Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part I – Time alignment.

Antony W. Rix (1), Michael P. Hollier (1), Andries P. Hekstra (2), and John G. Beerends (2)

(1) Psytechnics Limited, 23 Museum Street, Ipswich IP1 1HN, United Kingdom.

Psytechnics was formerly part of BT Laboratories.

(2) Royal PTT Nederland NV, NL-2260 Leidschendam, The Netherlands. A. P. Hekstra is now with Philips Research (WY-61), Prof.Holstlaan 4, NL - 5656 AA Eindhoven.

To be submitted to J.AES.

Abstract

A new model for perceptual evaluation of speech quality (PESQ) was recently standardised by the ITU-T as recommendation P.862. Unlike previous codec assessment models, such as PSQM and MNB (ITU-T P.861), PESQ is able to predict subjective quality with good correlation in a very wide range of conditions, that may include coding distortions, errors, noise, filtering, delay and variable delay.

This paper introduces time delay identification techniques, and outlines some causes of variable delay, before describing the processes that are integrated into PESQ and specified in P.862. More information on the structure of PESQ, and performance results, can be found in the accompanying paper on the PESQ psychoacoustic model.

1 Introduction

A key limitation of early perceptual models for assessing the subjective quality of codecs [1, 2, 3, 4, 5, 6] is that they do not include a method of identifying delay. This makes them unsuitable for end-to-end measurement applications, in which the delay is unknown.

Furthermore, simple methods for delay identification do not perform satisfactorily when the system under test includes coding distortions and filtering. Matters are further complicated by the fact that a growing number of communications systems may introduce variable delay, in particular due to packet-based transmission.

The ITU-T has now standardised a new model that was required to show good performance for a very wide range of applications, both in codec assessment and also end-to-end testing of networks of all types. This model, PESQ, was approved as ITU-T Recommendation P.862 in February 2001 [7, 8].

This paper describes the techniques used in PAMS [9] and PESQ [7, 8] for identification of delay and variable delay. Section 2 outlines the structure of intrusive perceptual quality assessment models and explains why time alignment is necessary. In section 3 the key assumptions made – that delay is piecewise constant in time – are outlined. Section 4 presents a range of techniques for estimation of a constant time delay, illustrating some of the problems with the most commonly-used techniques, and introducing the histogram-based approach used to provide robust delay estimates for PESQ.

Section 5 presents an overview of the types of delay variation that may be encountered in typical audio transmission systems. The algorithms used in PESQ for robust identification of variable delay are described in section 6. Finally, the context of these developments is discussed in section 7.

The accompanying paper on the psychoacoustic model of PESQ [8] provides more detail on where time alignment is implemented in PESQ. [8] also presents performance results comparing PESQ with PSQM and MNB; for this comparison, both PESQ and MNB were extended with a histogram-based constant delay identification algorithm similar to that described in section 4.2.

Although the superior performance of PESQ is due only in part to the new time alignment algorithm, PESQ was found to be much more accurate than PSQM and MNB in almost every condition. ITU-T recommendation P.861, which specified these models [3] was therefore withdrawn as P.862 came into force [7].

2 Perceptual comparison for audible error identification

Many authors have proposed models based on the general structure shown in Figure 1. Quality prediction is based on a comparison of perceptual representations of a reference and a degraded signal. This concept was introduced by Karjalainen [1] and has formed the core of most subsequent models, including the Perceptual Speech Quality Measure (PSQM) [2, 3], the Perceptual Analysis/Measurement System (PAMS) [4, 9], Measuring Normalizing Blocks (MNB) [5] and Perceptual Evaluation of Speech Quality (PESQ) [8, 7]. See [8] for more details on the structure of PESQ.

The auditory transform is a frame-by-frame representation, calculated in most of these models by spectrum estimation using the windowed Fast Fourier Transform (FFT) followed by mapping to perceptual

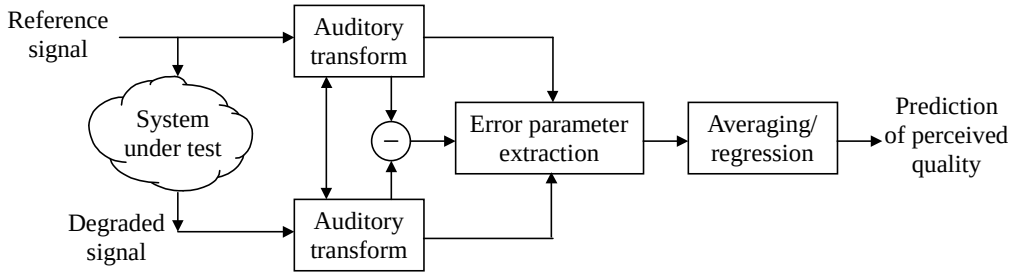


Figure 1: Comparison of perceptual transforms for quality prediction

frequency and loudness scales. A typical frame length is 32ms, used in PSQM and PESQ.

Error parameters are extracted by comparing the two signals frame by frame. However, because of artefacts due to the windowed FFT, and the time-varying nature of signals such as speech, false errors are detected if the frames are mis-aligned by even a small fraction of the frame length. Delay therefore needs to have been cancelled out before processing as models are very sensitive to it.

This delay sensitivity is illustrated by Figure 2, which shows the effect on PSQM score of delay (time mis-alignment) with three different systems. With a delay offset of 10ms, the PSQM score given to system A drops to the PSQM score given to system C at zero delay. A delay of as little as 2ms produces a decrease in measured quality on the order of 0.1 PSQM score.

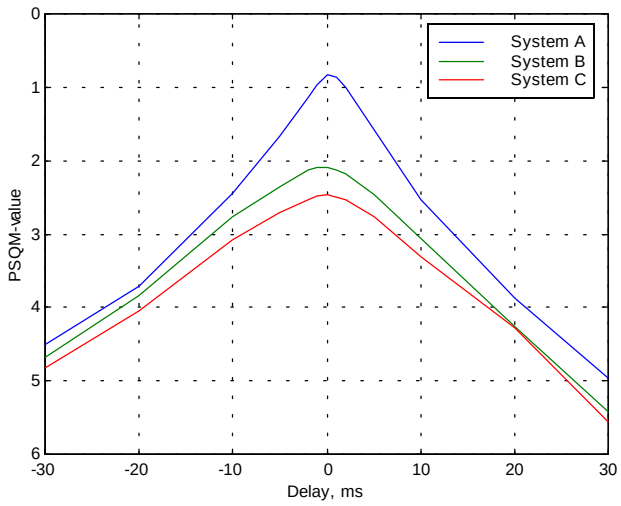


Figure 2: Effect of delay on PSQM score

3 The piecewise constant delay assumption

Before delay estimation techniques can be introduced, this section outlines assumptions made in this paper about the delay of the system under test. Specifically, delay is assumed to be constant in time for a given section of signal i.e. piecewise constant.

In general it cannot be assumed that delay is constant with frequency. Finite impulse response (FIR) filters designed synthetically are usually linear-phase. In this case the delay is constant at all frequencies, unambiguous, and is relatively simple to detect. However some FIR filters, most infinite impulse response (IIR) filters and most practical analogue electronic systems are usually non-linear phase. In this case the most relevant delay measure for perceptual assessment is the group delay, which can be thought of as the delay applied to the modulation envelope. Group delay for these systems varies with frequency and its distribution is important.

To illustrate this, Figure 3 plots the frequency response and group delay for a standard fixed telephone handset. The delay at the frequencies most relevant to speech perception – roughly 1–3kHz – is in the range 0.1–0.2ms. The maximum group delay, 2.0ms, occurs at 180Hz, but the handset's gain is 25dB below its maximum at this frequency. In the time domain, the peak of the impulse response of this handset is at a delay of 0.06ms.

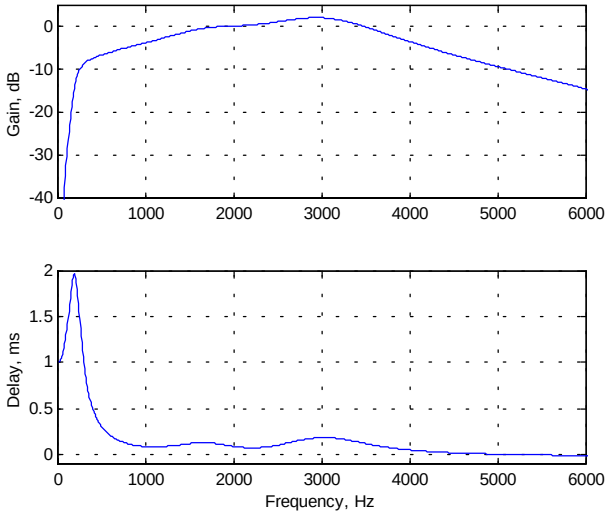


Figure 3: Frequency response and group delay of telephone handset

Given that perceptual models are sensitive to errors in delay on the order of 1ms, it appears in this case that either the location of the maximum of the impulse response, or the mean group delay in the 500–3,000 Hz band, would provide good estimators of delay. However, the maximum delay, or some method that is biased heavily towards this value, may be less useful as a delay estimator.

4 Estimation of constant time delay

This section examines a number of techniques that can be used to estimate the delay of a system. Transfer function estimation (TFE) provides a complete picture of a linear time invariant system. For delay estimation, TFE is often simplified to a cross-correlation of the reference and degraded signals, but this may lead to biased estimates of delay in the case of systems with non-linear phase. Both transfer function estimation and cross-correlation work best when the system is linear and time invariant, but may produce inaccurate results with systems that include low bit-rate coding, errors and noise.

Several authors have examined alternative methods for delay identification that are more robust in these applications. Zglinski embedded an easy-to-locate chirp at the start of each measurement [6]. This allows real-time processing of the remainder of the signal, but it can lead to problems with assessing systems that use silence suppression – voice activity detectors are triggered by the chirp – and/or with low bit-rate coders that are designed to transmit speech only. In addition, Zglinski's method only measures delay at the location of the chirp, making it unable to identify variable delay.

Generally, most other systems have preferred an offline approach. The key motivation for this is that the use of the entire signals improves the reliability of the delay estimation. For example, Voran proposed a two-stage approach, with envelope based alignment followed by a search based on maximising correlation between the power spectral density of frames in each signal [5]. Another two stage algorithm using a histogram-based method, developed for PAMS [9] and PESQ [8], is described in section 4.2.

4.1 Techniques for time delay estimation with linear systems

4.1.1 Transfer function estimation

A number of methods are available for identifying a constant delay. The most common techniques use a simplified version of the transfer function estimation method. If the system under test is stable, linear and time invariant, its frequency-domain transfer function $H(f)$ can be calculated by

$$H(f) = \frac{P_{XY}(f)}{P_{XX}(f)}$$

where $P_{XY}(f)$ is the complex cross-power spectrum between reference signal $x(k)$ and degraded signal $y(k)$, and $P_{XX}(f)$ is the power spectrum of the reference signal. These spectra can be calculated efficiently using the windowed FFT method. The time-domain impulse response $h(k)$ is estimated by taking the inverse FFT of $H(f)$. If the system is linear-phase the delay is given by the index of the maximum of $h(k)$, and $h(k)$ will be even-symmetric in time either side of this point.

4.1.2 Limitations of group delay estimation

Systems that are not linear phase exhibit frequency-dependent group delay. It is possible to estimate the group delay by examining the derivative of the phase of $H(f)$, but this is often not a very stable function and is particularly sensitive to coding errors. As in the example introduced in section 3, in practical communications systems the group delay in the passband is usually within a few samples of the delay found by locating the index of maximum absolute amplitude in $h(k)$. (Sign is sometimes not preserved in telephone networks.)

4.1.3 Windowed cross-correlation method

A common simplification is to discard the spectrum of the reference signal, $P_{XX}(f)$, from the calculation, and assume $H(f)=P_{XY}(f)$. This is equivalent to estimating the impulse response $h(k)$ by cross-correlation of the reference and degraded signals using windowed frames. The result will be identical to transfer function estimation if the reference signal is white noise, i.e. samples are independent and identically distributed (IID), but it is not appropriate to evaluate low bit-rate speech coders using white noise or other synthetic signals that bear little similarity to speech.

This method is simpler to compute, as $P_{XX}(f)$ is no longer required. It can also be more stable, because the aliasing inherent to the windowed FFT method can sometimes lead to large errors in estimating $P_{XX}(f)$, particularly when $x(k)$ is highly coloured – which is often the case with speech.

However, there are certain problems with this method. Depending on the window function used, windowing can cause aliasing or bias in the estimation of $h(k)$. If the system has a non-linear phase response and the reference is not white, the omission of the normalisation by $P_{XX}(f)$ means that the method tends to identify the delay in the frequency range at which $x(k)$ contains the most energy, which may not be perceptually relevant.

4.1.4 Crude delay estimation

Both transfer function estimation and the windowed cross-correlation method are only able to identify delay within the frame length used for the windowed FFT (typically within $\pm N/2$ samples for an N-point FFT). Although in communications applications the duration of the impulse response is normally within a few milliseconds, the bulk delay may be much larger than this. It is therefore necessary to estimate the delay to within some fraction of N samples; the appropriate offset is then used to eliminate the estimated delay from the calculation of $h(k)$. Two alternative methods are commonly used to achieve this:

- signal matching – locating a distinctive signal component such as a chirp
- correlation of signal envelopes, usually after decimation.

The accuracy of this crude delay estimation becomes less critical as N is increased. However, this increases computational complexity and reduces the noise rejection

and stability of the transfer function estimate. A good balance between these factors is usually achieved with a frame length on the order of 50ms.

4.1.5 Whole-signal cross-correlation

As an alternative to crude alignment followed by windowed cross-correlation or transfer function estimation, it is possible to compute the cross-correlation of the whole files. If the system is linear, this usually gives very similar results to envelope alignment followed by windowed cross-correlation. However, it is quite computationally inefficient to cross-correlate the whole signals if they are long (more than a second or so), and this method is particularly sensitive to noise, non-linearity and time variance in the system.

4.1.6 Effect of non-linear or non-stationary components of communications systems

In speech communications systems, however, it is not satisfactory to assume that the system is linear or time-invariant. Whilst the filtering introduced by analogue components may be roughly constant, the instantaneous transfer function varies from frame to frame due to non-linearity in the coding or errors introduced in the coded bitstream (such as radio errors in a mobile path).

Furthermore, during silent intervals, some systems cease transmission. The receiver re-synthesises a comfort noise signal, of similar level and spectrum, to prevent the connection from sounding 'dead'. The sent and received signals during such periods are independent and therefore impossible to align.

To illustrate the effect of coding errors, Figure 4 presents the impulse response estimates using the methods described above for a simulated mobile phone connection. This includes an approximation to the acoustic and analogue path from mouth to microphone, realistic radio errors, and background noise. The group delay introduced by the acoustic/analogue path ranges between 0.5–0.9ms in the frequency range 500–3,000Hz, and the delay of the digital path has already been eliminated. Frames of 512 samples (32ms at 16kHz sample rate) were used, with a Hann window function. The (known) impulse response of the linear part of the system is shown as a dotted line in Figure 4.

In this example, both cross-correlation based methods give estimates that are in error by about 1ms. The windowed cross-correlation method produces a smooth curve with a maximum at 1.7ms, shown as the dashed line in Figure 4. However, there is little to distinguish this peak from the peak of opposite sign and 96% of the amplitude, at 3.1ms delay; slightly different radio errors could easily lead to this value being detected instead. The shape of the whole-signal cross correlation function (not plotted) is very similar at this scale, but it could also be used for delay identification outside the ±16ms range of the windowed method without the need for a crude delay estimate.

In contrast, the maximum amplitude of the impulse response calculated using the TFE method (the solid line) is at a delay of 0.7ms. This falls within the range of the group delay of the system. However, due to non-linear distortions introduced by the codec and radio path, the impulse response estimated using this method is not a good approximation to the actual impulse response of the linear part of the phone. In some cases this method has been found to give delay estimates that are less reliable than those obtained using the cross-correlation method.

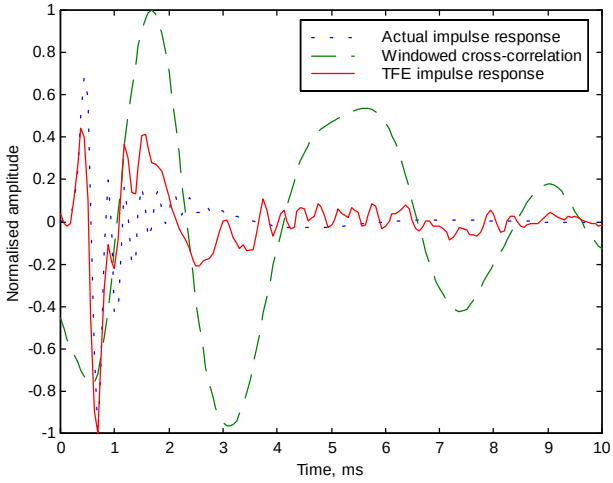


Figure 4: Example impulse response estimates

4.2 New histogram-based algorithm for delay identification

The example in Figure 4 illustrates some of the problems with both transfer function estimation and cross-correlation as methods of delay identification. A new time alignment algorithm was therefore developed and tested against a large corpus of speech material containing measurements or simulations of a very wide range of networks. The method uses an envelope-based crude delay estimate followed by fine-scale delay identification using a weighted histogram of the frame-by-frame delay.

4.2.1 Pre-processing

As mentioned in section 4.1.3, correlation-based methods may give delay estimates that are biased towards the parts of the frequency range in which one or both of the signals has greatest energy. Depending on talker age and gender, between 60% and 90% of the energy of natural speech lies below 500Hz. For perceptual quality assessment, however, the frequency range 1–3kHz is more important as it is thought to contain most of the parts of speech that are important for intelligibility. Both input signals are therefore processed through a filter that attenuates strongly below 500Hz. In PESQ, this filtering is in addition to the application of a filter that models the listening handset, which further serves to bandlimit the signals.

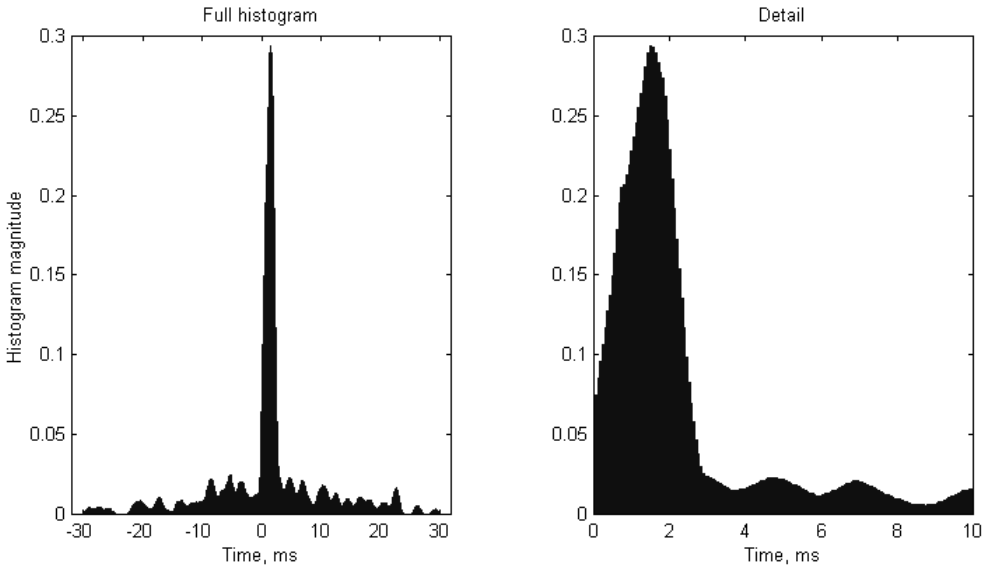


Figure 5: Histogram-based delay estimate

4.2.2 Crude envelope-based delay estimation

Decimated envelopes of the reference and degraded signals are calculated by measuring the power of the filtered signals over 4ms non-overlapping frames. The location of maximum cross-correlation between the envelopes is taken as a crude estimate of the delay. It is important to note that the accuracy of this estimate depends on there being sufficient information in the envelope. Fortunately this is the case with speech, which is highly non-stationary at this timescale. Provided that the signals contain at least 500ms of speech, this crude delay estimate is usually accurate to within ±8ms.

4.2.3 Construction of the weighted delay histogram

The transfer function estimation and cross-correlation based approaches described in section 4.1 assume that the system under test is stationary, and calculate functions that are averaged over the whole duration of the measurement. If this assumption is not valid, the resulting delay estimates may be highly inaccurate.

An alternative method was developed, which constructs a histogram of the frame-by-frame delay estimates using the following steps:

- The crude delay estimate calculated by envelope cross-correlation is eliminated.
- The signals are divided into 64ms, 75% overlapping frames using the Hann window.
- The index of maximum absolute cross-correlation is found for each frame.
- The value of maximum cross-correlation, raised to the power 0.125, is taken as a weight for each frame, to give slight emphasis to the louder sections of the signals.
- A weighted histogram of the delay estimates is constructed by adding this weight to the histogram bin given by the index of maximum cross-correlation.
- The histogram is normalised by the sum of the weights at all indices.
- The histogram is smoothed by convolution with a triangular kernel of half-width 1ms and peak value 1.0.
- The delay estimate is given by the location of the peak of the smoothed histogram.

The histogram produced using this method is plotted in Figure 5 for the simulated mobile phone connection described in section 4.1.6. In this case the delay estimate is 1.5ms, which is slightly higher than the 0.5–0.9ms group delay in the 500–3000Hz frequency range. Compared to the basic cross-correlation methods, however, this estimate is slightly more accurate and, as there is only one large peak, it is much less ambiguous.

The histogram-based method also allows a confidence measure between 0 and 1 to be calculated from the value at the peak – in this example, 0.29. This is a more robust approach than using correlation coefficient, which varies between –1 and 1, and is particularly sensitive to non-stationarity in the system. The confidence value in this case is quite low, a result of the poor quality of the connection, the highly non-linear phase of the analogue filter, and the added noise.

5 Variable delay in communications systems

The methods outlined in the previous section can be used to estimate a single value of delay. However, there are circumstances in which the delay of the system may vary, either continuously or in discrete steps. This section considers an number of sources of variable delay.

5.1 Packet-based transmission

It is becoming increasing common to use packet-switched networks, typically based on Internet protocol (IP) and/or asynchronous transfer mode (ATM), for carrying real-time speech traffic. This has the potential for significant cost savings over traditional circuit-switched networks, due to the lower cost of switching equipment and the ability to run a single network for both voice and data. For the purposes of this paper, we will focus on voice over IP (VoIP).

In packet-based transmission, speech is compressed using a coding scheme such as ITU-T Recommendation G.711 or G.723.1, and divided into packets. In VoIP a typical packet length is 20ms. The packets are sent across the network, reassembled and decoded to a speech stream at the receiver. Packets may reach the receiver in a different order, because the route used to transmit the packets may change to one of different delay. Additionally, because many network components operate on a "best effort" basis and switch traffic from many streams, the time taken for each packet to travel may vary and some packets may be lost.

The variation in the delay of each packet is termed packet jitter. It is not usually a problem for data traffic, which is fairly insensitive to delay, allowing time to request that lost packets be re-sent. This is not, however, the case for two-way voice communication, which can be hampered by round-trip delay of as little as 50ms if there is echo present. Although the impairment due to delay can be reduced by use of appropriate echo cancellation, this cannot eliminate the effect of delay on conversational quality.

There are therefore two conflicting requirements. To prevent packet jitter from leading to packets being discarded because they arrive too late, the receiver must buffer the incoming data. The longer the buffer, the fewer packets are lost and the higher the (one-way) speech quality of the system. However, a longer buffer means greater round-trip delay, and correspondingly lower conversational quality. From a conversational perspective, the buffer needs to be as short as possible.

Packet jitter in the network can lead to variations in the delay in the audio path through several different mechanisms. Two of the most common are dynamic buffer resizing and excessive late packet dropping.

Dynamic buffer resizing during silence is a common method used in VoIP systems to deal with time-varying levels of packet jitter whilst attempting to minimise bulk delay. The buffer length is changed during silent intervals and leads to delay variations in silence.

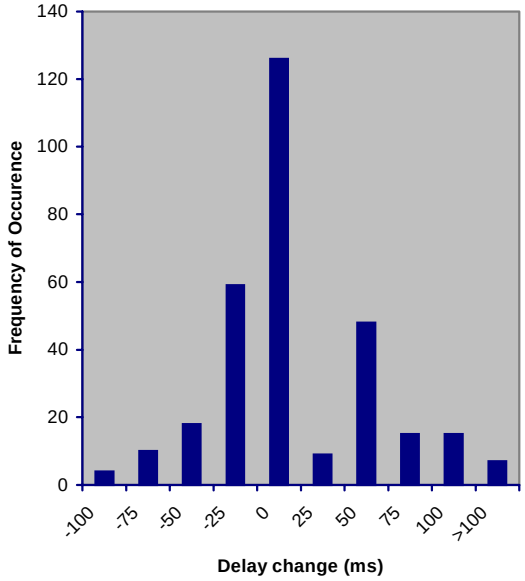
Excessive late packet dropping is a less frequent effect. In this case a large change occurs in the packet delay during a speech event – for example, a route alteration that delays subsequent packets by 100ms more than the previous route. Following this change the first few packets arrive too late and the buffer becomes empty. The typical result is the insertion of silence in the middle of the speech event, with playback re-starting only after a few new packets have arrived in the buffer.

In both of these cases the delay changes occur in discrete steps, and delay is constant in between these changes. It is therefore assumed for PESQ that the delay of the system is piecewise constant.

5.2 Typical characteristics of VoIP variable delay

To give an idea of the delay variations encountered, measurements of delay changes on packet-based networks

(a) VoIP field trial



(b) PSTN/IP Gateway test

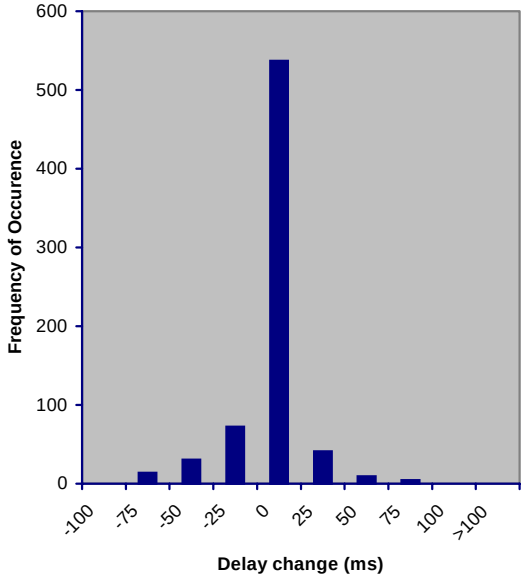


Figure 6: Measured delay variations in VoIP

are presented in Figure 6. This shows the distribution of changes in the one-way end-to-end audio delay during a series of 16-second measurements with artificial speech across two types of network. Figure 6(a) gives results from a field trial of a PC-based VoIP system. Figure 6(b) shows the corresponding distribution measured across a PSTN to Internet Gateway (PIG) system. Delay changes as large as ±100ms were encountered, and in the VoIP field trial a delay change of 25ms or more in magnitude occurred in over 50% of measurements.

5.3 Parametric coding

Because there is significant natural variation in the temporal structure of speech, it is possible to continuously warp the local timebase of speech with relatively little perceived effect. This is now being exploited by very low bit-rate coders such as waveform interpolation (WI), mixed-excitation linear prediction (MELP) and harmonic vector excitation coding (HVXC), typically at bit-rates of 4kbit/s and below. In some coders the amount of local time-warping is limited to e.g. ±1ms per 20ms frame, but in other cases the time-warping can accumulate to many tens of milliseconds over a speech utterance.

Development of these coders is still an active research area. As yet, such low bit-rate coders have gained only limited acceptance, and they are certainly not in widespread use in telephone network equipment.

For VoIP there is little incentive to encode speech at such low bit-rates. In the absence of header compression, the packetisation overhead of real-time transmission over IP typically adds 16kbit/s for 20ms packets. In this case even an 8kbit/s coder is responsible for only a third of the capacity used by a voice call.

It is therefore difficult to comment on how important these coders will be in the context of intrusive quality assessment. However, given that they may violate the assumption of piecewise constant delay, it will be necessary to re-evaluate the accuracy of models such as PESQ if and when these coders become common in communications applications.

5.4 Other sources of variable delay

Whenever a signal is carried outside the digital domain, in analogue connections or storage media such as tape, there is the potential for the sampling rate at the receiver to vary from that used in transmission/recording, for example if the clocks are not synchronised or the media can be played back at a different rate. This can also lead to continuous variation in delay.

Sample rate jitter is generally well controlled in communications networks. As many components have little or no ability to buffer signals, this is essential to prevent severe distortions from being introduced. Typically digital clocks will be stable to well within 0.1% deviation. In comparison to non-linear coding distortions, this causes little problem for delay identification as long as samples are no more than a few seconds in duration.

However, playback rate in analogue tape transport is may not always be so stable. Tape speed can in extreme cases vary by more than 1% over relatively short timescales. In high-quality audio applications where tape speed is well controlled, it is sometimes possible to perform frame-by-frame time alignment to compensate for this [10]. In this case the delay changes are sufficiently small for a variable delay alignment model such as that outlined in section 6.

Difficulty in tracking the carrier frequency in certain radio systems, especially those using analogue single side-band (SSB) modulation, can also lead to phase shifts that cause not only variations in time base but also in the frequency of the signals. To minimise these distortions, SSB systems now generally use high-quality digital clocks and, as for clock asynchrony, they are unlikely to pose a problem.

5.5 Perceptual effect of delay variation

Based on the effect on perceived speech quality, the different classes of delay variation can be summarised as follows.

Delay changes in silence are normally inaudible. In certain cases the system may introduce some type of discontinuity, for example a short period of digital silence corresponding to an increase in delay, but even this will usually be minor. Large delay changes may have some impact on conversational quality, but this is of similar magnitude to the effect of large constant delay.

Step delay changes during speech are normally audible, although it may be possible for a sophisticated concealment algorithm to hide a delay change. Even a small change in delay can cause an annoying discontinuity. To assess this in PESQ, the frames on either side of a deletion or insertion are processed to identify any audible discontinuities, but entire frames which were deleted are ignored [8].

The effect of c**ontinuous delay changes during speech** depends strongly on how the temporal structure of speech is affected. If the pitch is made constant, the resultant "robotisation" of the speech is sometimes found by subjects to be annoying. However, as long as some natural variation in pitch remains, it can be difficult to detect any impairment. Continuous warping of the time axis by less than about 1% is normally inaudible with speech signals.

6 Robust identification of variable delay

The time alignment method outlined in section 4.2 was adapted to allow the identification of delay changes, both in silent periods and during speech [9]. A further extension, re-alignment based on identification of very bad frames, was developed for use with PESQ [8].

6.1 Utterance identification

For the purposes of time alignment, an utterance was defined as a continuous section of speech of at least

300ms duration, containing no silent period longer than 200ms. This identification is performed on the reference signal by a voice activity detector (VAD) with an adaptive threshold to make the speech/non-speech decision robust to noise.

6.2 Utterance delay estimation

Firstly, a crude delay estimate is calculated across the entire signals using the envelope correlation method described in section 4.2.2. This allows large time offsets, due to constant delay or poor synchronisation in the measurement process, to be eliminated.

The crude delay of each utterance is then identified by applying the envelope correlation method to sections in the reference and degraded files corresponding to that utterance. After eliminating this delay, fine delay and confidence estimation is performed by applying the weighted histogram method detailed in section 4.2.3.

Finally, the utterance boundaries are placed in the middle of each silent period or, for the first and last utterances, at the start and end of the file. Because each utterance is aligned separately, this method directly accounts for delay changes during silence.

6.3 Utterance splitting

However, delay estimation of whole utterances is not able to take account of delay changes during speech. If these occur, the delay estimate is usually the delay of the longest section, but large false errors can be observed for the sections that are of different delay. The utterance alignment method was therefore extended as follows.

Each utterance is divided in two and each section is processed through the same crude/fine delay estimation stages as before. This is repeated at a large number of division points. If there is evidence for a delay change – an increase in confidence and an absolute change in delay of 4ms or greater – the division that produces greatest confidence is selected. In this case the utterance is split in two at this point.

Because it is possible that multiple delay changes occur in each utterance, the process is recursively applied to each new half before proceeding to the remaining utterances.

6.4 Use of delay information in PESQ

The PESQ psychoacoustic model operates on 32ms, 50% overlapping frames. The utterance delay estimates are converted to frame-by-frame delay according to the delay of the utterance in which each frame begins (defined by its location in the reference signal).

If a decrease in delay has occurred, there will be a section in the reference signal which has been deleted. Any frames that have been entirely deleted are discarded from processing.

6.5 Bad frame identification and re-alignment

In some cases the process of utterance splitting fails to identify a delay change within speech. This is clearly visible in the psychoacoustic domain, as the mis-aligned sections show high levels of disturbance [8]. To improve the accuracy of PESQ in these cases it was found necessary to re-align these sections.

A bad frame is defined as a frame with a symmetric disturbance $D_n > 30$. Bad frames separated by less than four non-bad frames are grouped together into bad intervals. After eliminating the delay estimated previously, the portions of the reference and degraded signals which correspond to the bad interval are re-aligned using simple cross-correlation.

For each frame in the bad intervals the psychoacoustic processing is re-calculated using the new delay estimate. If the result is a decrease in disturbance, the new disturbance values are used. In general this results in disturbance estimates that are closer to the values that would have been obtained if the previous time alignment had been correct.

7 Discussion

With low bit-rate coding, errors, and the potential for delay to vary during a test, time alignment is a difficult problem. Common ‘classical’ techniques, such as transfer function estimation or cross-correlation, are not robust and can give biased, or highly inaccurate, estimates in many cases. The new techniques for delay identification that have been introduced in this paper work well across a large database of simulated and measured network conditions including many different systems and types of delay change.

The piecewise constant delay assumption appears to be valid for many applications, including common variable-delay communications systems such as VoIP.

During the development of PESQ it was found that introducing these time alignment techniques described in this paper made a significant improvement to the model’s accuracy, especially with variable delay conditions. This sets PESQ apart from earlier models, such as PSQM and MNB, that did not include time alignment processes and were therefore unsuitable for use in end-to-end measurement applications.

In contrast, PESQ was evaluated against a demanding set of conditions including variable delay, filtering, coding and errors – sometimes all at the same time. The results presented in the accompanying paper [8] illustrate that PESQ gives accurate quality scores in many different applications.

8 Acknowledgements

Thanks are due to ITU-T study group 12 for organising and driving the recent competition, and in particular the other proponents (Ascom, Deutsche Telekom and

Ericsson) who contributed valuable test data and provided stiff competition. The authors would like to acknowledge the assistance of many of their colleagues at BT and KPN, and thank the companies who acted as independent validation laboratories: AT&T, Lucent Technologies, Nortel Networks, and especially France Telecom R&D. A. W. Rix is also supported by the Royal Commission for the Exhibition of 1851.

9 References

[1] Karjalainen, M. *A new auditory model for the evaluation of sound quality of audio systems*, IEEE ICASSP, 608–611, 1985.

[2] J. G. Beerends and J. A. Stemerdink. *A perceptual speech-quality measure based on a psychoacoustic sound representation*, J. AES, 42 (3), p. 115, 1994.

[3] *Objective quality measurement of telephone-band (300–3400 Hz) speech codecs*, ITU-T Recommendation P.861, February 1998.

[4] M. P. Hollier, M. O. Hawksford and D. R. Guard. *Algorithms for assessing the subjectivity of perceptually weighted audible errors*, J. AES, 43 (12), p. 1041, 1995.

[5] Voran, S. *Objective estimation of perceived speech quality – Part I: Development of the measuring normalizing block technique*, IEEE Transactions on Speech and Audio Processing, 7 (4), 371–382, 1999.

[6] Zglinski, Z. and Di Pietro, G. "Method of transmission quality estimation of a speech transmission link", European Patent no. EP0644674, 1995.

[7] *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, ITU-T Recommendation P.862, February 2001.

[8] Hekstra, A. P., Beerends, J. G., Rix, A. W. and Hollier, M. P. *Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model*, submitted to J.AES, June 2001.

[9] Rix, A. W. and Hollier, M. P. *The perceptual analysis measurement system for robust end-to-end speech quality assessment*, IEEE ICASSP, June 2000.

[10] Herre, J. Personal communication, February 2001.