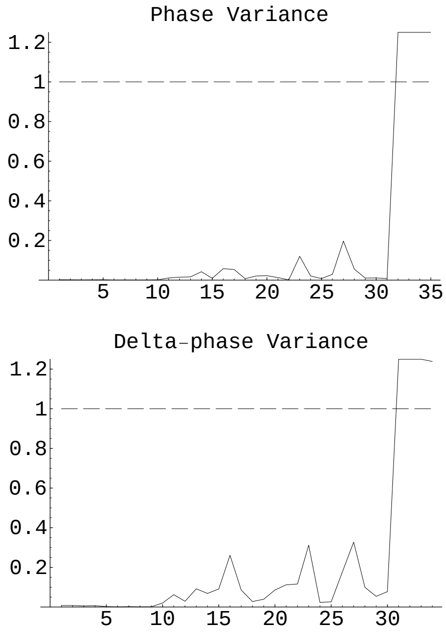


Figure 11. Phase variance and delta-phase variance vs harmonic number.

References

[1] D. McAllister, R. Rodman, and D. Bitzer. Lip Synchronization as an Aid to the Hearing Impared. In *Proceedings of the American Voice Input/Output Society*, pages 233–248, 1997.

[2] D. McAllister, R. Rodman, D. Bitzer, and A. Freeman. Lip Synchronization for Animation. In *SIGGRAPH 97 Visual Proceedings*, pages 255–226, 1997.

[3] D. McAllister, R. Rodman, D. Bitzer, and A. Freeman. Lip Synchronization of Speech. In *Proceedings of the Audio Visual Speech Processing Conference '97*, pages 133–136, 1997.

[4] Rodman, R., McAllister, D., Bitzer, D., and Chappell, D.. A High-Resolution Glottal Pulse Tracker. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP2000)*, October 2000. URL *http://www.multimedia.ncsu.edu/research/voiceio/Papers/GlottalPulseTrackerPaperForBeijing.pdf*.

[5] J.L. Flanagan and R.M. Golden. Phase vocoder. *The Bell Systems Technical Journal* **45**, 1493–1509 (1966).

[6] R.H. Mannell. The effects of phase information on the intelligibility of channel vocoded speech. In *Proceedings of the Third Australian International Conference on Speech Science and Technology*, Melbourne, November 1990. URL *http://www.ling.mq.edu.au/rmannell/research/sst90/*.

[7] A.V. Oppenheim, J.S. Lim, G. Kopec, and S.C. Pohlig. Phase in speech and pictures. In *Proceedings of the ICASSP*, pages 632–637, 1979.

[8] Jamie Taylor. *Using Phase Characteristics of Speech to Distinguish /m/ and /n/*. Ph.D. thesis, Department of Computer Science, North Carolina State University, Raleigh, NC 27695, 2006.

[9] E.G. Wever. *Theory of Hearing*. Dover Publications, New York, 1949. 1970 edition