Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I–Temporal Alignment

JOHN G. BEERENDS,[1] *AES Fellow*, CHRISTIAN SCHMIDMER[2], JENS BERGER[3], MATTHIAS OBERMANN[2], RAPHAEL ULLMANN[3], JOACHIM POMY[2], AND MICHAEL KEYHL,[2] *AES Member*

[1] *TNO, P. O. Box 5050, NL-2600 GB Delft, The Netherlands*
[2] *OPTICOM GmbH, Nägelsbachstrasse 38, D - 91052 Erlangen, Germany*
[3] *SwissQual AG, Allmendweg 8, CH-4528 Zuchwil, Switzerland*

In two closely related papers we present POLQA (Perceptual Objective Listening Quality Assessment), the third generation perceptual objective speech quality measurement algorithm, standardized by the International Telecommunication Union (ITU-T) as Recommendation P.863 in 2011. The algorithm is composed of two separate parts, a temporal alignment that finds speech parts that belong together and a perceptual model that builds an internal representation of the aligned input and output of the device under test. This paper (Part I) provides the basics of the POLQA approach and outlines the core elements of the underlying temporal alignment. The newly developed alignment approach allows assessing the latest Voice over IP technology that often introduces sudden align jumps (mostly in silent intervals) as well as slowly changing time scalings (mostly during speech activity), either using a pitch preserving technique like PSOLA (Pitch Synchronous Overlap Add) or a straight forward technique equivalent to sample rate changes.

0 INTRODUCTION

During the past decades objective speech quality measurement methods have been developed and deployed using a perceptual measurement approach. In this approach a perception-based algorithm simulates the behavior of a subject that rates the quality of an audio fragment in a listening test. For speech quality one mostly uses the so-called absolute category rating listening test, where subjects judge the quality of a degraded speech fragment without having access to the clean reference speech fragment. Listening tests carried out within the International Telecommunication Union (ITU) mostly use an absolute category rating (ACR) five-point opinion scale [1], [2] that is consequently also used in the objective speech quality measurement methods that were standardized by the ITU, PSQM (Perceptual Speech Quality Measure, ITU-T Rec. P.861, 1996) [3], [4], and its follow-up PESQ (Perceptual Evaluation of Speech Quality, ITU-T Rec. P.862, 2000) [5] – [9]. The focus of these measurement standards is on narrowband speech quality (audio bandwidth 100–3500 Hz) [3] – [8], although a wideband extension (50–7000 Hz) was devised in 2005 [9]. PESQ provides for very good correlations with subjective listening tests on narrowband speech data and acceptable correlations for wideband data.

As new (wideband) voice services are being rolled out that often introduce new time warping distortions and other new (wideband) distortions, prediction of the perceived quality by PESQ is becoming unreliable [10]. Therefore ITU-T (ITU-Telecom sector) Study Group 12 initiated the standardization of a new speech quality assessment algorithm as a technology update of PESQ. In order to be able to develop a future proof measurement standard, the audio bandwidth was extended beyond the current 7 kHz wideband standard toward 14 kHz (super-wideband speech). High fidelity reference speech file recordings were made in a low noise, low reverberation environment using 48 kHz sample rate with a voice bandwidth of 14 kHz.

Six proponents submitted candidate algorithms to the ITU-T for benchmarking, from which three were found to meet the requirements and were thus selected for the final standardization. The benchmark was based on a wide comparison with data from subjective tests, including an extensive number of newly designed databases, that were unknown to the candidate algorithms. The selection was mainly carried out on the basis of the prediction error in terms of mean opinion scores (MOS), taking into account the 95% confidence interval of each MOS score. To provide for a unique measurement standard, the three selected candidate algorithms from OPTICOM, SwissQual, and TNO were further integrated into a joint model. The joint model was found to outperform each of the underlying models. According to the original working title this model is referred to as POLQA (Perceptual Objective Listening Quality Assessment), and it was accepted in January 2011 by ITU-T as the new speech quality measurement standard Rec. P.863 for narrowband, wideband, and super-wideband speech quality assessment [11].

It should be noted that although POLQA operates at a sampling frequency of 48 kHz when run in super-wideband mode, this should not be misinterpreted that it could be applied to music signals in general. Due to the absolute category rating subjective tests [1], [2] used in the development of POLQA, where subjects do not get a reference signal, and due to the focus on speech quality, one would need to further develop the underlying perceptual and cognitive model before it can be applied to music. For the assessment of music signals at 48 kHz sampling frequency, the PEAQ (Perceptual Evaluation of Audio Quality) [12] algorithm according to ITU-R BS.1387 [13] still represents the state-of-the-art standard of a perceptual objective quality measure.

This paper (Part I) provides an overview of the temporal alignment used in the POLQA standard, including the general requirements, a short introduction to the subjective test methodology, and basics of the measurement approach (Sections 1, 2, and 3). The perceptual model, an overview of the subjective tests and the performance of the standard in comparison to the performance of PESQ are given in Part II.

1 REQUIREMENTS FOR THE FOLLOW UP OF PESQ P.862

A first requirement for the follow up of PESQ P.862 is that it has to be technically compatible with existing and previously standardized speech quality measures using the so-called full-reference approach. In this approach an undisturbed reference signal is compared with the degraded output signal to be scored and the system under test is considered as a black box, there is no further information available besides the reference and degraded speech signal. The method should predict the speech quality on a five-point MOS scale as used in absolute category rating listening tests in order to have the closest possible match between objective and subjective measurements.

Furthermore, it is clear that the new method should be more accurate than PESQ and that it should allow assessing degradations introduced by new speech processing technologies and degradations for which PESQ was not designed:

- New and advanced coding technologies;
- Voice quality enhancement devices;
- Time stretching and compression techniques;
- Influence of the acoustic coupling devices and the room acoustics during insertion / recording;
- Influence of the loudness of presentation of the speech signal;
- Influence of linear distortions and spectral shaping (“frequency response”);
- Influence of bandwidth (intermediate bandwidth to common telephony bands).

This requires a huge amount of test data where degraded speech is scored by subjects using a wide variety of different types of distortion. Besides the new types of degradation mentioned, the degradations for which PESQ is known to provide accurate predictions also have to be included in the set of POLQA requirements. This leads to the following set of degradations for which POLQA has to provide accurate results.

- Single and tandemed speech codecs as used in telecommunication scenarios today;
- Packet loss and concealment strategies (packet switched connections);
- Frame- and bit-errors (wireless connections);
- Interruptions (such as unconcealed packet loss or handover in GSM);
- Front-end-clipping (temporal clipping);
- Amplitude clipping (overload, saturation);
- Effects of speech processing systems such as noise reduction systems and echo cancellers on clean speech;
- Effects of speech processing systems such as noise reduction systems (adaptation phase and converged state) and echo cancellers on pre-noised speech;
- Effects of speech-coding systems on pre-noised speech;
- Variable delay (Voice-over-IP, video-telephony) and time warping;
- Gain variations;
- Influence of linear distortions (spectral shaping), also time variant;
- Non-linear distortions produced by the microphone / transducer at acoustical interfaces;
- Voice enhancement systems in networks and terminals and their effects on listening quality;
- Reverberations caused by hands-free test setups in defined acoustical environments.

The new method has to deal with all these types of distortion and has to correctly assess the relative ranking across different distortion types.

Due to the wide range of different distortion types, special requirements on the setup of the subjective experiments are needed. In order to produce training and validation data for the model, dedicated subjective tests were set up where

Table 1. ACR listening quality opinion scale [1], [2] used in the development of POLQA. The average score over a large set of subjects is called MOS-LQS (Mean Opinion Score Listening Quality Subjective).

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

different distortion types in a well balanced ratio were presented to human listeners.

The objective measurement method jointly developed by OPTICOM, SwissQual, and TNO under the name POLQA (Perceptual Objective Listening Quality Assessment) fulfills all the requirements set to the follow up of PESQ P.862 and shows very good performance over a wide range of distortion types.

2 SUBJECTIVE TEST METHODOLOGY

The development of a good objective speech quality measurement method requires large amounts of reliable subjective data. These data consist of reference and degraded speech files with a subjective quality score on each degradation [1], [2]. As the focus of the POLQA development was on super-wideband speech, whereas most subjective speech quality tests today are for narrowband or wideband speech, new experimental procedures were developed. In addition to extended requirements for the wider bandwidth and the lower noise floor than usually required for telephony tests, the constraints in the test design were much more challenging to minimize any context effects in the subjective experiments. A complete description of the subjective test procedure is beyond the scope of this paper but the main points are given in the next paragraphs and more details will be presented in Part II of this paper.

Generally the tests were conducted very similar to P.800, using at least four different speakers per test. Results are expressed in terms of Mean Opinion Scores for Listening Quality Subjective (MOS-LQS) on an absolute category rating (ACR) five-point opinion scale [1] [2] (see Table 1). Listening was performed using high quality headphones. To reduce context effects a relatively large set of common test conditions was included in every newly conducted experiment.

3 BASICS OF THE POLQA APPROACH

The basic idea behind the POLQA algorithm is the same as used in the PSQM and PESQ algorithms. They all use the same full reference approach where a reference input and degraded output signal are mapped onto an internal representation using a model of human perception (see Fig. 1). The difference between the two internal representations is used by a cognitive model to predict the perceived speech quality of the degraded signal. This perceived listening quality is expressed in terms of MOS-LQO (Mean Opinion Score Listening Quality Objective). The MOS-LQO obtained by POLQA was shown to have a very high correlation with the MOS-LQS (Mean Opinion Score Listening Quality Subjective), which is the average quality score over a large set of votes by human subjects using the five-point ACR (Absolute Category Rating) opinion scale [1], [2] of Table 1.

It is important to realize that in ACR type listening experiments, subjects cannot directly compare the degraded signals to the reference speech signals in order to judge the quality. They might therefore also experience different qualities for different original reference voice recordings if those are presented as if they were degraded signals. This will lead to the fact that different reference voice recordings may get different MOS scores although they are not degraded by any distortion.

An important improvement in the POLQA approach is that quality differences in the reference recordings are compensated by a process that the authors have coined as “idealization.” In this idealization process, the timbre of the voice is changed toward a global preferred timbre and low levels of noise that are always present in a recording are suppressed. The internal representations that are used by the POLQA cognitive model to predict the perceived speech quality are therefore calculated on the basis of an idealized input signal representation (see Fig. 2) that uses the psychophysical equivalents of frequency (measured in Barks) and intensity (loudness measured in Sones).

A consequence of this “idealization” approach is that a technically transparent system can be judged as being perceptually non-transparent when poor reference recordings are used. This is a result of using P.800 ACR experiments for subjective voice quality assessments where subjects do not use a direct comparison between a reference and degraded speech file but only rate a degraded file on the basis of an unknown internal “ideal.” One should also be aware of the fact that any perceptual measurement approach does not directly measure the quality of the system under test but instead measures the quality of the output signal of the system. The quality of the system can thus only be measured by averaging the quality over a large set of relevant test signals (speech and/or music). The perceptual measurement approach is necessary in characterizing modern voice and audio processing systems since these systems are strongly time-varying and non-linear.

One distortion type often found in modern voice and audio systems that is completely out of the scope of PESQ is global temporal compression and expansion, e.g., as found in many Voice-over-IP systems (VoIP). This compression and expansion can be implemented in two very different methods. One method does preserve the pitch frequency of human voice, while the other method simply varies the sample rate and thus also the pitch frequency of the speech signal. For POLQA in order to be able to deal with non-pitch preserving time compression and expansion a sample rate detector was developed, and any global differences in

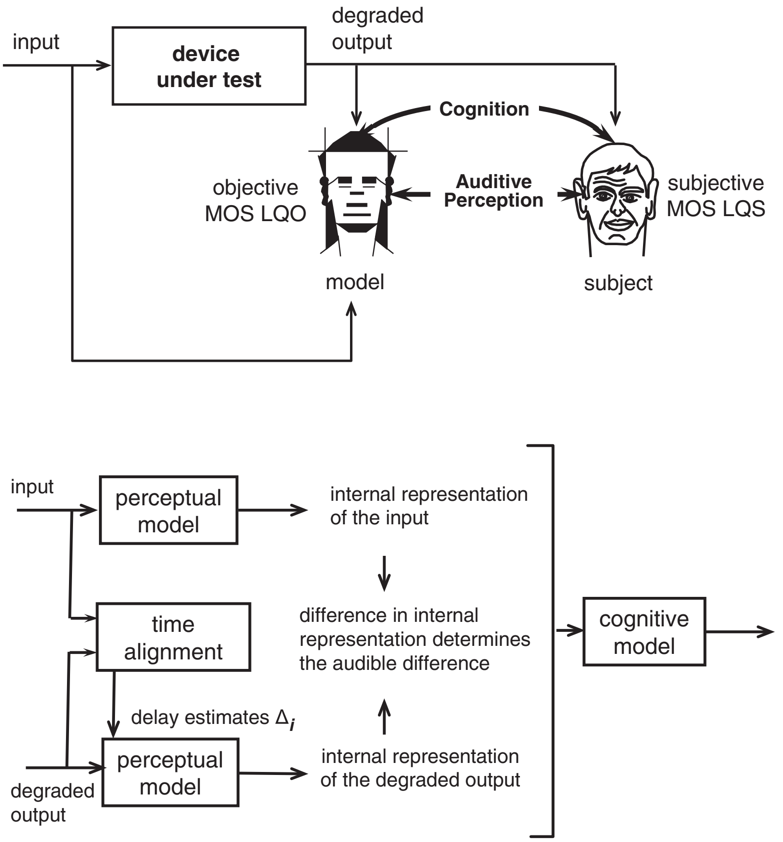

Fig. 1. Overview of the basic philosophy used in PSQM and PESQ. A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the degraded output with the original input, using alignment information as derived from the time signals in the time alignment module. The subjective MOS is referred to as MOS-LQS (Mean Opinion Score Listening Quality Subjective), the objective score as MOS-LQO (Mean Opinion Score Listening Quality Objective).

sample rate between the reference and the degraded signal are compensated by a sample rate conversion. In a first step, the delay variation between the two input signals is determined and the relative difference between the global sample rate of the two signals is estimated (see Fig. 3).

The global sample rate estimation is based on the delay information calculated by the temporal alignment. If the global sample rate differs by more than approximately 0.5%, the signal with the higher sample rate is down sampled. The result is stored together with an average delay reliability indicator, which is a measure for the quality of the delay estimation. The delay estimation results after the global resampling step are compared to the results without resampling and the most reliable one is finally chosen. Once the correct delay is determined (Section 4) and the sample rate differences have been compensated, the signals and the delay information are passed on to the perceptual model (see Part II of the paper), which calculates the perceptibility as well as the annoyance of the degradations and maps them to a MOS scale.

4 POLQA TEMPORAL ALIGNMENT OF THE REFERENCE AND DEGRADED SIGNAL

The temporal alignment is a challenging problem and we start by describing a number of these challenges (Section 4.1) while the basic concepts and elements are given in Section 4.2. In the next section (Section 4.3) the pre-alignment module is described, which has two different approaches, a fast pre-alignment for signals with fixed or at least piecewise fixed delays (Section 4.3.1) and a thorough pre-alignment (Section 4.3.2) for signals with highly variable delays. The result of the pre-alignment is a rather rough estimation of the delay for a few sections of the input signals. This rough estimation is further refined in the subsequent steps coarse alignment (Section 4.4) and fine alignment (Section 4.5). Section 4.6 describes how sections with an almost constant delay are joined together before passing on the information to the psychoacoustic model.

Finally, Sections 4.7 and 4.8 give a more detailed description of the resampling that is necessary when the reference and degraded file show a global time compression or expansion. This resampling is intertwined with the temporal alignment as explained in Section 3, the basics of the POLQA approach (see Fig. 3).

4.1 Challenges

The mathematical problem of temporally aligning two signals can be reduced to the problem of determining the temporal offset between the two signals at any given point in time. This very general task is well researched and simple for almost identical time series. Usually, detecting the

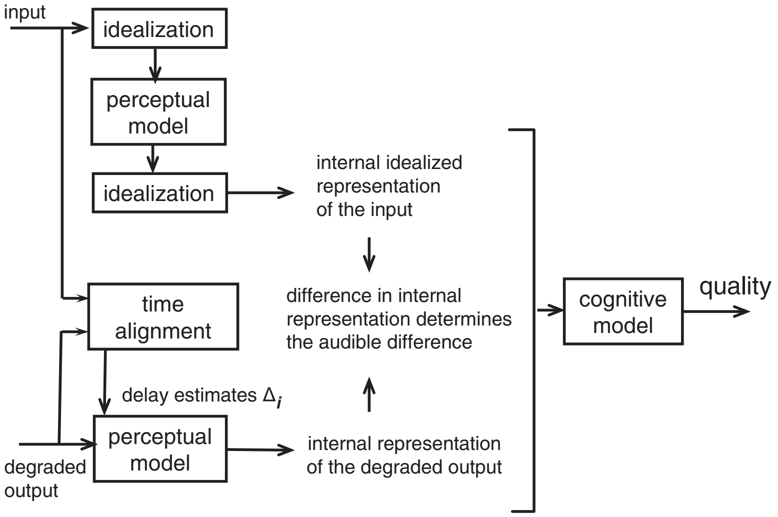
input
idealization
perceptual model
idealization
internal idealized representation of the input
difference in internal representation determines the audible difference
time alignment
delay estimates Δ_i
perceptual model
degraded output
internal representation of the degraded output
cognitive model
quality

Fig. 2. Overview of the basic philosophy used in POLQA. A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the degraded output of the device under test with the idealized input, using alignment information as derived from the time signals in the time alignment module.

maximum of a cross-correlation function between those signals will be sufficient, since the position of the maximum is identical to the delay offset. The task gets a little more difficult if one of the two signals is distorted and contains, e.g., dropouts. In this case it helps to calculate several cross-correlation functions at different positions in the signals, build a histogram of the found positions of the maxima, and finally search the peak in this histogram.

This method works reliably for signals that have a delay (= temporal offset) that is at least piecewise constant. Each section with a constant delay should thereby be at least as long as the cross-correlation length. Further difficulties

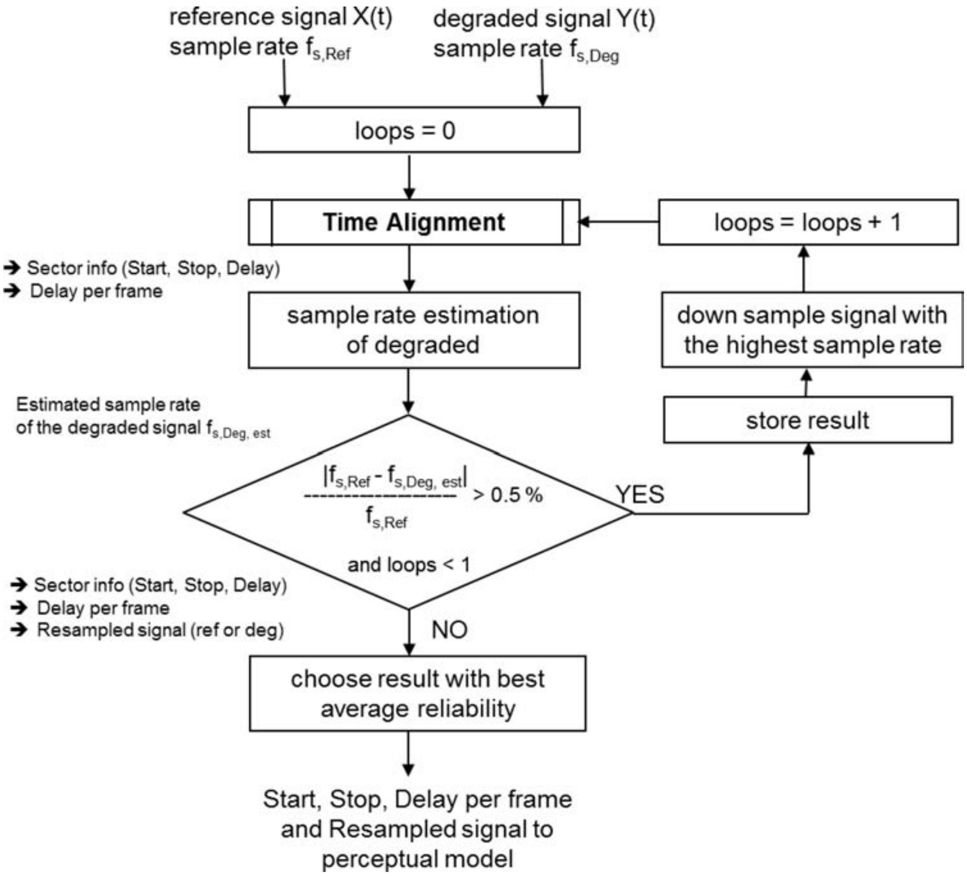
reference signal X(t)
sample rate $f_{s,Ref}$
degraded signal Y(t)
sample rate $f_{s,Deg}$
loops = 0
Time Alignment
loops = loops + 1
➔ Sector info (Start, Stop, Delay)
➔ Delay per frame
sample rate estimation of degraded
down sample signal with the highest sample rate
store result
Estimated sample rate of the degraded signal $f_{s,Deg,\,est}$
$\frac{|f_{s,Ref} - f_{s,Deg,\,est}|}{f_{s,Ref}} > 0.5\,\%$
and loops < 1
YES
➔ Sector info (Start, Stop, Delay)
➔ Delay per frame
➔ Resampled signal (ref or deg)
NO
choose result with best average reliability
Start, Stop, Delay per frame and Resampled signal to perceptual model

Fig. 3. Overview of the resampling strategy used in POLQA. If the sample rates differ by more than 0.5% a down sampling of the signal with the higher sample rate is carried out once, and the result with the best average reliability is chosen.

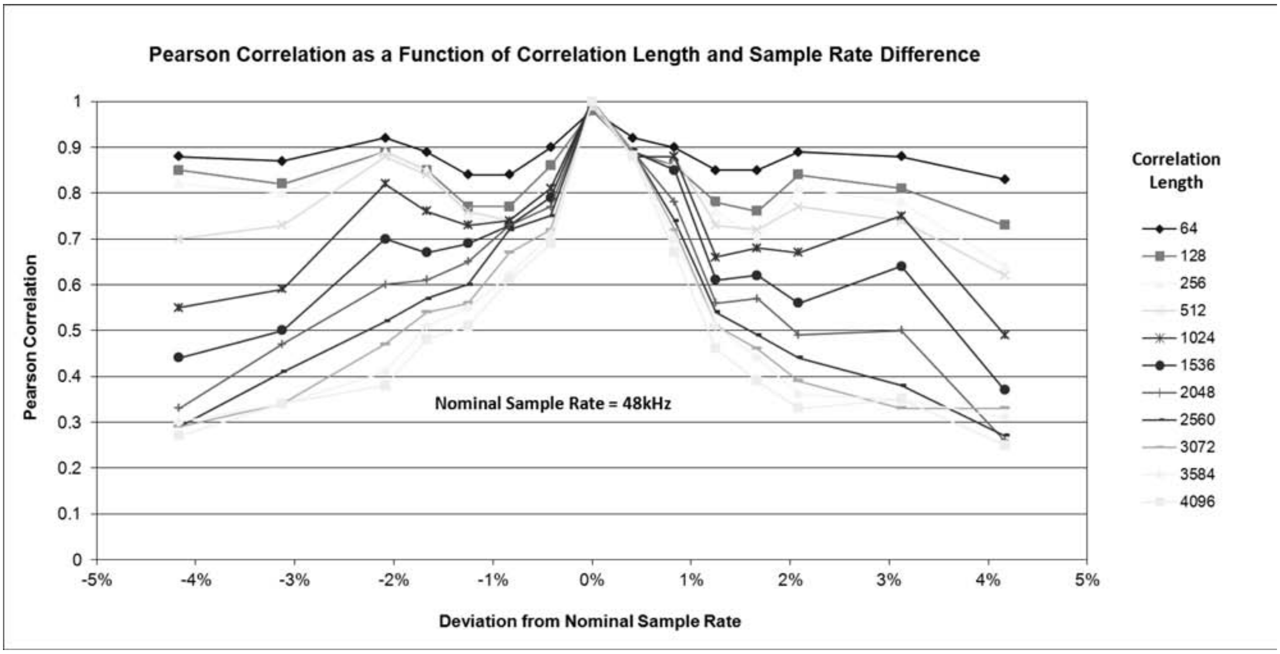

Fig. 4. The average Pearson correlation as a function of temporal compression or stretching (expressed as sample rate variation from a nominal frequency of 48 kHz) and correlation length for a typical speech signal without any additional distortions.

arise if the signals that shall be temporally aligned contain periodic sections, which is typically the case for, e.g., voiced sounds of speech. In this case the cross-correlation function will show multiple maxima. The simplest solution to overcome this is to use rather long cross-correlations, so that voiced and non-voiced sections are spanned by one cross-correlation function. A simplified version of this method is successfully used in, e.g., PESQ P.862 [5] – [9], where this method is implemented by searching each entire speech utterance of the reference signal in the degraded signal. This specific implementation can handle different delays for each utterance plus, under some circumstances, a maximum of one delay variation per utterance. As far as the application to speech signals is concerned, the described methods have the two major limitations of requiring fairly long signal sections with constant delay and the need of calculating cross-correlation functions over rather long windows. This, however, conflicts severely with the task of finding the correct delay between two signals that have non-constant delays, as is typically the case with temporally compressed or stretched signals, where the delay varies even on a frame by frame basis. Those effects are described in more detail in the next section.

4.1.1 Temporal Compression / Stretching of the Degraded Signal (Time Scaling)

In the past, time scaling was a mere technical problem in measurement equipment that was missing synchronization between DA and AD converter clocks and could be overcome by proper system design. Today, however, time scaling is frequently used as a method to compensate for packet losses and to conceal the effects of jitter buffer length adaptations in the transmission system itself. Also, some codecs produce slightly varying delay.

Time scaling is applied in two very different flavors. The first and simpler method is to up- or down sample sections of the signal. While being relatively simple, this method however has the big disadvantage that it also modifies the pitch frequency of the speech signal, which may be perceived as a distortion. The second flavor, time scaling with pitch preservation, is more advanced since it achieves the same signal compression or stretching by preserving the pitch frequency of human speech. In general this is carried out by repeating parts of the signal and "gluing them together" by using an overlap and add algorithm. The disadvantage of the second method is, however, that it will usually fail for non-speech signals. From a temporal alignment point of view those two methods require very different alignment concepts.

From a signal processing point of view, time scaling without pitch correction can be modeled as looking at two signals with slightly differing or varying sample rates. For simplification it can be assumed that the ratio of the two sample rates is section-wise constant. To properly align those signals, it is enough to detect the difference between the sample rates and to reverse the up or down sampling accordingly ("resampling"). The challenge here is the detection of the sample rate difference, which is explained in detail in Section 4.7. Without resampling, the temporal alignment that is based on the cross-correlation between two signals is very likely to fail since the signals to be compared have increasingly lower similarity with increasing amounts of stretching or compression. The longer the signal windows used for correlation are, the stronger this effect will be. Fig. 4 illustrates this problem. Here, the correlation between a reference signal with 48 kHz sample rate and resampled versions of the same signal are presented. The length of the correlation was used as a parameter. It can be seen that the achievable peak correlation for a certain percentage of temporal compression decreases dramatically with increasing correlation length. If this would be used in the alignment of a speech quality measurement algorithm, the probability of misaligned signal sections would increase significantly.

Time scaling with pitch correction or time scaling introduced by PSOLA (Pitch Synchronous Overlap Add) like algorithms can be handled much easier. In this case the temporal and spectral structure and especially the envelope of very short signal sections are mostly preserved. The delay can therefore be determined by normal correlation-based temporal alignment routines without major difficulties. Of course, the alignment routine must still be capable of calculating different delay values for many, very short signal sections. Generally, however, the delay variation from frame to frame caused by time scaling is very small, typically a few samples only. It is, therefore, not required to resample such signals; resampling would even skew the measurement results, since it would again change the pitch frequency of one signal and thus influence the perceptual model.

4.1.2 Summary of the Requirements

To summarize, the requirements for a delay search method suitable for a perceptual measurement algorithm today are thus:

- Correlation lengths must be short;
- The signal section that is searched must be long enough to avoid problems with periodic signal sections;
- Delay offsets in the range of a few seconds must be handled;
- The method must be suitable to handle delays that vary every few milliseconds;
- The method must be very robust against stationary and non-stationary background noise, transmission errors like, e.g., packet loss, and coding distortions;
- In order to cope with time scaling effects, the method must be able to properly align signals where the sample rates of the reference and the degraded signal differ slightly.

To overcome those apparently contradictive requirements, POLQA uses some novel concepts. One central point is to conduct the temporal alignment in several steps, where the found delay is refined from step to step. This stepwise refinement permits using fairly short search ranges in each step, since only the relative offset to the best found delay so far has to be searched.

This concept is achieved by using down sampled versions of the signals or feature vectors in the early stages of the alignment. This reduces the required correlation length at the cost of accuracy (if 128 samples are down sampled to one single sample, the alignment can't be more accurate than ±128 samples). To gradually improve the accuracy, the amount of down sampling is then stepwise reduced and the delay is refined in each step. While this sounds simple, the actual implementation is rather complex in order to make the method robust against transmission errors and especially high and non-stationary noise levels. The part of the algorithm that is most important for this robustness is the pre-alignment (Section 4.3) since it has to deal with the largest delay search range. The stepwise refinement of the delay happens mostly in the coarse alignment (Section 4.4) and the fine alignment (Section 4.5). The latter two steps only refine the initial delay values that were determined by the pre-alignment. The main problems for the coarse and the fine alignment are periodic signal sections and effects caused by non-stationary noise.

Table 2. Macro frame sizes used in the temporal alignment.

Sample rate:	48 kHz	16 kHz	8 kHz
Macro frame size [samples]	1024	512	256

4.2 Basic Concepts and Elements

The global concept of the temporal alignment algorithm is to limit the search ranges as much as possible and to stepwise refine the delay estimate. To achieve this, the early stages of the alignment are operating on heavily down sampled vectors. Each subsequent step uses less down sampling than the previous one and only needs to search a more accurate delay value around the last found delay. Consequently, the early stages of this hierarchical alignment approach are the most critical ones since they permit the longest search range, which is equivalent to a higher risk of false delay estimations.

The temporal alignment consists of the major blocks filtering, pre-alignment, coarse alignment, fine alignment, and section combination. The input signals are split into equidistant macro frames, the length of which is dependent on the input sample rate (see Table 2) and not necessarily the same as the ones used in the perceptual model. The delay is determined for each macro frame. The calculated delay is defined as the delay of the reference signal relative to the degraded signal.

The pre-alignment determines the active speech sections of the signals, calculates an initial delay estimate per macro frame, and an estimated search range required for the delay of each macro frame (i.e., a theoretical minimum and maximum variation of the detected initial delay). This pre-alignment exists in two versions, one that is optimized for speed and is generally sufficient to solve simple alignment problems (e.g., those without time scaling), and one version that performs a more thorough delay search in multiple dimensions and at the cost of computational efficiency. The choice between the two pre-alignments is made on the basis of a reliability estimation (see Fig. 5).

After the pre-alignment a coarse alignment is carried out with an iterative refinement of the delay per macro frame, using a multidimensional search and a Viterbi-like backtracking algorithm to filter the detected delays

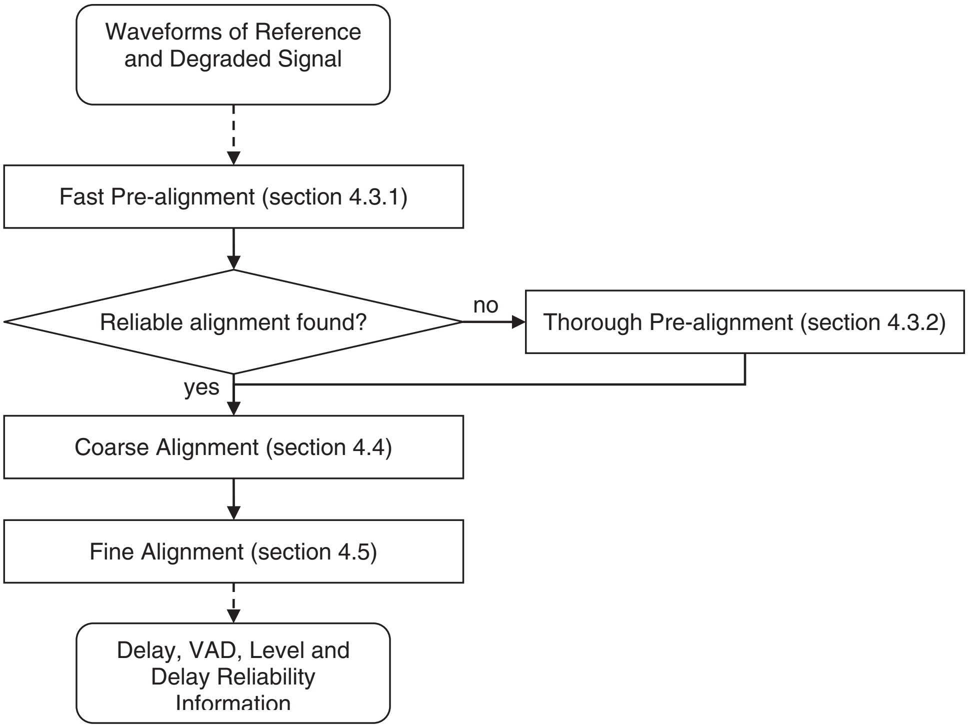
Waveforms of Reference and Degraded Signal
Fast Pre-alignment (section 4.3.1)
Reliable alignment found?
no
Thorough Pre-alignment (section 4.3.2)
yes
Coarse Alignment (section 4.4)
Fine Alignment (section 4.5)
Delay, VAD, Level and Delay Reliability Information

Fig. 5: Overview of temporal alignment modules in POLQA.

(Section 4.4.2). The resolution of the coarse alignment is increased from step to step in order to keep the required correlation lengths and search ranges small.

The fine alignment finally determines the sample exact delay of each macro frame directly on the input signals. Again, a backtracking algorithm is used to smooth out short extreme delay variations. The search range of the fine alignment is determined by the accuracy of the last iteration of the coarse alignment. In a final step, all sections with almost identical delays are combined to form the so-called "Section Information."

4.2.1 General Delay Search Method and Delay Reliability Measure

Most of the modules related to the temporal alignment use the same method to find the lag (= delay, offset) between two signal sections. This method is based on the analysis of a histogram that is created by calculating the cross-correlation function between two windows taken from the signal sections, entering the found peak value into the histogram, shifting both windows by a small amount, and repeating this step again. Once the histogram contains enough values, it is filtered and the peak is determined. The position of this peak in the histogram is equivalent to the delay offset between the two signal sections.

In most steps of the alignment a vector indicating the reliability of the best so far found delay for each signal section is maintained. Typically the Pearson correlation between the associated reference and degraded signal is used as the reliability measure.

4.2.2 Bandpass Filter

In order to minimize the impact of noise on the alignment both the reference and degraded signal are bandpass filtered. As most speech energy lies in the range of 300 to 3500 Hz this part of the signal will provide the most reliable delay estimation. Background noise often has a 1/x-shaped energy distribution along the frequency axis in which case the filtering out of frequencies below 300 Hz significantly reduces the impact of this noise. The final filter shape depends on the operating mode of the model (narrowband or super-wideband). In the super-wideband operating mode, the signals are bandpass filtered to 320 Hz up to 3400 Hz. In the narrowband operating mode, the signals are bandpass filtered to 290 Hz up to 3300 Hz. This filtering ensures that the alignment operates on the speech signals and not on the background noise, etc. The exact values were optimized to provide optimal performance of the model.

4.3 Pre-Alignment

4.3.1 Fast Pre-Alignment for Fixed or Piecewise Fixed Delays

In many cases, the delay in the degraded signal is fixed or piecewise fixed. In case of local distortions, like temporal clipping or concealed frame loss, a temporal alignment module designed to detect very small delay changes may be too sensitive and incorrectly measure variable delay. A delay estimation with assumed fixed or piecewise fixed delays is therefore attempted first. The assumed simple delay characteristic allows for a computationally inexpensive delay estimation.

The fast pre-alignment module runs once at the beginning of the temporal alignment procedure. It uses sequential cross-correlation to subdivide the degraded speech signal into segments with constant delay (“matched segments”), and thus allows constraining the delay search range for these segments in subsequent modules of the temporal alignment. Larger delay search ranges will only be used in segments for which no sufficient cross-correlation maximum (“match reliability”) was found. To prevent misaligning signals that feature continuous variable delay or resampling, POLQA reverts to thorough pre-alignment (see Fig. 5) if less than 75% of active speech can be matched reliably, or if the determined piecewise constant delays change continuously throughout the signal. If resampling was detected, POLQA directly applies thorough pre-alignment when temporal alignment is re-entered (see Fig. 3), since the basic assumption of a simple delay characteristic no longer holds.

The fast pre-alignment consists of the following steps:

- Down sample a copy of the input signals to 8 kHz for further processing, apply a 700–3000 Hz bandpass filter and rescale the digital active speech level (ASL) to –26 dBov (Section 4.3.1.1);
- Compute an estimate of the average signal delay by calculating the cross-correlation of the signal envelopes (Section 4.3.1.2);
- Further process the input signals by normalizing the power of active speech portions using a sliding window, and thresholding the remaining inactive signal portions to zero (Section 4.3.1.3);
- Subdivide the reference speech signal into smaller segments. Determine the delay of each corresponding segment in the degraded speech signal by means of a weighted cross-correlation of the power-normalized signals (Section 4.3.1.4);
- Set the delay search range of each segment based on the obtained cross-correlation maximum. The delay search range of speech pauses is inferred from the search ranges of neighboring active speech (Section 4.3.1.6);
- Detect possible outliers in the measured segment delays through statistical analysis and avoid misalignments by extending the delay search range for those segments (Section 4.3.1.5).

4.3.1.1 Preprocessing The average energy distribution of speech is highly skewed toward low frequencies, it is thus important to avoid an excessive bias of the calculated cross-correlations by these frequencies. On the other hand, the perceptually relevant parts of voiced sounds are located in the frequency range below 5 kHz, while unvoiced sounds are spectrally flat for the most part (and less critical to perceptual speech quality estimation).

A 700–3000 Hz bandpass filter is used as a middle ground between these two considerations. This is in addition to the bandpass filter of Section 4.2.2. Both input signals are rescaled to a digital active speech level (ASL) of –26 dBov after filtering.

As part of preprocessing, a common active speech threshold *thr* for both input signals is estimated:

$$thr = \min \times \left(\frac{-26 + 3 \cdot \max\left(noiseLev_{ref}, noiseLev_{\text{deg}}\right)}{4}, -29 \right) [dBov] \qquad (1)$$

where *noiseLev* denotes the average noise level in dBov of the respective signal. This threshold will be used to exclude signal parts with insufficient signal-to-noise ratio (SNR) from the cross-correlation process.

4.3.1.2 Average Delay Estimation An envelope-based delay estimation is first used to determine the average signal delay. The RMS power envelope is computed in intervals of 180 ms frames with a 50% overlap between frames for both filtered signals. The previously determined active speech threshold *thr* is then subtracted from the envelope values (also measured in dBov), resulting in positive values for speech active parts. The remaining negative envelope values for non-active speech are clipped to zero. Level variations, e.g., due to active gain control systems, are compensated through sliding window power normalization of the processed envelopes using a window length of 450 ms.

The location of the cross-correlation maximum of the processed envelopes is used as a first estimation of the average signal delay. This first estimation already reduces the number of cross-correlation lags to compute in the following calculations.

4.3.1.3 Sliding Window Power Normalization As in the envelope-based delay estimate, sliding window power normalization is used to reduce the influence of level variations on the cross-correlation of both signals. The normalization is carried out by calculating the RMS of the signal amplitudes in a sliding window of 26.625 ms length and normalizing (i.e., dividing) the sample in the center of the current window position by this RMS value. Signal samples with a window RMS value below the active speech threshold *thr* are clipped to zero amplitude (digital silence). The signal level of speech often decreases gradually at the end of an utterance; therefore a threshold hysteresis is used to continue normalization for an additional 70 ms of speech even after the window RMS has dropped below *thr*.

4.3.1.4 Segment-Wise Delay Estimation This step determines the delay and delay search range for each active speech segment in the reference, i.e., the segments are “matched” with corresponding portions in the degraded signal. The approach for this step is based on the following considerations:

- There is no guarantee that the reference speech content is present in its entirety in the degraded signal. In particular, some speech may have been muted, covered by noise, or simply lost. However it is safe to assume that the monotonicity of the speech content is preserved, i.e., the position in time of an earlier

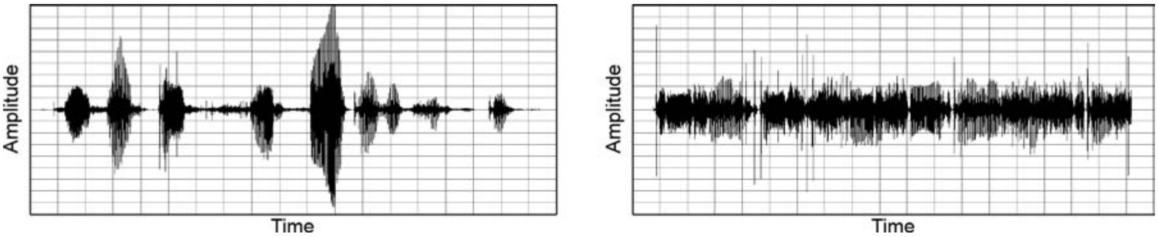

Fig. 6. Example of a preprocessed speech signal before (left) and after sliding window power normalization (right). Note the almost flat signal envelope of active speech after normalization.

speech portion in the reference signal remains earlier in the degraded signal.

- In case of consecutive frame losses, error concealment strategies usually consist in repeating the last decoded frame while gradually decreasing the signal level. Thus the level of concealed lost frames will tend to drop faster below the cross-correlation threshold *thr* if the original level of the affected utterance was already low.
- Similarly, voice activity-switched processing like noise suppression and discontinuous transmission (DTX) are largely based on the level of the speech signal. Such processing is triggered by quiet speech parts (especially in case of strong background noise) that are thus more likely to suffer degradations by such systems.

The consequence of these considerations is that segments are matched in decreasing order of signal level in the reference. In other words, the fast pre-alignment starts with the signal segments that are thought to result in the most reliable match. Each successful match divides the remaining signal portions. Because of the assumed monotonicity, the range of possible delays for these portions is limited by the delays of surrounding matched segments.

Therefore the range of possible delays (and hence the risk of misalignments) is decreased as the matching process gradually progresses toward more degradation-prone segments.

The signal envelope of the preprocessed reference signal and a signal level threshold *curThr*, which is about 6 dB greater than *thr*, are used to select a segment. The algorithm looks for a continuous unmatched piece of the reference signal of 64 ms to 1.5 s length with an envelope level greater than *curThr* and where the sum of the envelope frame levels is the highest (the "loudest unmatched segment"). This segment in the power-normalized reference is cross-correlated with available (i.e., not yet matched) portions of the power-normalized degraded signal. Depending on the value of the obtained cross-correlation maximum, the matched segment will be categorized as either *assigned* or *unsure*. If no unmatched portions are available in the degraded signal anymore, the segment will be categorized as *missing*. One-hundred-eighty degree phase shifts in the degraded signal are handled by using the absolute values of the cross-correlation vectors. After matching all segments with an envelope level above *curThr*, the matching procedure continues with the segments with envelope levels above the active speech threshold *thr*. Fig. 7 provides a schematic overview of the segment-wise matching process.

When the matching process starts, segments with an insufficient cross-correlation maximum are ignored and kept for later matching. This is in line with the aforementioned approach of avoiding misalignments by starting with the most reliable matches. Once the matching process has progressed to segments of envelope levels below *curThr*, these skipped segments are matched using the delays of surrounding segments and marked as *unsure* for possible later correction. Any remaining unmatched parts of the reference signal, including speech pauses, are matched as *unsure* segments using the same method as well.

The cross-correlation calculation step in the flowchart of Fig. 7 is kept computationally inexpensive by not using the method of accumulating multiple cross-correlation maxima in a histogram described in Section 4.2.1. Rather, it applies a simple bell-shaped weighting on the cross-correlation results before searching for the maximum absolute value. The location of the weighting function maximum corresponds to the expected delay for the current segment, inferred from the averaged delays of the two nearest surrounding matched segments. The measured delay is thus biased by the delays of previously matched segments.

The amplitude of the weighting function (strength of bias) is increased when the distance of the surrounding segments to the current one is small (as there is less room for delay changes to occur), or when the delay difference between these two segments is small (better agreement of averaged delays). As the number of matched segments increases, the delay calculation for later, more error-prone segments is therefore influenced more and more by previously matched segments.

4.3.1.5 Statistical Post-Processing of Calculated Segment Delays The fast pre-alignment works by first matching signal segments that are assumed to have the lowest risk of misalignment. A post-processing step at the end of the segment-wise matching process is used to detect and correct misalignments in the first matched segments, for which a cross-correlation weighting using previously matched segments is not possible. This post-processing uses simple descriptive statistics to detect outliers in the delays of segments. The reference signal is split into two halves at the position of the longest speech pause and the first and third quartiles Q_1 and Q_3 of all *assigned* segment delays within

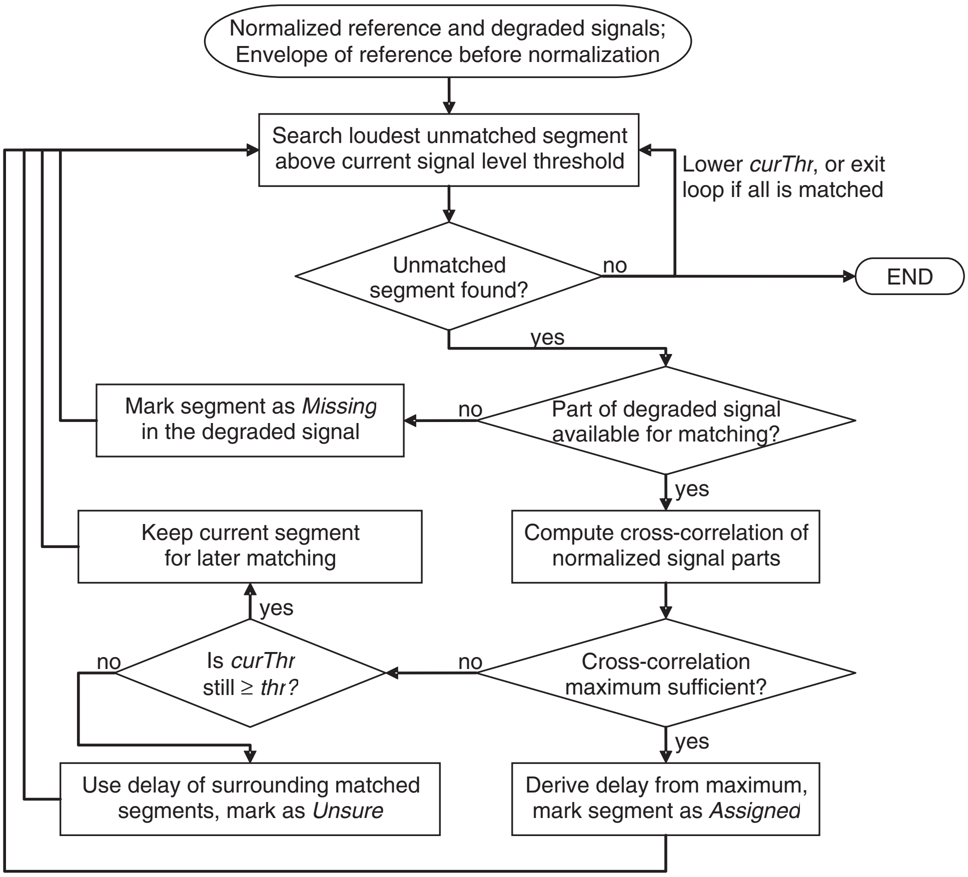
Normalized reference and degraded signals; Envelope of reference before normalization
Search loudest unmatched segment above current signal level threshold
Lower curThr, or exit loop if all is matched
Unmatched segment found?
no
END
yes
Mark segment as Missing in the degraded signal
no
Part of degraded signal available for matching?
yes
Keep current segment for later matching
Compute cross-correlation of normalized signal parts
yes
no
Is curThr still ≥ thr?
no
Cross-correlation maximum sufficient?
yes
Use delay of surrounding matched segments, mark as Unsure
Derive delay from maximum, mark segment as Assigned

Fig. 7. Flowchart of the segment-wise matching process.

each half are computed. The statistical lower and upper outer fences of the delay values are then given by

$$\begin{aligned} lower\ fence &= Q_1 - 3(Q_3 - Q_1) \\ upper\ fence &= Q_3 + 3(Q_3 - Q_1) \end{aligned} \tag{2}$$

for each half. Segments of type *assigned* or *unsure* with delays outside these fences are flagged as misalignments. Adjacent segments of type *missing*, typically a by-product of misaligned speech segments, are flagged as well. Each group of consecutive flagged segments is replaced by a single segment of type *unsure* with a new delay value inferred from surrounding matched segments.

4.3.1.6 Creation of Per Macro Frame Information The determined *assigned*, *unsure,* and *missing* segments are used to generate the per macro frame delay information vectors needed for the subsequent temporal alignment modules in POLQA. The degraded signal is traversed in increments of the macro frame length and the segment corresponding to each macro frame is determined.

- For macro frames that fall within segments of type *assigned*, the delay search range is set to zero. Correspondingly, the delay reliability measure for these frames is set to 1.0.
- If the corresponding segment is of type *unsure*, the closest surrounding *assigned* segments are searched. Differences between the estimated delay of these segments and the current *unsure* delay set the delay search range, while the delay reliability measure is set to the cross-correlation maximum that was obtained during the segment-wise matching process.

The following modules of the POLQA temporal alignment only search the reference for correspondences of the degraded signal, thus segments of type *missing* are not needed for generating the per macro frame delay information vectors. They are merely used as auxiliaries during the segment-wise matching process.

The degraded signal may also contain inserted parts (e.g., extended speech pauses), which do not correspond to any segment. For macro frames in such parts of the degraded signal, the averaged delay of the closest surrounding *assigned* segments is used, and the delay search range is set to about half the macro frame size. Finally, the delay reliability measure for these frames is set to zero.

4.3.2 Thorough Pre-Alignment

The thorough pre-alignment calculates the same information as the fast pre-alignment does. However, the thorough version uses a significantly more robust algorithm—at the cost of significantly higher processing requirements. This part of the algorithm is only used if the delay estimation from the fast pre-alignment was not considered sufficiently reliable, as described in Section 4.3.1.

The key element is the determination of an initial delay estimate at so-called reparse points. Reparse points are the start of continuous signal sections that are expected to contain no significant delay variation (significant meaning more than approximately 400 ms). Typically these points are located where the signal transits from pause to active speech. The signal sections following a reparse point are called reparse sections.

For each reparse point the reparse section information is calculated. This section information stores the position of the beginning and end of the section, an initial value for the delay of said section as well as an indication of the reliability of the estimated delay and its accuracy, i.e., an upper and lower bound within which the accurate delay is expected to be found. The general process is depicted in Fig. 1. The details of each step are explained in the subsequent subsections.

The robustness of thorough pre-alignment is mostly achieved by a "brute force" method. The delays are determined by trying to correlate several feature vectors (fractal dimension and signal energy) using different correlation lengths at varying positions within the sections and finally choosing the version that resulted in the highest delay reliability. In order to save processing time, some processing steps are skipped as soon as a sufficiently reliable delay has been found.

If the initial assumption that no major delay variations occur within a reparse section is violated, then the subsequent alignment steps will most likely not be able to correctly refine the delay, and the resulting MOS values will be very pessimistic. However, in general such gross delay jumps during active speech are also very annoying degradations and the resulting subjective scores will be very poor as well.

4.3.2.1 Determination of the Delay Limits. This simple step tries to identify some reasonable upper and lower limit for the overall delay search range. The decision is based on the following assumptions:

- The reference file has at least 40% activity and consists of at least two sentences;
- The total amount of silence is split into at least two sections (typically three);
- Not more than 50% of the silence fall before the start or after the end of active speech;
- The active speech part is not cut off at either end due to the delay.

This results in a search range according to the following formulae:

$$\begin{aligned} DelayHigh &= \max\left(2.5s, \frac{StartsampleIndex_{\mathrm{Ref}}}{SampleRate}\right) \\ DelayLow &= \max(-2.5s, -(F_{len,\mathrm{Ref}} * 0.2 \\ &\quad + F_{len,Deg} - F_{len,\mathrm{Ref}}) * 0.8) \end{aligned} \tag{3}$$

With $F_{len,\mathrm{Ref}}$ being the length of the reference signal and $F_{len,\ Deg}$ being the length of the degraded signal in seconds. *StartsampleIndex* is the index of the detected start of the reference signal; this is usually 0, but may be a later point in the signal if it starts with a very long silent period.

4.3.2.2 Overall Delay Estimation. The overall delay is estimated in three steps. First, an attempt is made to match entire signals by calculating the logarithmic power of the signals, averaged over frames with a length that is half the macro frame length, and determining the delay between those vectors using the method described in Section 4.2.1. This results in a first estimate of the overall delay and the overall reliability. The best achievable accuracy of the delay estimate in samples is defined by the window size used to average the signal energies. In a second and third step, the same method is applied to the first and the second half of the signals independently, which results in two more estimates for delay and reliability. If all three delays are of the same range, then the overall delay is accepted and marked as reliable. The tolerance for acceptable delays is reduced for long signals with poor reliability values.

4.3.2.3 Identification of Reparse Points. In this step a voice activity detection algorithm (VAD) is used to determine for each macro frame of the reference and the degraded signal whether it contains active speech or silence. The beginning of sections of consecutive active macro frames is called a reparse point, since those resemble the points at which the delay measurement is completely restarted. The so-called reparse section information contains for each reparse point the position of the active section's start point, the length of the active section, a coarse delay estimate, a reliability indicator for the detected delay, and an estimated range within which the exact delay can most likely be found. The delay determined in this step is simply the difference of the detected reparse section start in the degraded signal minus the start point of the corresponding section in the reference signal and may be very unreliable.

For the reference signal the active frame detection works very reliably; but especially for degraded noisy signals, the detected active sections may be very inaccurate. Therefore, the plain VAD information is by far not sufficient in order to allocate combinations of reference and degraded active sections. Instead, a rather complex method is used. Up to three sets of potential allocations are therefore investigated for each reparse point:

- VAD1: A matched set of allocations plainly based on the VAD information. If the length of the next reference and degraded section does not differ by more than 120 ms for signals with an SNR below 35 dB, or if the length differs by less than 480 ms for signals with better SNR, then the VAD1 section information is marked as valid. This potential allocation calculation does not use correlation.
- Corr1: The reference section is searched in the degraded signal by using the VAD information of the reference signal only and the overall delay estimate as a hint on where to search. The search uses cross-correlation and is performed twice, once at the beginning of the active section and once slightly delayed. The better of the two results is stored. If the sections were long enough and a reasonable match could be

found, then the Corr1 information is marked as the valid allocation set.

- Corr2: The reference section is searched in the degraded signal by using the VAD information of the reference signal and the VAD information of the degraded signal as hints on where to search. Again, this search uses two times a cross-correlation, once at the beginning of the active section and once slightly delayed. The better of the two results is stored. If the sections were long enough and a reasonable match could be found, then the Corr2 information is marked as the valid allocation set.

Now, the best of VAD1, Corr1, and Corr2 is chosen. If the found match is very reliable and the found degraded section is much longer than the reference section, it will be split and a new degraded section is added to the list of sections to be allocated.

All operations in this section are performed on the power of the signals, using frames of the same size as in the overall delay estimation in Section 4.3.2.2.

4.3.2.4 Initial Delay Calculation at Each Reparse Point So far, the delay of each reparse section is very coarse only since it is mostly based on the information retrieved from the VAD. For each active section this delay is now refined by a multidimensional search. The dimensions used are the logarithmic power and fractal dimension. For each dimension the search is performed over two segments from each reparse section and using two different frame sizes. This results in eight estimates for the delay of each section. The estimate with the highest reliability is chosen. If the reliability of a section as it was determined by the reparse point identification is already very high, the search is skipped entirely for this section.

4.3.2.5 Determination of Active Macro Frame Flags from the Reparse Section Information In principle, the VAD information of the degraded signal is used directly in order to mark individual macro frames as active or pause. In addition, however, all macro frames that are outside any detected reparse section are set to pause, regardless of the VAD information. This avoids the wrong treatment of such sections for very noisy signals, where the VAD information might be misleading. The result is a vector that contains for each macro frame a flag that is set when the macro frame is active and is cleared when the macro frame is a speech pause.

4.3.2.6 Creation of Per Macro Frame Information from the Reparse Section Information This process is rather simple. All it does is to copy the information from the reparse section information to the corresponding per macro frame information, making sure that delay changes occur in the middle of the pause between two active sections. This step generates the following vectors:

- *DelayPerMacroFrame*, which contains the estimated initial delay for each macro frame;
- *ReliabilityPerMacroFrame*, which contains an indication of the reliability of the delay estimate for each macro frame;
- *SearchRangeLow*, which contains for each macro frame the lower bound of the range in which the exact delay is expected;
- *SearchRangeHigh*, which contains for each macro frame the upper bound of the range in which the exact delay is expected.

This is exactly the same information as calculated by the fast pre-alignment method and all subsequent alignment operations, starting with the coarse alignment, will work on those vectors.

4.4 Coarse Alignment

The coarse alignment performs a stepwise refinement of the delay per macro frame. This is implemented by subdividing each signal into smaller subsections (“feature frames”) and by calculating one characteristic value (“feature”) for each of those subsections. The resulting vectors are called feature vectors. Feature frames are again equidistant and their length is reduced from iteration to iteration. The length of the feature frames is independent from the macro frame length. Feature frames are generally much shorter. Due to the iterative length reduction, the accuracy of the estimated delay increases with each iteration, but at the same time the explored search range is reduced. Multiple feature vectors are calculated and for each macro frame the most suitable feature is used to determine the delay value for that macro frame.

The result of the coarse alignment is a vector with the delay per macro frame, expressed in samples, with an accuracy that depends on the feature frame length used in the final iteration.

In more detail, the coarse alignment works as follows:

Starting with the lowest resolution (i.e., the longest feature frame length), all feature vectors are calculated for the active sections of both the reference and the degraded signal. The features used are the energy per feature frame and the fractal dimension per feature frame (see Section 4.4.1). Now, the so-called correlation matrix is computed for each feature. This matrix is organized in correlation vectors per macro frame. The correlation vectors contain for each macro frame the correlation between the reference and the degraded feature vector for all possible time lags between *SearchRangeLow* and *SearchRangeHigh* around the best delay per macro frame computed so far. In the first iteration, *SearchRangeLow* and *SearchRangeHigh* are results from the pre-alignment. In subsequent iterations, the search range is determined by the feature frame length of the previous iteration. The resolution of the correlation vectors is identical to the resolution of the feature vectors, i.e., the resolution of the correlation vectors is increased with each iteration and thus a more accurate delay estimate can be determined. The resulting matrix is of the format $N_{cv} \times N_{mf}$, with N_{cv} being the number of possible lags tested in each correlation vector and N_{mf} being the number of macro frames. Next, the correlation matrices for all features are combined by selecting for each macro frame the correlation vector from the feature, which yields the maximum correlation for this macro frame. The position of the maximum correlation in the correlation vectors is equivalent to the optimum correction of the delay per macro frame required to achieve a better match between the two signals.

A problem in searching the position of the maximum correlation in this manner is that it often leads to wild delay variations since speech signals are frequently periodic and in some cases, e.g., when packet loss occurs, a wrong delay would lead to a better correlation than the correct delay. Therefore, a backtracking algorithm similar to Viterbi's algorithm is used to find the best possible path through the resulting combined correlation matrix (see Section 4.4.2). This algorithm starts with the last macro frame and traces the ideal path back to the first macro frame. For each macro frame, the correlations for all lags are weighted with a penalty factor, which depends on the history of the backtracking. This penalty factor is used to penalize larger delay variations.

The coarse alignment is the step requiring the highest computing power of the entire temporal alignment. Therefore, an efficient implementation is essential. In order to speed up the processing, calculations are often skipped for sections with reliable delays or for sections for which the delay can be extrapolated from the previous iteration.

4.4.1 Fractal Dimension

The fractal dimension of a signal can be seen as a measure of the signal's complexity. Very noisy signals will show a high fractal dimension *FD* per frame, while a sine tone will result in a very low *FD* value per frame.

In POLQA Sevcik's formula [14] is used to calculate the FD_f of each feature frame *f*:

$$\begin{aligned} Dist_i &= (Sample_i - Sample_{i+1})^2 \\ L_f &= \sum_{i=0}^{N} \sqrt{Dist_i + \left(\frac{1}{N-1}\right)^2} \\ FD_f &= 1 + \frac{\ln(L_f) + \ln(2)}{\ln(2 * (N-1))} \end{aligned} \tag{4}$$

where *N* is the number of samples in the feature frame *f* and $Sample_i$ is the value of the i'th sample of the signal (ranging from –32768 to 32767). The final feature vectors, which are based on the fractal dimension, are DC filtered in order to avoid problems with the subsequent computations [14].

4.4.2 Backtracking Algorithm

The backtracking algorithm, which is used to determine the optimal path through the correlation matrix, is very similar to Viterbi's algorithm. In POLQA, it is assumed that the correlation in each element $R_{m,f}$ of the correlation vector is similar to the probability of a delay-offset *f* of the macro frame *m*. All elements of the correlation matrix are first converted to a value, which can be interpreted as the logarithmic probability that macro frame *m* shows a delay offset of *f* feature frames. The result of this calculation is for each macro frame a probability vector p_m with *f* elements:

$$p_{m,f} = -\log_{10}(1 - R_{m,f})$$

The challenge is now to find the optimal path through that matrix that yields the highest overall probability, without having to try all possible combinations. To do so, the algorithm starts with the probability vector of the last macro frame *m* and searches the index of the element with the highest probability, $p_{m,f}$, giving a first path probability pp_m. Next, a penalty is added to all elements of the probability vector P_{m-1}. This penalty is weakest for delay offsets, which would result in the same absolute delay as that for macro frame *m* and it is strongest for large delay variations. This penalty reduces the likelihood of larger delay variations. Now, the element from P_{m-1} resulting in the highest combined probability pp_{m-1} is chosen:

$$pp_{m-1} = pp_m + p_{m-1,f} + Penalty(f)$$

For each step, the index *f* of the chosen optimum is stored. This index is equivalent to the offset of the best delay at the current feature resolution that has to be added to the last optimal delay value for each macro frame [15].

4.5 Fine Alignment

The fine alignment operates directly on both the reference and the degraded signal at the maximum possible resolution and determines the exact delay of each macro frame expressed in samples. The required search range is drastically limited due to the previous alignment steps. Therefore, it is possible to predict the accurate delay values using very short correlations without compromising the accuracy of the prediction.

The result of the fine alignment is the sample accurate delay value of each macro frame.

4.6 Joining Sections with Constant Delay

In this step all sections with identical delay are combined, which means one set of information (delay, reliability, start, stop, speech activity) is stored for the entire section.

In a second step, each section *n+1* is combined with section *n*

- If section *n+1* contains active speech **and** if the delay for both sections differs by less than 0.3 ms

or

- If section *n+1* consists of a speech pause **and** if the delay for both sections differs by less than 15 ms.

The resulting section information is passed on to the psychoacoustic model.

4.7. Sample Rate Ratio Detection

The sample rate ratio detection is required to compensate for perceptually irrelevant differences in the playout speed between the reference and the degraded signal. Such differences may have various reasons and may be intentional (e.g., time scaling due to jitter buffer adaptation) or unintentional (e.g., due to unsynchronized A/D or D/A converters in partly analog equipment). The resulting effect in any case is the same and can be described as a difference in the sample rate of two signals in the range of very few percent. It is important to note that this is not about the nominal sample rate but about the effective sample rate relative to another signal.

The detection of this effect as implemented in POLQA is based on the delay per macro frame vector and the detected active sections of the speech signals, as determined by the temporal alignment. The theory behind the algorithm is that sample rate differences will lead to delay changes, which are proportional to the ratio of the effective sample rates. The sophisticated part is to separate delay variations caused by sample rate differences from those caused by distortions like packet loss or jitter buffer adjustments. POLQA performs this by calculating a histogram of all delay variations that might be caused by a sample rate difference. Consequently, since sample rate differences cause many short delay variations rather than few large delay variations, only results of relatively small delay changes are used to build the histogram. The calculated histogram describes the distribution of delay variations per macro frame, which means that each detected delay variation is divided by the duration of the preceding section during which no delay change was detected. After filtering out unreliable peaks from the histogram, the position of the remaining peak value indicates the ratio of the sample rates. In order to calculate the exact value, the number of samples *NumAvg* stored in the histogram is counted; the weighted average *AvgBin* of all values is calculated, and the sample rate ratio *SRRatio* is derived from this value:

$$\begin{aligned} NumAvg &= \sum_{i=0}^{NumBins} Bin_i \\ AvgBin &= \frac{1}{NumAvg} \sum_{i=0}^{NumBins} Bin_i * i \\ SRRatio &= \frac{1}{1 - AvgBin + CenterBinIndex} \end{aligned} \tag{5}$$

With *NumBins* being the number of bins in the histogram and Bin_i representing the frequency of occurrence of delay step *i*. *CenterBinIndex* is the index of the bin, which corresponds to a delay of zero.

The resulting sample rate ratio is only valid if enough values were counted in the histogram. In all other cases, a ratio of 1.0 is reported.

4.8 Resampling

If the detected absolute sample rate difference is larger than 0.5%, the signal with the higher sample rate will be down sampled and the entire processing starts from the beginning. This happens at most once to avoid excessive looping in case of signals for which the sample rate ratio cannot be determined in a reliable manner.

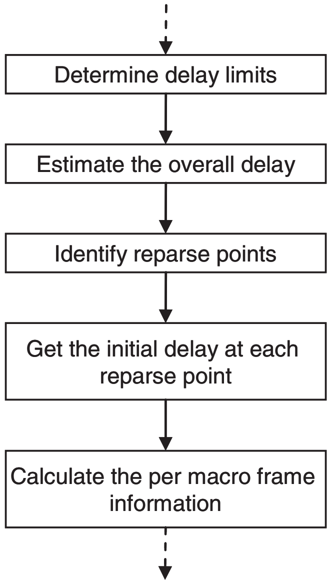
Determine delay limits
Estimate the overall delay
Identify reparse points
Get the initial delay at each reparse point
Calculate the per macro frame information

Fig. 8. Overview of the thorough pre-alignment.

Even if the sample rate determination cannot be made with perfect accuracy, e.g., in case of signals with additional variable delay, the detected sample rate ratio is still accurate enough to bring the signals back to the safe operating range of the temporal alignment.

5 VALIDATION OF THE TEMPORAL ALIGNMENT

In order to validate the temporal alignment, a measure for the “fitness” of two aligned signals had to be found. Simple methods, like, e.g., calculating the RMSE between the two aligned signals are not sufficient, since the algorithm had to be tested with file pairs where one signal is severely distorted. To assess the POLQA alignment, its own perceptual model was used as a measure for the similarity of the two input signals after the alignment. This perceptual model is fairly tolerant toward, e.g., speech coding, as long as the two compared signals sound similar, and it is thus an excellent measure for signal similarity. The next consideration is then how to make the assessment of the temporal alignment independent from the absolute accuracy of the perceptual model. By looking at the structure of the alignment algorithm it becomes obvious that the sample rate detection is not only a central element, but also the last step in the processing chain of the alignment. If the sample rate detection worked correctly, then most likely the temporal alignment itself also worked. The idea was now to measure each signal pair 60 times and to slightly vary the sample rate of the degraded signal for each run. If the entire alignment process worked well, the resulting MOS-LQO values should all be in the same range.

The result of this procedure can be seen in Fig. 9. The lines in this chart were derived by processing 21567 file pairs (POLQA Set 2, narrowband, wideband, and superwideband databases with a very large variety of distortions). Each file pair was processed 60 times while the degraded signal was resampled to +/–3% around its nominal sample

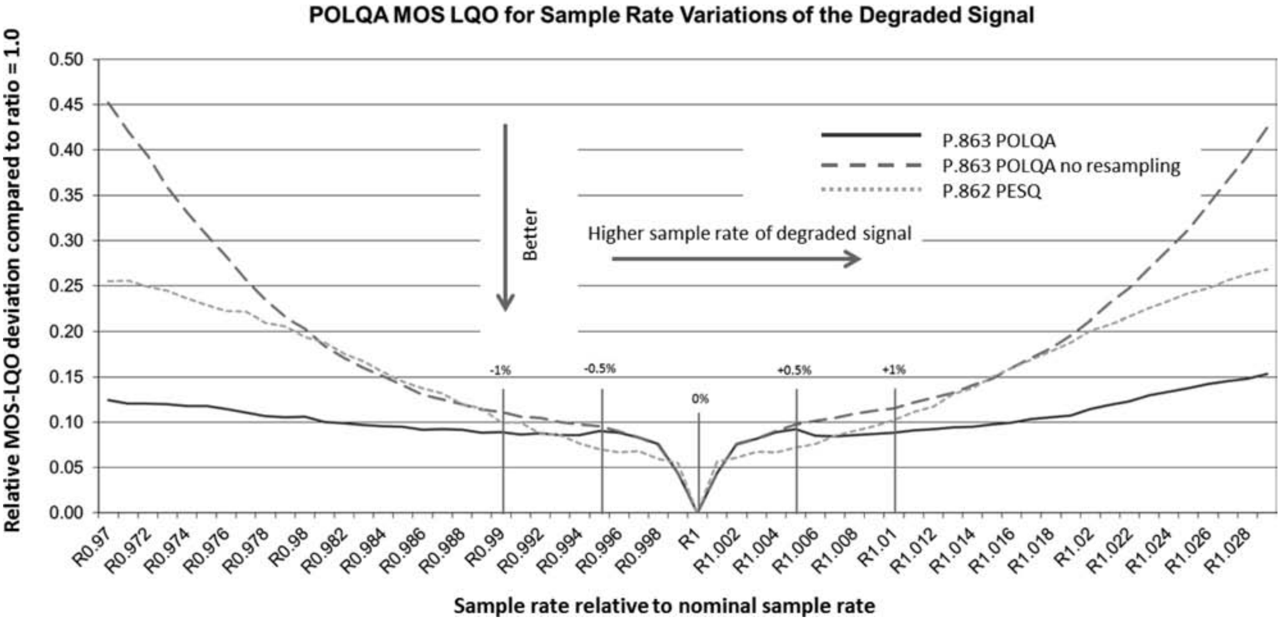

Fig. 9. Validation of the ability of the POLQA temporal alignment to handle time scaling effects.

rate in steps of 0.1%. For each measurement point the relative deviation of the resulting MOS-LQO compared from the MOS-LQO measured at the nominal sample rate was calculated. For each sample rate ratio this relative deviation was averaged and this average is shown as one line in Fig. 9. The different lines were derived from different alignment versions. The dark solid line applies the resampling to the degraded signal when the measured sample rate difference exceeds 0.5%. This is the version of the alignment that conforms to P.863. For comparison we also give the same results for the POLQA alignment without resampling (dashed line). Also shown in this chart are the same results for PESQ (dotted line), though one must clearly state that this test is far from anything for which PESQ has been designed. It only serves as a comparison to an alignment algorithm that expects piecewise constant delays.

Our conclusion from this comparison is that algorithms like PESQ can handle sample rate deviations of at most +/–1%. The temporal alignment of POLQA instead can handle sample rate deviations up to +/–3% using the resampling strategy. Tests outside of this range have not yet been performed, since 3% time scaling without pitch correction are already slightly perceivable. The numbers presented here hide a lot of information since they are the result of averaging 1,294,020 data points down to only 60 data points, but they are a very clear indication that the temporal alignment of POLQA works as expected. However, when looking at the chart, some open questions remain. It is unclear yet why a higher sample rate of the degraded signal has a much stronger effect on the MOS deviation than a lower sample rate. Our current assumption is that we are seeing some interaction between the perceptual model and the temporal alignment here. Especially the frame size used may be of influence on this. Also not yet sufficiently explained is the relatively strong variation as soon as any modification of the sample rate happens. We assume that this is caused by the sample rate conversion of the degraded signal as such and therefore due to the test setup.

6 CONCLUSIONS FOR PART I, POLQA TEMPORAL ALIGNMENT

The new temporal alignment that forms an integral part of the third generation objective speech quality assessment method POLQA, standardized by the ITU-T (International Telecommunication Union, Telecom sector) as Rec. P.863 is significantly more complex than the second generation, PESQ (Rec. P.862). However, it could be shown that the new algorithm allows for the alignment of a wide variety of complex distortions for which PESQ is known to fail, such as multiple delay variations within utterances as well as temporal stretching and compression of the degraded signal.

When this new alignment is used in combination with the new advanced perceptual model as described in Part II, it provides a new measurement standard that outperforms PESQ in assessing any kind of speech quality degradation making it the ideal tool for all speech quality measurements, from low end to HD voice communication in today's and future Voice-over-IP based and mobile networks.

7 REFERENCES

[1] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunication Union, Helsinki (1993); revised Geneva, Switzerland (1996).

[2] ITU-T Rec. P.830, "Subjective Performance Assessment of Telephone-Band and Wideband Digital Codecs," International Telecommunication Union, Geneva, Switzerland (1996 Feb.).

[3] ITU-T Rec. P.861, “Objective Quality Measurement of Telephone Band (300–3400 Hz) Speech Codecs,” International Telecommunication Union, Geneva, Switzerland(1996 Aug.).

[4] J. G. Beerends and J. A. Stemerdink, A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation,” *J. Audio Eng. Soc.*, vol. 42, pp. 115–123, (1994 Mar.).

[5] ITU-T Rec.P.862, “Perceptual Evaluation of Speech Quality (PESQ), An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs,” International Telecommunication Union, Geneva, Switzerland (2001 Feb.).

[6] ITU-T Rec. P.862.1, “Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO,” International Telecommunication Union, Geneva, Switzerland (2003 Nov.).

[7] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, “PESQ, the New ITU Standard for Objective Measurement of Perceived Speech Quality, Part I – Time alignment”, *J. Audio Eng. Soc.*, vol. 50, pp. 755–764 (2002 Oct).

[8] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, “PESQ, the new ITU standard for objective measurement of perceived speech quality, Part II–Perceptual Model,” *J. Audio Eng. Soc.*, vol. 50, pp. 765–778 (2002 Oct.).

[9] ITU-T Rec. P.862.2, “Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs,” International Telecommunication Union, Geneva, Switzerland (2005 Nov.).

[10] B. C. Bispo, P. A. A. Esquef, L. W. P. Biscainho et al., “EW-PESQ: A Quality Assessment Method for Speech Signals Sampled at 48 kHz,” *J. Audio Eng. Soc.,* vol. 58, pp. 251–268 (2010 Apr.).

[11] ITU-T Rec. P.863, “Perceptual Objective Listening Quality Assessment,” Geneva, Switzerland (2011 Jan.).

[12] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, “PEAQ–The ITU-Standard for Objective Measurement of Perceived Audio Quality,” *J. Audio Eng. Soc.*, vol. 48, pp. 3–29 (2000 Jan./Feb.).

[13] ITU-R Rec. BS.1387-1, “Method for Objective Measurements of Perceived Audio Quality,” International Telecommunication Union, Geneva, Switzerland (1998–2001).

[14] C. Goh, B. Hamadicharef, G. T. Henderson, and E. C. Ifeachor “Comparison of Fractal Dimension Algorithms for the Computation of EEG Biomarkers for Dementia” (University of Plymouth, UK), ISBN: 0-86341-520-2 © 2005 IEE, CIMED2005 Proceedings.

[15] M. Barkowsky, J. Bialkowski, R. Bitto, and A. Kaup, “Temporal Registration Using 3D Phase Correlation and a Maximum Likelihood Approach in the Perceptual Evaluation of Video Quality”, *MMSP Conference* (2007).

THE AUTHORS

John G. Beerends

Christian Schmidmer

Jens Berger

Matthias Obermann

Raphael Ullmann

Joachim Pomy

Michael Keyhl

John Beerends received a degree in electrical engineering from the HTS (Polytechnic Institute) of The Hague, The Netherlands, in 1975. After working in industry for three years he studied physics and mathematics at the University of Leiden where he received an M.Sc. degree in 1984. In 1983 he was awarded a prize of DFl 45000 by Job Creation for an innovative idea in the field of electro acoustics.

From 1984 to 1989 he worked at the Institute for Perception Research where he received a Ph.D. from the Technical University of Eindhoven in 1989. The main part of his doctoral work, which deals with pitch perception, was published in the Journal of the Acoustical Society of America. The results of this work led to a patent on a pitch meter by the N.V. Philips Gloeilampenfabriek.

From 1986 to 1988 he worked on a psycho-acoustically optimized loudspeaker system for the Dutch loudspeaker manufacturer BNS. The system was introduced at the Dutch consumer exhibition FIRATO in 1988.

In 1989 he joined the KPN Research Laboratory in Leidschendam where he worked on audio and video quality assessment, audio-visual interaction, and on audio coding (speech and music). This work led to several patents and two measurement methods for objective, perceptual, assessment of audio quality which he developed together with Jan Stemerdink. The first one deals with the quality of telephone-band speech codecs and was standardized within ITU-T in 1996 as Recommendation P.861 (PSQM, Perceptual Speech Quality Measure). The second method deals with the quality of music codecs and was integrated into ITU-R Recommendation BS.1387 (1998, PEAQ, Perceptual Evaluation of Audio Quality). Most of the work on audio quality (speech, music and audiovisual interaction) was published within the Audio Engineering Society and the ITU.

From 1996 to 2002 he worked with Andries Hekstra on the objective measurement of the quality of video and speech. The work on video quality led to several patents and a measurement method for objective, perceptual, assessment of video quality, standardized in ITU-T Recommendation J.247 (PEVQ, Perceptual Evaluation of Video Quality). The work on speech quality, partly carried out together with researchers from British Telecom, was focussed on improvements of the PSQM method and was standardized as ITU-T Recommendation P.862 (PESQ, Perceptual Evaluation of Speech Quality).

In January 2003 he joined TNO, which took over the research activities from KPN, where he worked on the objective measurement of speech intelligibility, (super) wideband speech quality, degradation decomposition, hearing aid quality, videophone quality and data chirping techniques. Together with researchers from OPTICOM and SwissQual he developed P.863 (POLQA, Perceptual Objective Listening Quality Assessment), the follow up of P.862.

John Beerends is the (co-) author of more than 90 papers/ITU contributions and 30 patents. In 2003 he received an AES fellowship award for his work on audio and video quality measurement.

●

Christian Schmidmer studied electronic engineering at the University of Erlangen. After achieving his M.S. degree (Diplom) he spent five years as a scientist at the audio department of the famous Fraunhofer Institute for Integrated Circuits in Erlangen (the home of mp3), mostly dedicated to the research of psychoacoustics and the development of perceptual measurement tools as well as audio codecs, contributing to the development of mp3. In 1997 he joined OPTICOM as CTO and co-owner. OPTICOM's core business is the development and IPR management for voice, audio and video quality measurement algorithms. Christian Schmidmer is active in standardisation bodies like ITU, VQEG and ETSI. He is the author of many scientific publications and frequently presenting papers at conferences and workshops. He is one of the main developers behind the recommendations ITU-R BS.1387 / PEAQ (Perceptual Evaluation of Audio Quality), ITU-T P.563 / 3SQM (no-reference voice quality assessment) and ITU-T P.863 / POLQA (full reference voice quality assessment).

●

Jens Berger completed his Master studies in Communication Engineering in Dresden, Germany, in 1989 with a thesis about measurement probes on microprocessor systems. He started in the area of studio acoustics and specialized further in speech processing and quality prediction models at the Research Institute of Deutsche Telekom.

Jens received a Ph.D. degree in Electrical Engineering from the Technical University of Kiel, Kiel, Germany, in 1998. His dissertation focused on objective measurements of speech quality, discussing the modeling of complex transmission systems and auditory tests by means of digital signal processing methods. The results of this work led to patents on a quality prediction model and loudness estimations in speech signals.

Since 2003, he has been with SwissQual AG, a Rohde & Schwarz company, Zuchwil, Switzerland, as Head of Applied Research for SwissQual's Quality of Service product lines and is responsible for the signal analysis and objective quality prediction methods of audio and video services in telecommunication networks.

This work led to several ITU-T standards; together with researchers from OPTICOM and Psytechnics he developed P.563 for voice quality prediction in 2003. In 2011, together with researchers from OPTICOM, TNO and SwissQual, Jens developed P.863 for voice quality prediction. Also in 2011, ITU-T standardized J.341, which was submitted by SwissQual, for objective HDTV quality prediction.

For the past ten years, Jens has been Rapporteur within ITU-T SG12: "Perceptual-based objective methods for voice, audio and visual quality measurements in telecommunication services." He is further active in ETSI, VQEG and other international standardization bodies.

●

Matthias Obermann studied electrical engineering at the University of Erlangen-Nuernberg, Germany, where he received the Diplom degree (M.S.) in 2009. His thesis dealt with ROI detection in video signals for CCTV scenarios.

Since end of 2009 he is a research engineer at OPTICOM where he worked on integration of the ITU-T P.863 model. Lately he switched focus to perceptual video quality measurement algorithms based on pixel and transmission protocol information.

Matthias Obermann is active in ITU-T and VQEG groups.

mk@opticom.de

●

Raphael Ullmann received the M.S. degree in Microengineering from Ecole Polytechnique Fédérale de Lausanne, Switzerland, in 2006. His thesis dealt with psychoacoustic modeling for Advanced Audio Coding. From 2006 to 2012, he was an Applied Research Engineer at SwissQual AG in Zuchwil, Switzerland, where he developed new algorithms for degradation analysis and quality prediction of speech signals in telecommunication systems. He has been a key contributor to SwissQual's P.OLQA candidate algorithm and its subsequent inclusion in the ITU-T P.863 model.

Since 2012, he is pursuing a Ph.D. at Idiap Research Institute in Martigny, Switzerland. His research is focussed on the perception and information content of background signals in speech telecommunications.

●

Joachim received his Diploma in Communications Engineering from Darmstadt Technical University in 1984.

The same year he joined Telefonbau & Normalzeit in Frankfurt where he became responsible for PABX transmission planning, and approval, for mouth-to-ear voice quality and for perceptual evaluation of users' satisfaction.

Since 1988 he is actively involved in related standardization efforts in ITU-T, ETSI, MESAQIN, BITKOM, TIA, T1A1 and IEEE. In 2002 Joachim was awarded the ETSI internal Service Medal for his achievements with the Asia-Pacific-Telecommunity (APT).

In 2008 Joachim left his position in Frankfurt and started to work as a Freelance Consultant in Telecommunications & International Standards. He completed successfully ITU projects on QoS in Nepal, the Maldives and for Afghanistan; he gave QOS/QoE workshop presentations e.g. in Spain, Kenya, the Maldives, Lebanon and South-Korea.

Currently, Joachim is leader of the an ETSI project on Adaptation of the ETSI QoS Model to better consider results from field testing. Joachim has taken responsibility and leadership in ETSI STQ and ITU-T SG12 for more than one decade.

Together with researchers from OPTICOM, Swissqual and TNO he supported actively the standardization process of P.863 (POLQA, Perceptual Objective Listening Quality Assessment) by acting as the appointed editor of P.863 until its publication and now as a co-editor for the Application Guide to P.863.

Joachim is vice chair of ETSI STQ and (co-) rapporteur for multiple areas in ITU and ETSI; he is also the (co-) author of more than 150 contributions to ITU, ETSI and TIA.

●

Michael Keyhl studied electrical engineering at the University of Erlangen-Nuernberg, where he received the degree Dipl-Ing. (M.S.) in 1989. He joined the Audio group of the Fraunhofer-Institute in Erlangen, better known as the, Home of MP3'. Michael was part of the MP3 development team and a project manager of the very first Audio-on-Demand music delivery system in the early nineties. In the course of his work with Fraunhofer, his passion for high-quality sound drove his engagement in further advancing perceptual measurement systems, including the Noise-to-Mask ratio real-time tester, the world's premiere perceptual quality analysis tool for compressed audio.

This work laid the ground for founding OPTICOM in 1995, the first spin-off enterprise commercially developing perceptual audio analysis tools. OPTICOM soon teamed up with John Beerends' work on PSQM/ITU-T Recommendation P.861. Having been the sole point of PSQM licensing in the world since the definition of the standard in 1996, OPTICOM has been instrumental in the commercial introduction of PSQM and the technology of standardized perceptual measurement as such.

Together with the distinguished team he gathered at OPTICOM, including his partner and CTO Chris Schmidmer, Michael co-authored, steered and managed development and standardization, along with the successful deployment of up to now six International Standards for Voice, Audio and Video testing, including PEAQ, PESQ, P.563, PEVQ and now POLQA – the third generation of perceptual voice quality testing for telecommunications.

Michael has been continuously active in, or observing the work of the AES, EBU, ITU-R, ITU-T, ETSI, ISO/MPEG and others. He has been recognized by many talks, paper presentations, workshop panels and contributions to books and standards' publications, and has been granted a number of International patents.

Having filled the position of the CEO of OPTICOM until today, he returned to the University of Erlangen-Nuernberg and received an executive MBA in Business Management in 2006. He is a member of the University's MBA advisory board and giving lectures in technology licensing.

He enjoys traveling, half-marathon running, and guitar playing and is always in favor of listening to good music, live or through, never-perfect' HiFi Surround.

He is a member of the Audio Engineering Society.

(*) AES Member

(362 words)