Accurate Delay Measurement of Coded Speech Signals with Subsample Resolution

Wenliang Lu, D. Sen, and Shuai Wang

School of Electrical Engineering & Telecommunications
University of New South Wales, Australia

Abstract

The accurate estimation of the relative delay between two signals is of vital importance in systems that strive to objectively measure the quality of synthetic speech. The measured delay is used to align the two signals before subsequent analysis and thus inaccurate estimates lead to significant deviations from subjective measures of quality. In this paper, we explore four different methods in terms of their accuracy and complexity in calculating the delay between two signals. One method is the traditional cross-correlation method, along with three other techniques with sub-sample resolution. We test the methods which is investigated on a variety of different signals including sinusoids and speech.

1. Introduction

In intrusive objective speech quality measurement, offline analysis is performed on segments of both original and synthesized speech. The first pre-processing step is the precise alignment of the original and degraded samples(Voran, 1999). Simple cross-correlation in time domain provides a resolution of one sample. However, the synthesized signal produced by speech coding systems is most likely to have subsample delays(Quackenbush, Barnwell III, and Clements, 1988) - making cross-correlation an inadequate technique.

The goal of objective measurement of speech quality is to predict corresponding subjective measurements of quality. Unlike non-intrusive methods which do not require the reference (original) signal from which the synthesized signal was derived, intrusive systems work best when the original and synthesized signals are accurately aligned. This allows the comparison of the spectral content of the signal and makes it feasible to incorporate models of hearing to compute the amount of noise that is perceived by any listener. It is of no surprise, therefore, that better delay detection produces a better correlation prediction of subjective quality.

The current ITU standard for objective measurement of speech quality (P.862), PESQ(Perceptual Evaluation of Speech Quality)(ITU-T, 2002), measures the delay between the original and synthesized speech to a resolution of one sample. We found that when the actual delay between the original and synthesized signal differs by non-integer samples, PESQ can produce considerably different scores. This is especially pronounced in non-waveform coders for which the signals of the original and synthetic signals bear little resemblance.

In one test on PESQ, 10 different codec systems, each containing 6 speech sentences of 45 seconds long respectively, are passed through PESQ. Subsequently all the 60 speech sentences are delayed by 0.5 samples before the same PESQ measurement is conducted again. The artificially introduced delay obviously did not change the perceptual quality of the signals. Results are shown in figure 1.

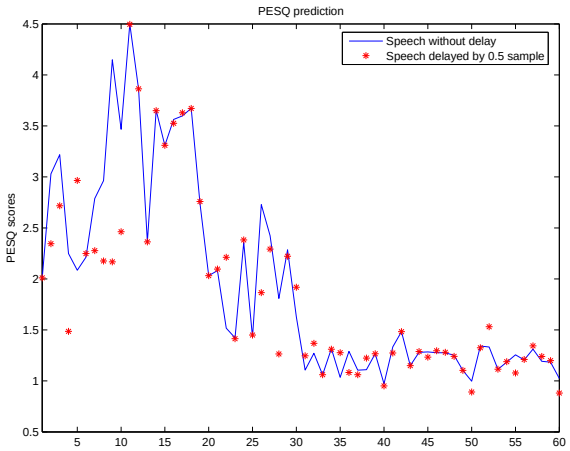

Figure 1: PESQ(ITU-T, 2002) scores for ten different systems, 6 speakers each.

For all 60 speech segments, only 5 have identical scores. The maximum difference between two scores is 1.98 MOS points which is significant since MOS scores range from 1 to 5. It is also of significance that highest discrepancies indicate a lower quality than when the signals were not artificially delayed - obviously indicating that the subsequent comparison of the time-frequency content is disrupted by an incorrect computation of the delay.

Even though recursive algorithms are adopted in PESQ to account for changing delays in the system, the results above indicate that PESQ fails to account for the artificial delay. It is hypothesized, therefore, that better results are possible if higher resolution alignment ($1/4$ or $1/16$ of a sample for example) can be achieved.

In this paper, we investigate three methods other than simple cross-correlation for finding subsample delays between signals.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

2. Methods and stimulations

2.1. Techniques

As stated above, intrusive objective measurement uses the original speech as reference. An intuitive use of the original signal is to calculate the masking threshold according to a psychoacoustic model. The masking threshold can be superimposed on the noise spectrum calculated by the difference between the original and synthesized spectrum to indicate the amount of noise energy above the masking threshold. In this paper, original signals are referred as $s_{orig}[n]$ while $s_{synth}[n]$ for degraded signals. It is assumed that $s_{synth}[n]$ is derived by passing $s_{orig}[n]$ through some system such as a speech coder. We assume $s_{orig}[n]$ and $s_{synth}[n]$ have the same length of L. These systems are usually non-linear and produce variable delay between the input and output. Four different algorithms are investigated in terms of their ability to calculate the delay between original and synthesized waveforms. Algorithms that were investigated are as follows:

1. The first method is the traditional normalized cross-correlation method(Knapp and Carter, 1976). The index of maximum cross-correlation between $s_{orig}[n]$ and $s_{synth}[n]$ is taken to be the amount of delay. This method obviously has resolution limited to one integer. In practical speech coders, the actual delay d_{real} is rarely an integer sample and the best result achieved by this method is the nearest integer to d_{real}.

2. The second method operates in the frequency domain where one signal can be delayed with any precision, integer or decimal, by adding a phase lag to this signal's digital Fourier transform. The delay is achieved, without any sampling rate change (usually requiring an upsampling followed by downsampling) in the time domain. To estimate the delay, we first delay the original signal $S_{orig}(\theta)$ to produce a delayed version $S_{orig\_d}(\theta)$.

$$S_{orig\_d}(\theta, d_{test}) = S_{orig}(\theta) * e^{-jd_{test}\theta\frac{2\pi}{L}} \quad (1)$$

Here L is the length of $s_{orig}[n]$. Subsequently a dot product $P_d(d_{test})$ is computed between $S_{orig\_d}(\theta)$ and conjugate of $S_{synth}(\theta)$.

$$P_d(d_{test}) = \sum_{\theta=0}^{2\pi} S_{orig\_d}(\theta, d_{test})\overline{S_{synth}(\theta)} \quad (2)$$

The dot product can be interpreted as zero lag ($R_{orig\_d,synth}[0]$) of cross-correlation between $s_{orig\_d}[n]$ and $s_{synth}[n]$ in the time domain. This procedure is repeated as within a predefined range of the delay $[d_{down}, d_{up}]$, with a step size of $1/2$, $1/4$ or $1/16$, depending on the desired resolution. The d_{test} for which $P_d(d_{test})$ is a maximum is used as the final estimate of delay. The predefined range $[d_{down}, d_{up}]$ should be chosen carefully to ensure that it covers the true delay. This algorithm does not require the use of $IDFT$, which will result in circular shift effect caused by the periodic extension implicit in the use of $DFT/IDFT$.

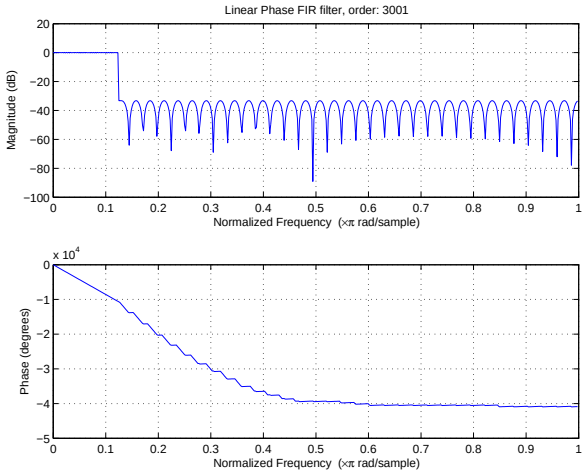

Figure 2: Frequency Response of Low Pass Filter

3. The third method of measuring a subsample resolution delay is implemented in time domain. The method is essentially the cross-correlation method of Method 1 above, but with an upsampling factor M applied to both $s_{orig}[n]$ and $s_{synth}[n]$, to produce an estimate of the delay $d_{detected}$, with a resolution of $1/M$. In practice, the characteristics of the linear phase low-pass filter used for the interpolation also has an affect on the result. For comparison, we have used a 3001th order FIR filter to ensure very narrow transient bands. The frequency response of the filter is shown in Figure 2.

4. The final method is similar to Method 2 but pads zeros after signals, before applying Method 2 on them. The length of signals increase from L to $2L-1$. Without zero-padding in time domain, Method 2's uses less than half time of Method 4 for same signals. However, technically Method 2 should have identical results as Method 4, as padding-zeros in the time domain in Method 4 leads to bandlimited interpolation in frequency domain, which does not provide extra information. Since the lengths of signals are doubled, the complexity of this algorithm is increased significantly.

2.2. Source signals and Delay generation

Four types of signals are used to test the various algorithms described in the previous section. The sampling rate f_s, of all signals was $8000Hz$. The signals are:

1. *Single Sinusoid*: This set of test stimuli were single sinusoids which ranged in frequency from $0.02fs$ to $0.45fs$. The multiple frequencies strove to disclose the performance of the methods as a function of frequency.

2. *Multiple Sinusoids*: This set of stimuli was created by combining sinusoids of different frequencies. There were 30 random frequency components for each combination tested.

3. *Speech signal*: Speech segments containing one sentence were tested.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

Signal	SNR needed
Single Sinusoid	$15dB$
Multi Sinusoids	$3dB$
Speech	$0dB$
Random Noise	$-9dB$

Table 1: Signals VS SNR

4. *Random noise*: Random noise with a Gaussian distribution and zero mean were used as the test stimuli.

To make the results more relevant, random delays were used to create the delayed signal for comparison.

There are several ways of producing a delayed signal $s_{synth}[n]$ from a known signal $s_{orig}[n]$. In this paper, we add to the DFT phase of original signal a linear phase lag, $-2\pi d_{real}\frac{k-1}{L}$, where $d_{real}[n]$ is the delay between $s_{orig}[n]$ and $s_{synth}[n]$, L is the length of original signal, and $k = 0, 1, .., L-1$. To avoid the effect of circular shifts, the original signal is padded with zeros and the delayed signal is recovered beginning at a certain starting index.

3. Results

3.1. Accuracy of methods

Figure 3 to 6 reveal the performance of the four different techniques applied to the four different stimuli as a function of SNR, respectively. Each procedure is repeated 32 times before mean of error are achieved. As Method 1 is the only method not expected to produce subsample resolution, its error is always larger than other methods for all signals, as can be anticipated. White noise of varying power were added to the stimuli to measure the performance as a function of SNR. As can be seen from Figure 3, for single sinusoid, when the SNR is less than $-20dB$, the $log|error|$ levels out at 4. As the SNR is increased, the error decreases before leveling out when the SNR is about $20dB$. The same trend can be observed for the other three sets of stimuli in that the error will reach a minimum when the SNR is slightly above $0dB$.

However, one fact that affects the performance in noise is the amount of frequency components of the stimuli. The four types of signals tested here have different amount of frequency components, and the smallest SNR needed to achieve best performance is directly related to that. Table 1 shows the minimum SNR required to achieve best performance for each stimuli. It can be observed that the greater the frequency content of the stimuli, the less the SNR required to reach best performance. As single sinusoidal signals have only one component in frequency, it tends to be affected more easily by noise. In comparison, random signals have the most number at components of the four method, and thus achieves the best results even when SNR is as low as $-8dB$.

As expected, Method 2 and Method 4 have the same results for all four types of signals.

3.2. Complexity

In this section we investigate the complexity of Methods 2, 3 and 4, as they are all methods that produce subsample resolution delays. From the algorithm, the complexity

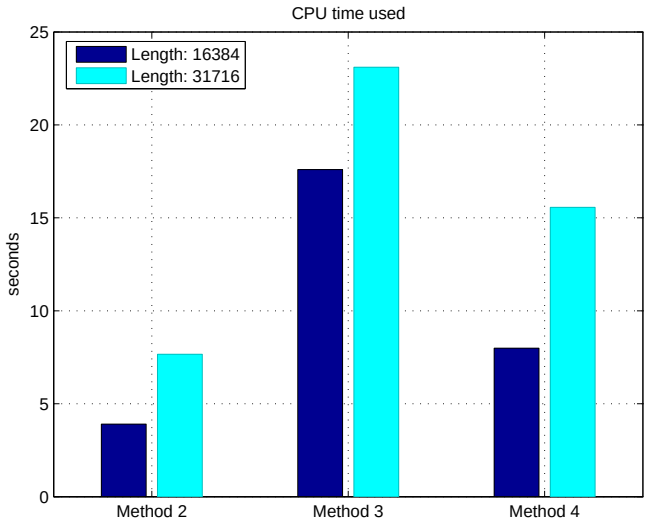

Figure 7: Complexity Comparisonfor Method 2, 3, 4

of method 2 is $(6\log_2 l + 2(e-b)m)l$ real additions and $(4\log_2 l + 4(e-b)m)l$ real multiplications. Here b and e indicate the beginning and end of detecting range, m represent subsample resolution, for example, 4 or 16 (for $1/4$ and $1/16$ resolution) and l is the length of signals involved.

Method 4 has almost identical performance to method 2. However, this method uses zeros padding which results in sequences twice as long as those used for Methods 1 to 3. This leads to a complexity of $(6\log_2 2l + 2(e-b)m)2l$ real additions and $(4\log_2 2l + 2(e-b)m)2l$ real multiplications. The method is thus more than twice as complex as Method 2.

Another important factor, the detecting range $e-b$, plays an important role in complexity of Method 2 and 4. e and b is determined in advance, and complexity is almost linear to $e-b$. Also, when m is high, for example, $m = 64$, even a small $e-b$ will lead to enormous computation. More strategies will be discussed in the next section.

Method 3 requires $(18\log_2 l + 2l_f m + 22)l$ real additions and $(12\log_2 l + 2l_f m + 20)l$ real multiplications, where l_f is the length of low pass filter used for interpolation. Obviously, when l is fixed, l_f is the main factor that control complexity of Method 3.

Figure 7 shows the CPU time used in a test when Methods 2, 3, 4 are applied to two different-length signals, respectively, on a PC with Pentium® 4 Processor (3.00GHz) and 2 GBytes of memory as well. When signal length increases to around double the size, complexity of Method 2 and 4 will almost double. Method 4 takes about twice the CPU time as Method 2 for the same signal, as expected. For Method 3, the complexity increase from 17 seconds to 23 seconds, a ratio of 1.35, which is quite small relative to the increase in signal size.

In a nutshell, Method 3 has an almost fixed computation complexity, while Method 2 and 4 can reduce their complexity with the recursive like strategy stated next section.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

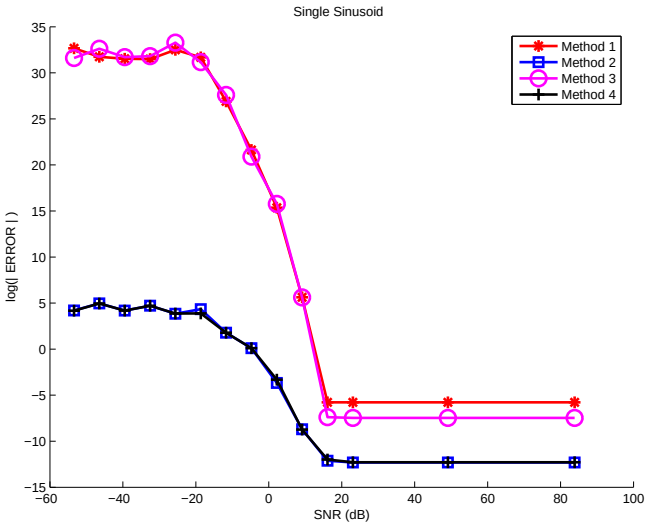

Figure 3: Mean error for single sinusoid

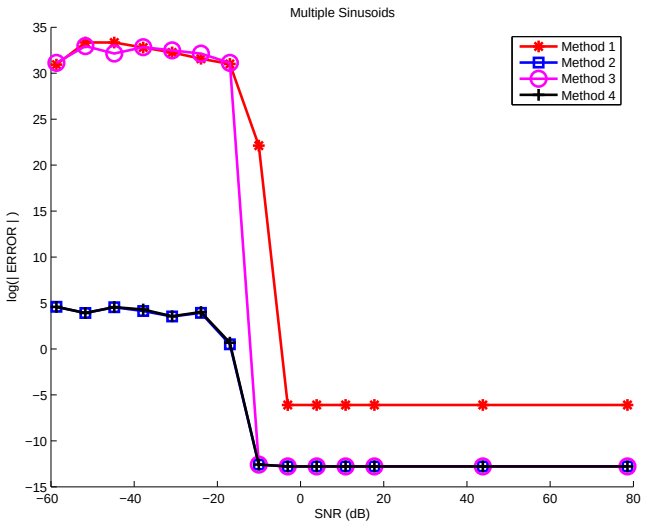

Figure 4: Mean error for multiple sinusoid

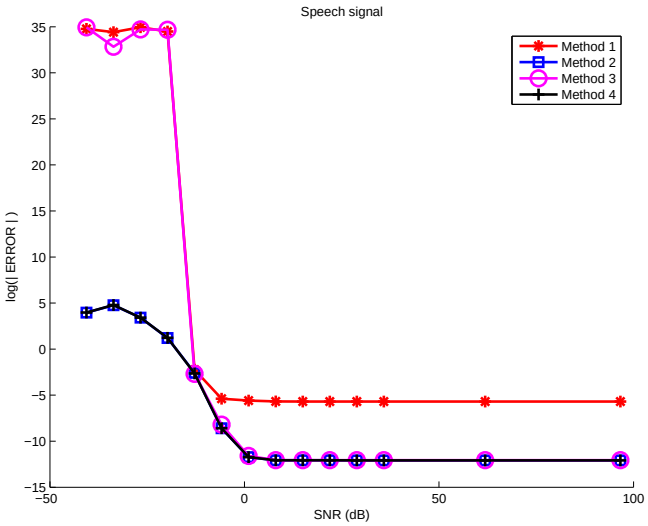

Figure 5: Mean error for speech signal

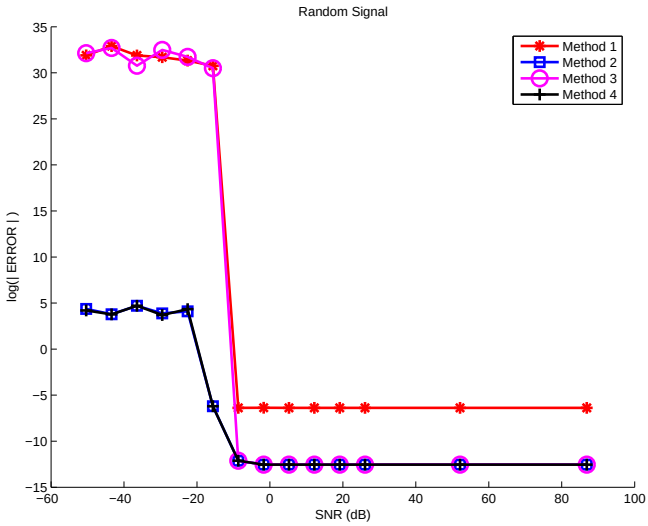

Figure 6: Mean error for random signal

4. Discussion

A slight disadvantage to Method 2 is the requirement of a preset range of testing delays. This range can be determined by an initial use of Method 1 to detect multiple candidate integer values around which to search for the actual delay. The size of the preset range has a linearly proportional impact on the computational complexity. However, if the range can be narrowed to a small range in advance, the complexity can be reduced significantly. A combination of Method 1 with Method 4 in a recursive strategy will lead to excellent accuracy with the least amount of computational complexity (for the same performance).

For low SNR, Method 3 produces the same error as Method 1. However, the error performance of Method 3 improves sharply at higher SNR. This characteristic can be attributed to the low-pass filter used in this algorithm. The filter requires a transition bandwidth as short as possible, at the cost of stop band ripple. To achieve this, a relatively high order, low pass filter is required. This, however, makes the method computationally expensive. However, when the SNR is high enough, the effect of the linear phase low-pass filter will disappear, except for the single sinusoid, which will be distorted by the filter and cannot achieve the expected results.

All the methods discussed in this paper are based on waveform similarity. In practical speech coding systems, the synthesized signal will not only be delayed but significantly altered in amplitude. This means that the error in the delay estimation may be larger for these systems.

5. Conclusion

In this paper, we have investigated three methods of aligning signals with sub-sample resolution. It was found that Method 2 along with Method 1 in a recursive strategy provides the best tradeoff between complexity and performance. Method 3 may be practical for signals with high SNR but at a cost of significantly higher complexity.

References

ITU-T (2002). Perceptual evaluation of speech quality(pesq), an objective method for end-to-end speech

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

quality assessment of narrow-band telephone networks and speech codecs.

Knapp, C. H. and G. C. Carter (1976). The generalized correlation method for estimation of time delay. *IEEE Tranactions on Acoustics, Speech, and Signal Processing*.

Quackenbush, S., T. Barnwell III, and M. Clements (1988). *Objective Measurement of Speech Quality*. Prentice Hall.

Voran, S. (Vol.7, No.4, July 1999). Objective estimation of perceive speech quality–part i: Development of the measuring normalizing block technique. *IEEE Transaction on Speech and Audio Processing*.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review