A SIMPLIFIED VERSION OF THE ITU ALGORITHM FOR OBJECTIVE MEASUREMENT OF SPEECH CODEC QUALITY

S. Voran

Institute for Telecommunication Sciences, National Telecommunications and Information Administration
325 Broadway
Boulder, Colorado 80303, USA, sv@bldrdoc.gov

ABSTRACT

ITU-T Recommendation P.861 describes an objective speech quality assessment algorithm for speech codecs [1]. This algorithm transforms codec input and output speech signals into a perceptual domain, compares them, and generates a noise disturbance value, which can be used to estimate perceived speech quality. The performance of this algorithm can be judged by the correlation between those estimates and actual listener opinions from formal subjective listening tests. We show that significant simplifications can be made to the P.861 algorithm with very minimal effect on its performance. Specifically, for the portions of the algorithm under study here, 64% of the floating point operations can be eliminated with only a 3.5% decrease in average correlation to listener opinions. The resulting simplified algorithm may offer a practical new objective function to drive parameter selections, excitation searches, and bit-allocations in speech and audio coders.

1. INTRODUCTION

ITU-T Recommendation P.861 describes an objective speech quality assessment algorithm for speech codecs [1]. For consistency reasons, this paper uses the same terminology and notation as the Recommendation. The algorithm (Figure 1) uses a listening environment model and models for human hearing to transform codec input and output speech signals into a perceptual domain. It then compares the two signals and generates a noise disturbance (ND) value, which is an estimate of perceptual distance between the codec input and output speech signals. This estimate can be used to predict the perceived quality of codec output speech. The algorithm operates on 32-ms speech frames, indexed by the variable "i." A critical band frequency domain representation is used through most of the algorithm. This representation involves 56 samples spaced at 0.312 critical bands from 50 Hz to 4 kHz and these samples are indexed by the variable "j."

The performance of the P.861 algorithm can be judged by the correlation between its estimates and actual listener opinions from formal subjective listening tests. This paper describes a sensitivity study that was undertaken to determine the relative importance of six different components of the P.861 algorithm. In this study we made simplifications and approximations in the shaded areas of Figure 1, both independently and jointly, resulting in 15 variants of the original algorithm. We measured the consequent changes in correlation values for seven different subjective tests.

2. COMPONENTS UNDER STUDY

2.1 IRS Filtering

The P.861 algorithm filters both speech signals according to the modified Intermediate Reference System (IRS) receiving characteristic [2]. This filtering action simulates the frequency response of a typical telephone handset earpiece. It has a bandpass characteristic, with -3 dB points near 400 Hz and 3200 Hz, and a fairly flat passband. To investigate the significance of this filtering, we replaced it with a rectangular bandpass filter that cuts off at 400 Hz and 3200 Hz. We also experimented with removing the filter entirely.

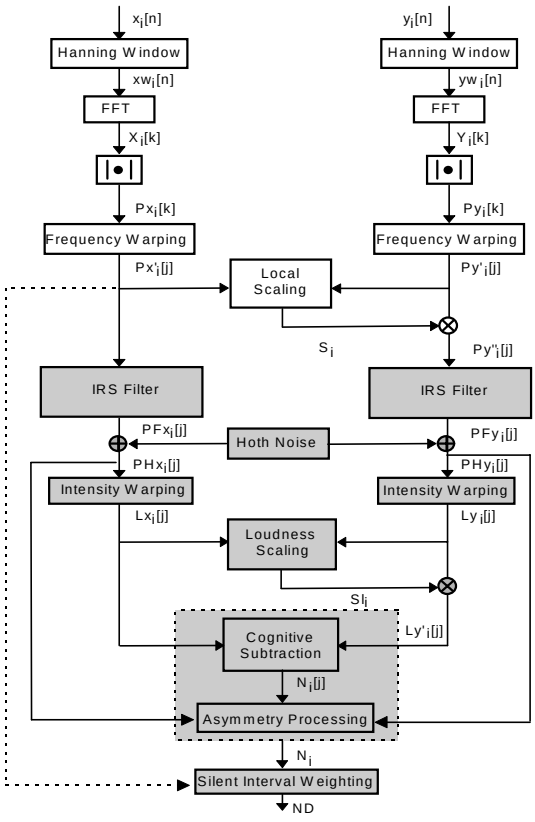

Figure 1. Block diagram of the P.861 algorithm.

2.2 Hoth Noise Injection

Hoth noise is injected into both speech signals to model a typical listening environment [3]. However, the injection level is set to 45 dBA, which is low enough to be of limited consequence. To investigate this possibility, we experimented with removing the Hoth noise.

2.3 Intensity Warping

The intensity warping function models the relationship between signal power and perceived loudness. The relationship given in P.861 is

$$Lx_i[j] = \max\left(0,\ \left(PHx_i[j] + P_0[j]\right)^{\gamma} - \left(2 \cdot P_0[j]\right)^{\gamma}\right), \qquad (1)$$

where $Lx_i[j]$ is loudness, $PHx_i[j]$ is power, $P_0[j]$ is the hearing threshold, and γ=0.001. For signals that are more than a few dB above the threshold of hearing, the relationship given in (1) is very nearly linear in logarithmic signal power. The intercept varies with frequency, but the slope does not. We investigated three simplifying approximations to (1). All are linear functions of signal power, measured in dB, and the slope of (1) is preserved. The first approximation nearly preserves the frequency-dependent intercepts in (1), the second approximation uses an averaged intercept, and the third approximation ignores the intercept issue completely:

$$\begin{aligned} Lx_i[j] &\approx a \cdot \max\left(0, \left[10 \cdot \log_{10}\left(\frac{PHx_i[j]}{P_0[j]}\right) - 3.2\right]\right), \\ Lx_i[j] &\approx a \cdot \max\left(0, \left[10 \cdot \log_{10}\left(PHx_i[j]\right) - 11\right]\right), \\ Lx_i[j] &\approx a \cdot \max\left(0, 10 \cdot \log_{10}\left(PHx_i[j]\right)\right), \text{ where } a = 0.056. \end{aligned} \qquad (2)$$

2.4 Loudness Scaling

The loudness scaling function forces the momentary compressed loudness of each frame of the two speech signals to match. We investigated the effect of turning this function off.

2.5 Distance Measure

A distance measure attempts to model how listeners compare two sounds, and is likely to be one of the more important parts of an objective audio quality assessment algorithm [4]. The P.861 algorithm uses a distance measure constructed from a cognitive subtraction stage followed by asymmetry processing. Together they are described by

$$\begin{aligned} N_i &= \sum_{j=1}^{56} N_i[j] \cdot C_i[j], \text{ where } N_i[j] = \max\left\{0, \left|Ly'_i[j] - Lx_i[j]\right| - .01\right\}, \\ \text{and } C_i[j] &= \min\left\{\left(\frac{PHy_i[j]+1}{PHx_i[j]+1}\right)^{0.2},\ 2\right\}. \end{aligned} \qquad (3)$$

We evaluated this equation for tones between 270 Hz and 4 kHz at a range of power levels $PHx_i[j]$ and $PHy_i[j]$ that corresponds to 30 - 100 dB SPL. Over these ranges, (3) is almost completely independent of frequency, and is well approximated by a function of the loudness difference, $Ly'_i[j] - Lx_i[j]$ alone. Our study of these loudness difference values over a wide range of speech codecs and channel conditions has revealed that their distribution is approximately Laplacian with a standard deviation of $\sigma \approx 0.15$ over most of the band. At the band edges, σ is even smaller. Thus 99% of the loudness difference magnitudes $|Ly'_i[j] - Lx_i[j]|$ are expected to be less than 0.70. Over this range, (3) can be approximated by a pair of linear equations:

$$\begin{aligned} N_i[j] &= 1.673 \cdot (Ly'_i[j] - Lx_i[j]), \text{ when } Ly'_i[j] \geq Lx_i[j], \\ &= 0.712 \cdot (Lx_i[j] - Ly'_i[j]), \text{ otherwise.} \end{aligned} \qquad (4)$$

2.6 Silent Interval Weighting

The P.861 algorithm performs a weighted average of ND values from multiple speech frames to generate a single ND value. In this weighted average, a weight of 1.0 is applied to all frames with an estimated level above 70 dB SPL, and a weight of 0.2 is applied to all frames below that level. Recommendation P.861 refers to this process as "silent interval weighting." As part of this study, we replaced the weight of 0.2 with a weight of 0.0. Under this option, only the loudest frames (those above 70 dB SPL) need to be processed and the rest can be ignored.

3. RESULTS

Section 2 describes the nine simplifying modifications we have made to the P.861 algorithm in this study. In this section we report the impact of those modifications on the ability of the algorithm to estimate perceived speech quality. We judged the P.861 ND values against mean opinion score (MOS) results from seven formal subjective tests. Together, these seven tests include 182, 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit-rates ranging from 2.4 - 64 kbps (Table 1). They include 19 hours of speech material from 3 languages.

The ND values generated by P.861 have a theoretical range from 0 to + ∞. In practice MOS values range from approximately 1.0 to 4.4. We used a logistic function to map ND values into that interval:

$$L(ND) = 0.99 + (5.01 - 0.99)/(1 + e^{a \cdot ND - b}), \qquad (5)$$

where a = 0.5431, and b = 1.6761.

We chose the coefficient of correlation between L(ND) and MOS as the figure of merit for the different versions of P.861 created in this study. For each device in a subjective test, we averaged L(ND) and MOS values over all available speech files and then calculated correlations using these averaged L(ND) and MOS values. We call these results "per-condition correlations." The constants a and b in (5) were chosen to maximize these correlations for the original P.861 algorithm across all seven tests.

3.1 Simplification of Single Components

Table 2 shows correlation results for the original P.861 algorithm and each of the nine modifications described in Section 2. We calculated percentage changes in these correlation values (referenced to the correlation of the original P.861 algorithm) and averaged them across algorithm modifications and across

subjective tests. The averages across tests allowed us to identify the loudness scaling factor as the most important component, but the average performance decrease associated with its removal is only 3.5%. Approximating the distance measure with (4) causes an average correlation drop of only 1.6%, and all other simplifications or approximations cause drops of 1% or less. Before averaging, the largest correlation drop was only 13%, and there was a correlation gain of 9% as well (see Table 2).

The averages across modifications make it clear that the modifications have bigger impacts on tests 1-4 than on tests 5-7. Tests 5-7 contain only higher rate coders that tend to preserve waveforms and error-free channel conditions. Thus they present easier estimation problems than tests 1-4. It is possible that tests 1-4 benefit more from the more precise modeling of the original P.861 algorithm. We note however, that even the largest averaged (across modifications) performance decreases are less than 2%.

3.2 Simplification of Multiple Components

Using the sensitivity results from Table 2, we created six additional variants of the P.861 algorithm by simplifying multiple components. These six versions are defined in Table 3. As one moves down Table 3, additional simplifications are made, complexity is reduced, and we would expect performance to decrease.

Results of this portion of the study are given in Table 4. The averaged percentage changes at the bottom of Table 4 confirm that performance does decrease with version number up to a point. However, together, the distance measure approximation and the removal of the loudness scaling factor seem to compensate for the other approximations, resulting in a highly simplified variant algorithm (version 6) with an average performance reduction of only 3.5%.

These results, along with counts of floating point operations, allowed us to plot a measured complexity-performance relationship for the group of variant P.861 algorithms. In Figure 2, the averages of Table 4 are plotted against the total number of floating point operations required by the shaded areas of Figure 1. The abscissa units in Figure 2 are kflops required to process one second of speech (2 channels, 8000 samples/s each channel). For these seven subjective tests, 64% of the floating point operations can be eliminated with only a 3.5% decrease in average correlation to listener opinions. Across all six versions considered, the averaged correlation decrease never exceeds 10.1%.

All six of these versions bring with them a separate, even more dramatic advantage. Since they all use a weighting of 0.0 for the ND values of frames below 70 dB SPL, those frames need not be processed at all. For the seven subjective tests used in this study, about 60% of all speech frames fall into this category. This means that only 40% of the speech needs to be processed. This *speech processing reduction* applies throughout the entire algorithm, and is in addition to the *floating point operation reductions* for the shaded areas in Figure 1.

4. SUMMARY

ITU-T Recommendation P.861 describes an objective speech quality assessment algorithm that is a useful tool in many situations. However, it appears that a portion of the algorithm complexity is not contributing much to the perceived speech quality estimates, at least for the seven subjective tests studied here. These tests include 182, 4-kHz bandwidth speech codecs, transmission systems, and reference conditions, with bit-rates ranging from 2.4 - 64 kbps.

While IRS filtering and Hoth noise injection do model listening conditions, removing either of these components results in an averaged correlation drop of 1% or less. The detailed intensity warping relationship in (1) does reproduce known tone and noise loudness perception results, but it is apparently not necessary to model perceived speech signal loudness in this application. That relationship can be replaced with the simpler relationships in (2) and averaged correlation drops are 0.6% or less. The loudness scaling factor seems more significant, as does the distance measure. However, the distance measure given in (3) can be greatly simplified as shown in (4), resulting in an averaged correlation drop of only 1.6%. Finally, the weighting placed on ND values for speech frames below 70 dB SPL can be reduced to 0.0, with almost no effect. This allows one to process only 40% of the frames in a speech signal and lose only 0.2% in correlation to MOS. In addition to this *speech processing reduction*, the simplifications and approximations described here allow one to eliminate 64% of the *floating point operations* in the shaded areas in Figure 1, at a cost of an average correlation drop of only 3.5%. Using these simplifications, the algorithm may now be a candidate for inclusion in speech and audio coders. It might provide feedback to parameter selection, excitation search, and bit-allocation algorithms to ensure that the highest possible signal quality is obtained at the lowest possible bit rate.

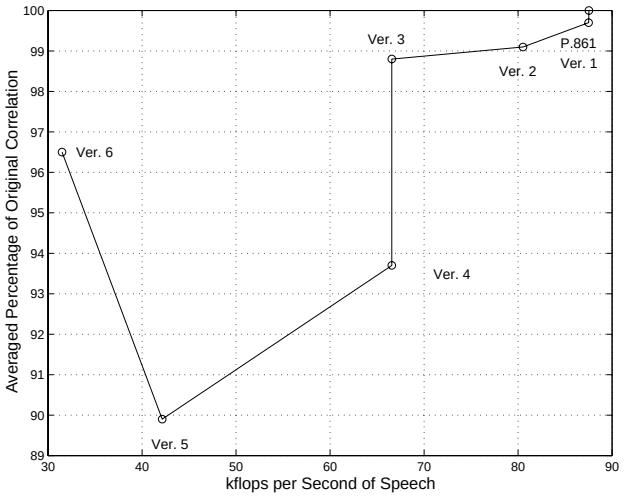

Figure 2. Complexity-performance trade-off.

5. REFERENCES

[1] ITU-T Rec. P.861, "Objective quality measurement of telephone-band speech codecs." Geneva, 1996.

[2] ITU-T Rec. P.830, "P.830 Subjective performance assessment of telephone-band and wideband digital codecs." Geneva, 1996.

[3] Hoth D.F. "Room noise spectra at subscribers' telephone locations." *Journal of the Acoustical Society of America*, 12:499-504, 1941.

[4] Voran S. "Estimation of perceived speech quality using measuring normalizing blocks." *Proceedings of the 1997 IEEE Speech Coding Workshop*, Pocono Manor PA, USA, 1997, pages 83-84.

Test*	1	2	3	4	5
Num. Conds.	22	35	27	38	20
Cond. List	PCM, ADPCM, APC, SELP, LPC, MNRU (Tandems)	PCM, CELP, AMPS , MNRU (Frame Erasures)	ADPCM, CVSD, VSELP, CELP, IMBE, STC, LPC, POTS, MNRU (Bit Errors)	ADPCM, CELP, VSELP, IMBE, AMBE, MNRU, (Mixed Tandems)	PCM, ADPCM, CELP, MNRU (Tandems)
Rates (kbps)	2.4-64	8-64	2.4-32	6.4-32	16-64
Talker/Cond.	4	6	6	8	4
Num. Files	176	1050	1994	2432	1440

*Tests 1-5 are in English; Tests 6 and 7 are identical to Test 5, but are in Japanese and Italian, respectively.

Table 1. Summary of conditions and speech material in seven subjective listening tests.

Test	P.861	IRS		No Hoth Noise	Intensity Warping Appxs.			No Loudness Scaling	Distance Measure Approx.	Silent Interval Weight = 0	Avg. % Change
		Rect.	None		1	2	3				
1	.929	.927	.893	.932	.925	.915	.913	.854	.947	.910	-1.7
2	.941	.923	.924	.936	.941	.941	.942	.903	.864	.935	-1.8
3	.795	.776	.792	.801	.794	.771	.769	.703	.869	.802	-1.1
4	.973	.969	.968	.956	.974	.977	.978	.969	.842	.981	-1.6
5	.985	.984	.985	.978	.985	.986	.987	.989	.978	.982	-0.1
6	.986	.985	.983	.985	.985	.985	.985	.976	.986	.983	-0.2
7	.976	.974	.976	.979	.976	.978	.979	.979	.978	.975	+0.1
Average % Change		-0.8	-1.0	-0.2	-0.1	-0.6	-0.6	-3.5	-1.6	-0.2	

Table 2. Per-condition correlation values and averaged percentage changes for nine different simplifications and approximations.

Version	IRS Filtering	Hoth Noise	Intensity Warping	Loudness Scaling	Distance Measure	Silent Interval Weight
1	Original	Original	Approximation 1	Original	Original	0.0
2	Original	Removed	Approximation 1	Original	Original	0.0
3	Rect. Appx.	Removed	Approximation 3	Original	Original	0.0
4	Removed	Removed	Approximation 3	Original	Original	0.0
5	Removed	Removed	Approximation 3	Original	Approximation	0.0
6	Removed	Removed	Approximation 3	Removed	Approximation	0.0

Table 3. Definitions for six simplified versions of the P.861 algorithm.

Test	P.861	Version 1	Version 2	Version 3	Version 4	Version 5	Version 6
1	.929	.908	.907	.914	.792	.677	.882
2	.941	.935	.923	.920	.851	.854	.906
3	.795	.802	.798	.779	.793	.883	.756
4	.973	.981	.974	.973	.940	.766	.944
5	.985	.981	.975	.974	.949	.917	.961
6	.986	.983	.977	.979	.919	.896	.955
7	.976	.974	.970	.971	.923	.895	.957
Average Percent Change		-0.3	-0.9	-1.2	-6.3	-10.1	-3.5

Table 4. Per-condition correlation values and averaged percentage changes for six simplified versions of P.861 algorithm.