LOW COMPLEXITY SINGLE MICROPHONE TONAL NOISE REDUCTION IN VEHICULAR TRAFFIC ENVIRONMENTS

Navin Chatlani *Christophe Beaugeant* *Peter Kroon*

Intel, Allentown, PA USA Intel, Sophia Antipolis, France Intel, Allentown, PA, USA

ABSTRACT

A low complexity single microphone Tonal Noise Reduction (TNR) technique is presented for speech enhancement. This method is particularly effective in noisy environments which contain tonal noise sources, such as vehicular horns and alarms. TNR was designed to have low complexity and low memory requirements for use with battery operated communication devices. TNR detects the probability of the presence of these tonal noises which contaminate the desired speech signals. These noises are then attenuated using the proposed system for noise suppression. This is particularly effective for noise sources with a harmonic spectral structure. The proposed TNR system is able to maintain a balance between the level of noise reduction and speech distortion. Listening tests were performed to confirm the results. TNR can be used together with a general noise reduction system as a post-processing stage by reducing the residual noise components.

Index Terms— Single microphone noise reduction, Speech enhancement, Mobile devices

1. INTRODUCTION

Single microphone speech enhancement systems in mobile communication devices are used to reduce the level of noise from noisy speech signals. A common problem in such speech enhancement systems is the reduction of traffic noise sources, such as vehicular horn sounds, which contaminate the desired speech signal. Vehicular horns are highly non-stationary and they have a tonal structure. The spectral characteristics of the horn signals depend on the horn source. Therefore, simple approaches such as comb filters to notch predefined frequencies are inadequate. In such highly non-stationary environments, the noise power must be tracked, even during speech activity. Single microphone noise estimation techniques which operate in the Short-Time Fourier Transform (STFT) domain are very popular, including noise estimation systems such as the Minimum Statistics (MS) [1]. These MS-based techniques estimate the noise spectrum based on the observation that the noisy signal power decays to values characteristic of the contaminating noise during speech pauses. The main challenge faced by these techniques is tracking the noise power during speech segments, especially for non-stationary noises. This would result in poor estimates during long speech segments with few pauses.

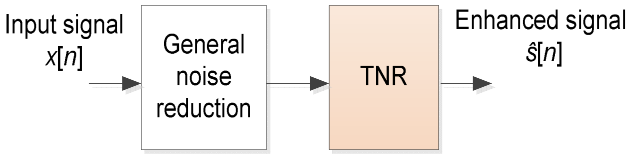


Fig. 1. Noise reduction solution incorporating TNR system.

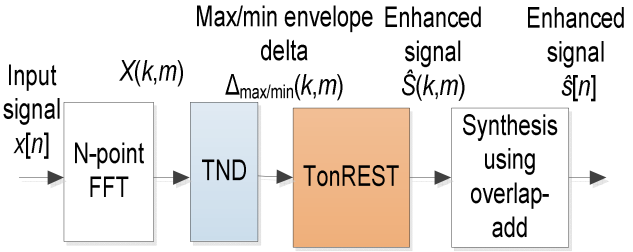


Fig. 2. TNR system diagram.

The noise reduction technique presented in this paper is targeted for portable, battery operated communication devices such as mobile phones, tablets, etc. Therefore, the system should have low complexity and low memory requirements. Previous approaches based on non-negative matrix factorization (NMF) [2-3] have been proposed to estimate the power of the tonal noise signals. These methods are computationally demanding and are not suitable for a system with low complexity requirements as targeted in this work. Assuming a proper noise estimate can be obtained, it can be used to filter the measured signal to reduce the noise and enhance the desired speech. The proposed TNR system can be used together with a general noise reduction system by applying it as a separate step, as shown in Fig. 1. As a result, it will be shown that TNR can be optimized and tuned separately.

In this paper, a TNR system is described to enhance noisy speech in the presence of tonal noises. The proposed system has been designed to be particularly effective at reducing vehicular horn noises. The TNR system is broken up into two sub-systems. In the first sub-system, a

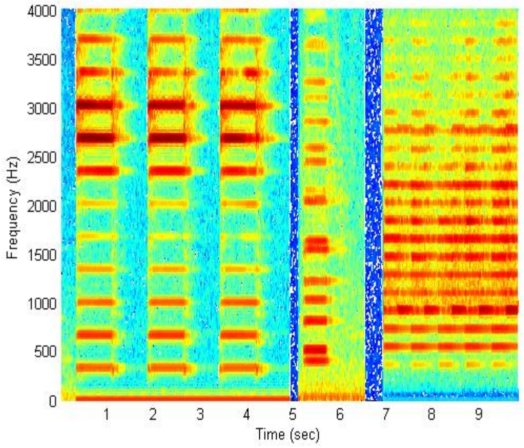


Fig. 3. Spectrogram illustrating the typical characteristics of horn and alarm sounds present in traffic noise.

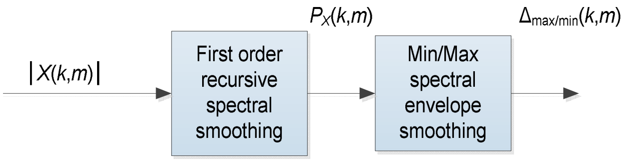

|X(k,m)|
First order recursive spectral smoothing
P_X(k,m)
Min/Max spectral envelope smoothing
Δ_max/min(k,m)

Fig. 4. Tonal Noise Detection (TND) system diagram.

technique is presented to perform Tonal Noise Detection (TND) to detect the interfering noise components. The second sub-system performs Tonal Reduction by Estimation (TonREST) of the desired speech and the detected noise components.

The paper is organized as follows. In Section 2, the TNR system is presented. Section 3 gives an evaluation of the system's performance and discussion of the results. Finally, conclusions are made in Section 4.

2. TONAL NOISE REDUCTION

The proposed TNR system is presented in Fig. 2. Consider the model described by:

$$x[n] = s[n] + d[n] \qquad (1)$$

where $x[n]$ is the noisy speech signal, $s[n]$ is the original noise-free speech signal, and $d[n]$ is the noise source which is assumed to be independent of the speech. The STFT of (1) may be written as:

$$X(k, m) = S(k, m) + D(k, m) \qquad (2)$$

for frequency bin k and time frame m.

In Fig. 2, it can be seen that the TNR system first performs TND to extract underlying signal characteristics which are used to detect the presence of tonal noise. The maximum/minimum envelope difference $\Delta_{max/min}(k, m)$ is used in the TonREST technique to attenuate the detected tonal noise components. The output enhanced signal $\hat{s}[n]$ is then reconstructed using circular convolution. The TND and the TonREST stages of the TNR system from Fig. 2 are described below. Voiced speech components have a harmonic structure which may be misclassified as tonal noise components. To reduce the impact of misclassification, the TNR system processes frequencies above the cut-off frequency f_c for detection and processing, in order to minimize distortion to the desired speech signal.

2.1 Tonal Noise Detection (TND) System Analysis

The TNR system has been designed to attenuate noise components, while minimizing distortion of the desired speech signal. The aim of the TND system is to extract characteristics of the noise components which can then be used for performing detection and classification of the desired speech and undesired noise components. The TND system was designed to be particularly effective at detecting tonal noise components such as vehicular horn sounds and alarms. Horn and alarm sounds have harmonic structures, with low fundamental frequencies which vary depending on the source of origin (e.g. type of car). Examples of the spectral and temporal characteristics of these noises are shown in Fig. 3.

It was observed that the power levels of these tonal sounds are either stationary for short time segments (signal dependent) or the power level decays with time. This characteristic is not the same for speech signals, as the power level fluctuates at a faster rate (4-6 syllables per second [4]) than the vehicular tonal noises. This observation for speech signals was confirmed in [5]. The TND system is based on these observations and is illustrated in more detail in Fig. 4. The minimum and maximum power envelopes of the noisy signal are tracked and the magnitude of their difference is used to classify either the desired speech or the target noise sources. The noisy speech spectral components $|X(k, m)|$ are first smoothed to form the smoothed noisy signal spectrum $P(k, m)$ by first order recursive averaging:

$$\begin{aligned} P(k, m) = (1 - \alpha)P(k, m - 1) \\ + \alpha|X(k, m)| \end{aligned} \qquad (3)$$

where α is a smoothing constant.

The minimum and maximum envelopes of $P(k, m)$ are tracked to determine the corresponding envelope nals $P_{max}(k, m)$ and $P_{min}(k, m)$. The maximum spectral envelope $P_{max}(k, m)$ decays when the signal energy level $P(k, m)$ remains constant or decreases in level as in (4). The spectral envelope $P_{max}(k, m)$ is tracked and updated when the signal energy $P(k, m)$ increases as in (5). The computation of $P_{max}(k, m)$ can be summarized as follows:

If $P(k, m) \leq P_{max}(k, m - 1)$

$$\begin{aligned} P_{max}(k, m) = (1 - \beta)P_{max}(k, m - 1) \\ + \beta|P(k, m)| \end{aligned} \qquad (4)$$

Else

$$P_{max}(k, m) = P(k, m) \qquad (5)$$

The minimum spectral envelope $P_{min}(k,m)$ increases when the signal energy level $P(k,m)$ remains constant or increases in level as in (6). The spectral envelope $P_{min}(k,m)$ is tracked and updated when the signal energy $P(k,m)$ increases as in (7). The computation of $P_{min}(k,m)$ can be summarized as follows:

If $P(k,m) \geq P_{min}(k,m-1)$

$$P_{min}(k,m) = (1-\beta)P_{min}(k,m-1) + \beta|P(k,m)| \quad (6)$$

Else

$$P_{min}(k,m) = P(k,m) \quad (7)$$

The final stage of the TND involves the computation of the difference between $P_{max}(k,m)$ and $P_{min}(k,m)$. This difference is denoted as $\Delta_{max/min}(k,m)$ and is determined as follows:

$$\Delta_{max/min}(k,m) = 10log_{10}P_{max}(k,m) - 10log_{10}P_{min}(k,m) \quad (8)$$

During tonal noise occurrences such as vehicular horn sounds, the second order statistics of these noises either remain relatively stationary or tend to decrease. From the above analysis of the TND technique, it can be seen that during noise instances which exhibit such behavior, the two spectral envelopes of $P_{max}(k,m)$ and $P_{min}(k,m)$ converge resulting in a decrease in $\Delta_{max/min}(k,m)$. Therefore, $\Delta_{max/min}(k,m)$ is used in TonREST to classify the signal components as desired speech or noise, before performing attenuation.

The time constants are set to determine the smoothing factors used in the recursive averaging in the top branch of the TND system from Fig. 4. These are set to low values to slow down the convergence of $P_{max}(k,m)$ and $P_{min}(k,m)$ and typical values are α=0.25 and β=0.01 in (3), (4) and (6). This results in low misdetections of speech as noise components.

2.2 Tonal Reduction by Estimation (TonREST) System Analysis

The TonREST system is designed to classify the input signal components of $X(k,m)$ as either speech or noise and perform noise reduction. The targeted traffic noise components have a tonal spectral structure and occupy the entire signal spectrum. Therefore, the first stage of TonREST in Fig. 5 involves the analysis of $X(k,m)$ to detect the spectral peaks $|X(i,m)|$ where i is denoted as the peak index. The corresponding spectral troughs $|X(j,m)|$ which surround the spectral peaks $|X(i,m)|$ are then detected. The trough index is denoted as j in the signal spectrum. The hypothesis H_1 is used to denote the presence of tonal noise. The difference of the maximum and minimum envelopes at the identified spectral peaks is denoted as $\Delta_{max/min}(i,m)$. This is used to estimate the tonal noise probability $p(i,m) = p\left(H_1\middle|\Delta_{max/min}(i,m)\right)$ corresponding to the detected spectral peaks.

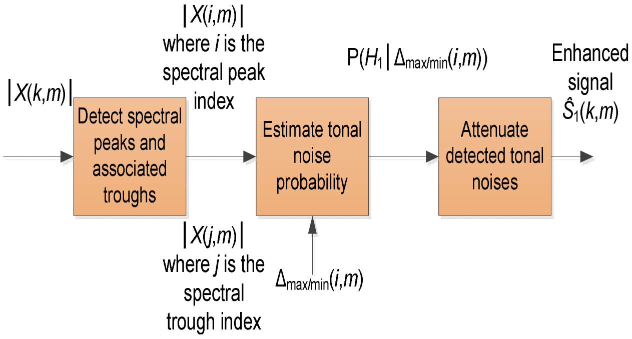

|X(k,m)|
Detect spectral peaks and associated troughs
|X(i,m)| where i is the spectral peak index
|X(j,m)| where j is the spectral trough index
Estimate tonal noise probability
$\Delta_{max/min}(i,m)$
$P(H_1|\Delta_{max/min}(i,m))$
Attenuate detected tonal noises
Enhanced signal $\hat{S}_1(k,m)$

Fig. 5. Tonal Reduction by Estimation (TonREST) system diagram.

The computed $\Delta_{max/min}(i,m)$ yields $p(i,m)$ and is defined below:

$$\begin{aligned} p(i,m) &= 0 \; for \; \Delta_{max/min}(i,m) > \tau_2 \\ &= \frac{\tau_2 - \Delta_{max/min}(i,m)}{\tau_2 - \tau_1} \\ &\quad for \; \tau_1 < \Delta_{max/min}(i,m) < \tau_2 \\ &= 1 \; for \; \Delta_{max/min}(i,m) < \tau_1 \end{aligned} \quad (9)$$

where the two thresholds τ_2 and τ_1 are set to control the boundaries for the signal classification as speech or noise. In this implementation, a good compromise between noise reduction and speech preservation can be achieved with values of τ_2=16 dB and τ_1=14 dB.

The final stage of TonREST involves the reduction of the detected tonal noises. For every spectral peak $|X(i,m)|$, a speech estimate $\lambda_S(i,m)$ is obtained from the surrounding spectral troughs $|X(j,m)|$which are less affected by the tonal noise components. $\lambda_S(i,m)$ is estimated as:

$$\lambda_S(i,m) = \left\{\frac{|X(j,m)| + |X(j+1,m)|}{K}\right\} \quad (10)$$

where K is set to control the amount of attenuation applied to the noisy signal. Therefore, larger values of K result in more noise attenuation. A typical value of K = 2 can be used. A noise estimate $\lambda_D(]j,j+1[,m)$ is hence derived as:

$$\lambda_D(]j,j+1[,m) = |X(]j,j+1[,m)| - \lambda_S(i,m) \quad (11)$$

where $]j,j+1[$ denotes the range of spectral troughs surrounding the examined peak i, excluding the endpoints. The magnitude of the enhanced speech $\lambda_S(]j,j+1[,m)$ is then recomputed by incorporating the estimated $p(i,m)$ to perform a weighted spectral subtraction to obtain the speech estimate as follows:

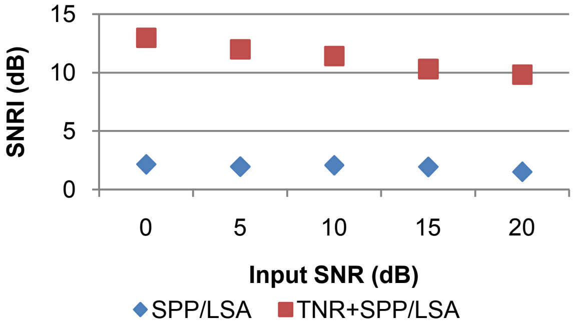


Fig. 6. SNRI of the TNR+SPP/LSA and SPP/LSA systems for various values of input SNR. The input signal is speech with traffic noise and SNRI indicates the signal to noise ratio improvement.

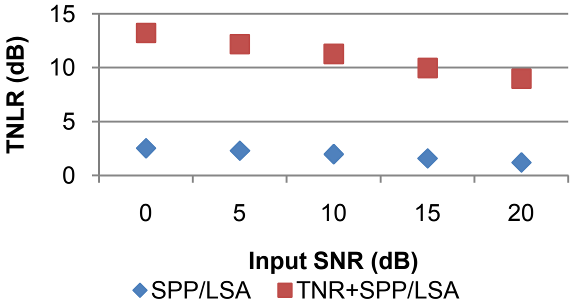


Fig. 7. TNLR of the TNR+SPP/LSA and SPP/LSA systems for various values of input SNR. The input signal is speech with traffic noise and TNLR indicates the total level of noise reduction performed during speech and non-speech segments.

$$\lambda_S(]j,j+1[,m) = |X(]j,j+1[,m)| - p(i,m)\lambda_D(]j,j+1[,m) \quad (12)$$

The weighting of the noise estimate λ_D in (12) with $p(i,m)$ allows the attenuation of the detected tonal noise components. The speech estimate from (12) is assigned to the following speech estimate:

$$|S(]j,j+1[,m)| = \lambda_S(]j,j+1[,m) \quad (13)$$

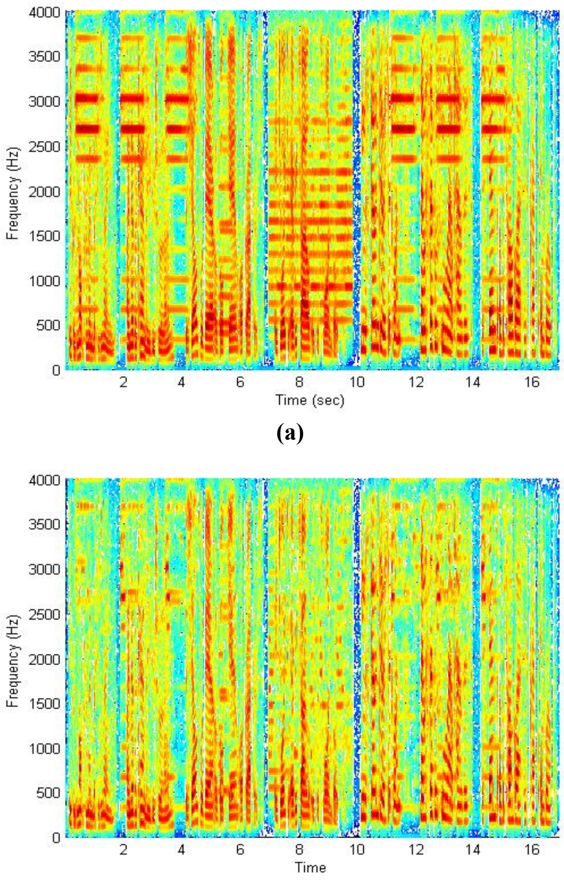
The enhanced speech signal is then estimated as follows:

$$\hat{S}(k,m) = |S(k,m)| exp\left(\sqrt{-1} * \theta_x(k,m)\right) \quad (14)$$

where $\theta_x(k,m)$ is the phase of the noisy speech.

3. RESULTS AND DISCUSSION

The performance of TNR was tested on databases containing multiple speech utterances and different noises including car, cafeteria, pub, road, train and office. In the case of non-traffic noises, the TNR system did not modify the signals as these signals did not have predominant interferences from tonal noises. In the following evaluation, the clean speech signals were corrupted with real recordings of traffic noise which was dominated by vehicular horn sounds and alarms. The SNR of the noisy signals range from 20 dB to 0 dB. A sampling frequency of 8 kHz was used for narrow-band processing. The signals were split up into frames of length 20 ms using a Hanning window with 50% overlap between consecutive frames and 256 point FFT. For comparison, the noisy signals were processed using the Speech Presence Probability (SPP) based noise estimation from [7]. The noisy speech was then enhanced using the Log Spectral Amplitude (LSA) estimation approach [8] and this competing approach is denoted as SPP/LSA. The LSA approach was used as it offers a steep attenuation characteristic. The noisy speech signals were also processed using the processing scheme illustrated in Fig. 1, where the aforementioned SPP/LSA system was used for general noise reduction. The TNR+SPP/LSA system was then used to process the signals above cut-off frequency f_c = 400 Hz. This cut-off frequency was chosen as a compromise between noise reduction and possible speech distortion.



(a)

(b)

Fig. 8. Demonstration of the TNR system at suppressing tonal components in traffic noise (a) Noisy speech signal at 5 dB SNR (b) Speech enhanced by TNR+SPP/LSA system.

In order to assess the relative performance of the noise reduction techniques, the objective measures of SNR Improvement (SNRI) and Total Noise Level Reduction (TNLR) [9] are used as the indicators for noise suppression effectiveness. SNRI measures the SNR improvement during speech activity. TNLR assesses the amount of noise reduction during both speech activity and speech pauses. In both cases of SNRI and TNLR, larger values indicate improved performance. The SNRI comparison of TNR+SPP/LSA and SPP/LSA is given in Fig. 6. The TNLR comparison of TNR+SPP/LSA and SPP/LSA is given in Fig. 7. The SPP/LSA system is effective at reducing general noise sources present in noisy speech signals. However, traffic noises are dominated by vehicular horns and alarms. These results demonstrate the effectiveness of the TNR system at performing significant reduction of the tonal noises present in traffic noise environments.

The spectrogram of a noisy speech recording made in the presence of traffic noise at 5 dB SNR is shown in Fig. 8(a). This signal was processed with the TNR+SPP/LSA system and the spectrogram of the enhanced signal is illustrated in Fig. 8(b). These results demonstrate the effectiveness of TNR at attenuating the tonal components present in traffic noise, while preserving the underlying speech content to minimize speech distortion.

It is important that the noise reduction system does not significantly distort clean speech signals. To assess the relative performance of the TNR system for clean speech, the objective measures of segmental SNR (segSNR) and Perceptual Evaluation of Speech Quality (PESQ) [10] are used. These measures indicate the amount of speech distortion introduced to clean speech signals processed by the TNR system. The above simulation set-up is used with the TNR system of Fig. 2. The results give the following values of: segSNR = 35.7 dB, PESQ = 4.2. These values illustrate TNR’s ability to preserve speech quality of clean speech signals.

TNR has low computational requirements of approximately 5 MCPS (8 kHz sampling frequency) when implemented on a fixed point DSP, and does not introduce additional latency when combined with a suitable noise reduction system. The performance of this technique was also verified for signals sampled at 16 kHz for wide-band processing and similar improvements were obtained.

4. CONCLUSIONS

In this paper, a novel technique is presented for the reduction of tonal noise interferers and improving the quality of speech conversations. Noise reduction is performed on spectral components only associated with the tonal noise and it typically does not impact any other type of encountered noises or speech. As a result, it does not introduce speech distortion that is commonly introduced by noise reduction techniques. This was confirmed in subjective listening tests which showed that the reduction of annoying tonal noises resulted in more comfortable listening experiences. The TNR system can be used together with classical noise reduction systems by applying it as a post-processing stage, and as such can also be optimized and tuned separately. Since it operates on a single channel, it can be used at the sending/uplink or the receiving/downlink side. The TNR system has low complexity because of its modular implementation. It has both low computational requirements and low memory requirements. Finally, many other acoustic enhancement techniques also operate in the frequency domain. This allows for computationally efficient implementations by combining the frequency to time transforms of various processing modules in the audio sub-system.

REFERENCES

[1] Martin, R., "Noise PSD Estimation based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, Jul. 2001

[2] Benayora, L., et al, "Single Sensor Source Separation based on Wiener Filtering and Multiple Window STFT", *Proc. Of International Workshop on Acoustic and Echo Noise Control (IWAENC)*, 2006

[3] Jeong, K. M., et al, "Mechanical Noise Suppression Based on Non-Negative Matrix Factorization and Multi-Band Spectral Subtraction for Digital Cameras", *IEEE Transactions on Consumer Electronics*, vol. 59, no. 2, pp. 296-302, May 2013

[4] Buchler, M., PhD thesis titled "Algorithms for Sound Classification in Hearing Instruments"

[5] Marzinzik, M., Kollmeier, B., "Speech Pause Detection for Noise Spectrum Estimation by Tracking Power Envelope Dynamics", *IEEE Transactions on Speech and Audio Processing*, vol. 10, Feb. 2002

[6] Loizou, P., 2007 Noise Compensation by Human Listeners. In: PRESS, C. (ed.), Speech Enhancement: Theory and Practice. FL, Ch. 4.2, pp. 78-79

[7] Gerkmann, T., Hendriks, R., "Noise power estimation based on the probability of speech presence", *Proc. Of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2011

[8] Ephraim, Y., Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", *IEEE Transactions on Acoustical Speech Signal Processing*, vol. ASSP-33, no. 2, pp. 443-445, Apr. 1985

[9] "Voice Enhancement Devices: Amendment 1:Revised Appendix II - Objective Measures for the Characterization of the basic functioning of noise reduction algorithms", *ITU-T Recommendation G.160 Amendment 1*, Nov. 2009

[10] "P.862.1:Mapping function for transforming P.862 raw result scores to MOS-LQO", *ITU-T Recommendation*, Oct. 2003