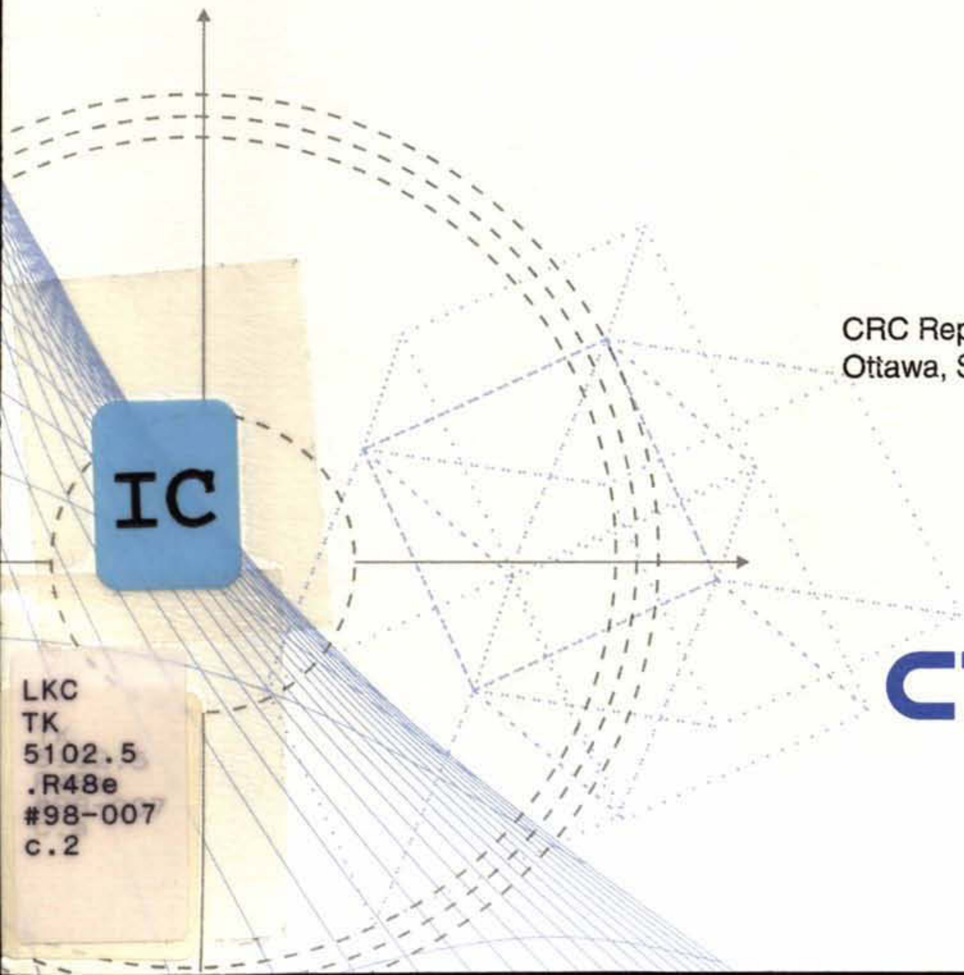
Objective Measurement of Perceived Audio Quality

William C. Treurniet
Advanced Sound Systems



CRC Report/TN No. CRC-TN-98-007
Ottawa, September 1998

CRC
Communications
Research Centre
Centre de recherches
sur les communications

Objective Measurement of Perceived Audio Quality

Acknowledgments

A number of people have contributed to the work leading to the objective audio quality measurement model described in this paper. Bruno Paillard, under contract to CRC, developed the original version of the peripheral ear model. Louis Thibault was instrumental in initiating and managing the contract, and performed an initial evaluation of the model. Seymour Shlien suggested changes to the values of several ear model parameters that improved performance, and Gilbert Soulodre contributed useful insights regarding signal processing techniques. Finally, Ted Grusec and Gilbert Soulodre at CRC, as well as investigators from labs in other countries, performed the experiments that provided the subjective listening test data used to calibrate the model.

Executive Summary

A measure of the subjective quality of audio data is usually obtained by asking human listeners to judge the degree of degradation in processed music or speech excerpts relative to the original audio signal. Since such ratings must be made under optimal listening conditions using careful experimental procedures, obtaining reliable subjective data is often expensive. Therefore, a method for objective assessment of audio quality is needed for situations where listening tests are impractical.

According to ITU-R Recommendation BS-1116, listeners in listening tests rate the degraded audio items as well as the corresponding reference items using a continuous quality scale ranging from one to five. The difference between the ratings for the reference and processed items is defined as the subjective difference grade (SDG) for the item, and the mean SDG over a number of listeners represents its subjective quality. Different types of audio distortions may occur sequentially, so variations in quality must be integrated over time. Therefore, prediction of the SDG requires an accurate model of both the peripheral auditory system as well as cognitive aspects of audio quality judgements. This note describes such a model for objective measurement of audio quality.

The model produces a number of variables based on differences between the reference and processed signals as well as the reference signal alone. These variables were mapped to an objective difference grade (ODG) using an optimization technique that minimizes the difference between the ODG distribution and the corresponding distribution of mean SDGs for the available data set.

The calibrated model was used to measure the quality of items in a new data set that was created to statistically compare several state-of-the-art codec systems. The correlation of 0.76 between SDGs and ODGs for the individual items was smaller than expected. Qualities of codecs were also determined by averaging the ODGs of items processed by each codec at each available bit rate. The correlation between the subjective and measured qualities of the codec systems was 0.91. The model successfully identified the best performing codecs in the test, and the pattern of results suggested directions for further work.

Table of Contents

List of Figures

List of Tables

1 Introduction

A measure of the subjective quality of audio data is usually obtained by asking human listeners to judge the degree of degradation in processed music or speech excerpts relative to the original audio signal. Since such ratings must be made under optimal listening conditions using careful experimental procedures, obtaining reliable subjective data is often expensive. Therefore, a method for objective assessment of audio quality is needed for situations where listening tests are impractical.

According to ITU-R Recommendation BS-1116, listeners in listening tests rate the degraded audio items as well as the corresponding reference items using a continuous quality scale ranging from one to five. The difference between the ratings for the reference and processed items is defined as the subjective difference grade (SDG) for the item, and the mean SDG over a number of listeners represents its subjective quality. Different types of audio distortions may occur sequentially, so variations in quality must be integrated over time. Therefore, prediction of the SDG requires an accurate model of both the peripheral auditory system as well as cognitive aspects of audio quality judgements.

This note describes *Perceval*, a computational model of audition developed to evaluate objectively the perceptual quality of audio signals [1]. The model simulates peripheral auditory processes and makes reasonable assumptions about higher level perceptual and cognitive processes in order to estimate the perceived quality of a test signal relative to a known reference signal.

A number of variables are produced based on differences between the reference and processed signals as well as the reference signal alone. These variables are mapped to an objective difference grade (ODG) using an optimization technique that minimizes the difference between the ODG distribution and the corresponding distribution of mean SDGs for the available data set.

A high level representation of the model is shown in Figure 1. In general, it compares time-aligned versions of a reference signal and a processed version of the same signal. The reference and possibly altered signals are each transformed to a basilar membrane (BM) representation, and the *basilar degradation* is defined as the difference between these representations. The basilar degradation, as well as other characteristics of the audio material, is analyzed further as a function of frequency and time by a *cognitive model.* The latter extracts perceptually relevant features that are used to compute a measure of quality.

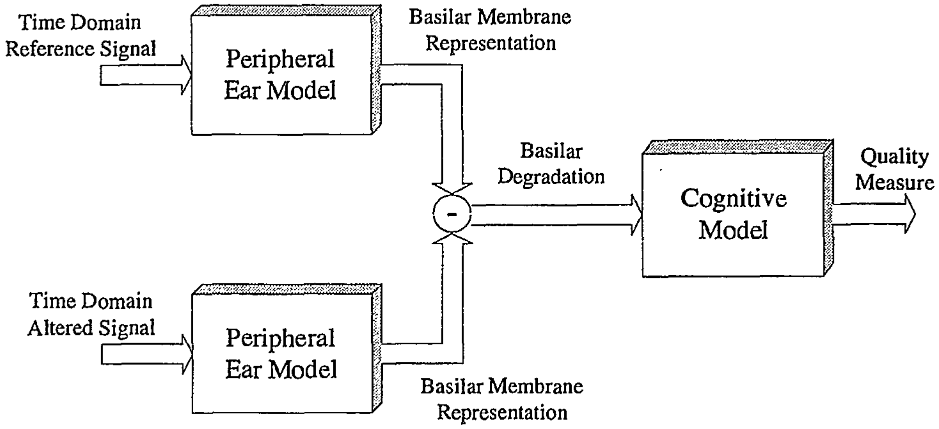


Figure 1. High-level representation of *Perceval*

2 Model Description

2.1 Peripheral Ear Model

Perceval's ear model is a frequency domain model designed primarily to deal with simultaneous masking effects. That is, it does not attempt to explicitly model forward or backward masking. It accounts for simultaneous masking by modeling the transfer functions of underlying physical processes. The result is a BM representation that is assumed to reflect only the audible signal.

The ear model assumes that the detectors on the basilar membrane are sensitive to the logarithm of the input energy and have low temporal resolution. Further, the mechanical phenomena of the inner ear are considered linear and invariant with respect to frequency and level of the input signal. This is an approximation since the response of the cochlear mechanism is somewhat sensitive to signal level.

The successive stages of the ear model are shown in Figure 2. A Hann window is applied to 2048 input data samples, and the Fast Fourier Transform (FFT) is applied to form a time-frequency representation. Successive windows of the input data overlap by 1024 samples. The energy spectrum is attenuated by a frequency dependent function that models the effect of the ear canal and the middle ear. The attenuated spectral energy values are mapped from the frequency scale to a pitch scale using an empirically derived frequency to pitch mapping function. The energy components are then convolved with a spreading function to simulate the dispersion of energy along the basilar membrane. Finally, an intrinsic frequency-dependent energy is added to each pitch component to simulate internal noise that is thought to account for the absolute threshold of hearing. Conversion of the energy to decibels results in the basilar membrane representation of the signal.

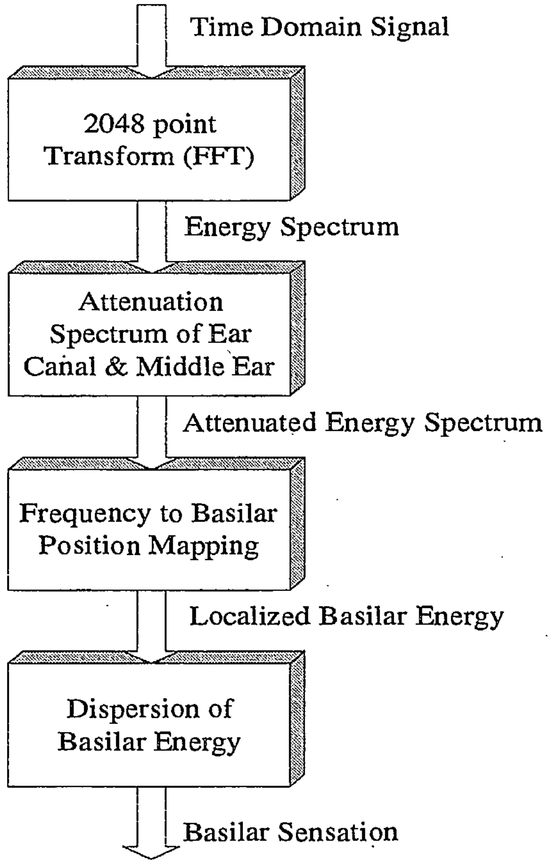

Time Domain Signal
2048 point Transform (FFT)
Energy Spectrum
Attenuation Spectrum of Ear Canal & Middle Ear
Attenuated Energy Spectrum
Frequency to Basilar Position Mapping
Localized Basilar Energy
Dispersion of Basilar Energy
Basilar Sensation

Figure 2. Peripheral ear model in *Perceval*

The following computations are performed following the transformation to the frequency domain.

- The energy spectrum is multiplied by the attenuation spectrum of a low pass filter that models the effect of the ear canal and the middle ear. The attenuation spectrum, described by the following equation (where f is in kHz), is slightly modified from that presented in [3] in order to extend the high frequency cutoff.

$$A_{dB} = -6.5\, e^{(-0.6(f-3.3)^2)} + 10^{-4} f^{3.6}$$

- The attenuated spectral energy values are transformed using a non-linear mapping function from the frequency scale to a basilar membrane pitch scale. The mapping function is a modification of one proposed in [5] in order to improve resolution at higher frequencies:

$$p = f/(af + b)$$

The frequency *f* is in Hz. The shape of the function is easily changed by choosing different values for *a* and *b*. We choose a= 9.0e-05, and b= 2.6.

It is this mapping that determines the number of pitch units in the range up to the Nyquist frequency.

- The basilar membrane components are convolved with a constant spreading function to simulate the dispersion of energy along the membrane. The spreading function applied to a pure tone results in an asymmetric triangular excitation pattern with slopes of 24 and -4 dB/Bark on the low and high frequency sides, respectively. Twenty-five Barks (a traditional unit of pitch) cover the total pitch range. The spreading function is implemented by sequentially applying two recursive filters,

  $$H_1(z) = 1/(1-a/z) \quad \text{and} \quad H_2(z) = 1/(1-bz),$$

  where the *a* and *b* coefficients are the reciprocals of the slopes of the spreading function on the dB scale.

- A small intrinsic frequency-dependent energy is added to each of the basilar membrane components to account for the absolute hearing threshold. The intrinsic energy for the *i*th pitch unit is as follows [3]:

  $$E_i = 3.65\, f_i^{-0.8}$$

- The basilar sensation vector is obtained by converting the basilar membrane component energies to decibels. Note that the energy is not converted to subjective loudness, or sones [9]. The compressive sone scale represents perceived loudness more accurately than the decibel scale. However, small differences in loudness between a reference and a test signal should not be affected dramatically by different forms of the non-linear energy transformation function.

2.2 Cognitive Model

Since the BM representation produced by the model is expected to represent only supraliminal aspects of the audio signal, this information should be sufficient to simulate results of listening experiments. However, the perceptual salience of audible basilar degradations can vary depending on a number of contextual factors. Therefore, the reference BM representation and the basilar degradation vectors are processed in various ways according to reasonable assumptions about human cognitive processing. The result is a number of variables, described below, that together produce a perceptual quality rating. A value for each variable is computed for each of three adjacent frequency ranges - 0 to 1000 Hz, 1000 to 5000 Hz, and 5000 to 18000 Hz. An exception is the measure of harmonic structure of spectrum error that is calculated using the entire audible range of frequencies.

The following features, described below, were found useful for predicting the quality of an audio sequence: average distortion level, maximum distortion level, average reference level, reference level at maximum distortion, coefficient of variation of distortion, correlation between reference and distortion patterns, and harmonic structure in the distortion.

A total of 19 variables result from these seven features when the three pitch regions are taken into account. The variables are mapped to a mean quality rating of that audio sequence as measured in listening tests. Non-linear interactions among the variables are required because the average and maximum errors should be weighted differentially as a function of the coefficient of variation. A multilayer neural network with semi-linear activation functions was applied to allow this possibility.

The feature calculations and the mapping process implemented by the neural network constitute a task-specific model of auditory cognition.

2.2.1 Average Distortion Level

For each analysis frame, the model provides a basilar error vector that describes the extent of degradation over the entire range of auditory frequencies. A positive error represents energy added to the reference signal, while a negative error represents energy taken away. A single scalar estimate of degradation for the entire sequence of frames could be obtained by integrating the vector elements over time and frequency. However, the perceptibility of distortions is likely modified by the characteristics of the current distortion as well as temporally adjacent distortions. The measured error was modified according to the following criteria.

2.2.1.1 Perceptual Inertia

A particular distortion is considered inaudible if it is not consistent with the immediate context provided by preceding distortions. This effect might be called perceptual inertia. That is, if the sign of the current error is opposite to the sign of the average error over a short time interval, the error is considered inaudible. The duration of this memory is close to 80 msec, which is the approximate time for the asymptotic integration of loudness of a constant energy stimulus by human listeners [6].

In practice, the energy is accumulated over time, and data from several successive frames determine the state of the memory. At each time step, the window is shifted one frame and each basilar degradation component is summed algebraically over the duration of the window. Clearly, the magnitudes of the window sums depend on the size of the distortions, and whether their signs change within the window. The signs of the sums indicate the state of the memory at that extended instant in time.

The content of the memory is updated with the distortions obtained from processing the current frame. However, the distortion that is output at each time step is the rectified input, modified according to the relation of the input to the signs of the window sums. If the input distortion is positive and the same sign as the window sum, the output is the same as the input. If the sign is different, the corresponding output is set to zero since the input does not continue the trend in the memory at that position. In particular, the output distortion at the *i*th position, D_i, is assigned a value depending on the sign of the *i*th window mean, W_i and the *i*th input distortion, E_i.

If (SGN(E_i) EQ SGN(W_i) AND E_i GT 0.0) $D_i = E_i$

If (SGN(E_i) NE SGN(W_i)) $D_i = 0.0$

2.2.1.2 Perceptual Asymmetry

Negative distortions are treated somewhat differently. There are indications in the literature on perception [2][4] that information added to a visual or auditory display is more readily identified than information taken away. Accordingly, *Perceval* weighs less heavily the relatively small distortions resulting from energy removed from, rather than added to, the signal being processed. Because it is considered less noticeable, a small negative distortion receives less weight than a positive distortion of the same magnitude. As the magnitude of the error increases, however, the importance of the sign of the error should decrease. The size of the error at which the weight approaches unity was somewhat arbitrarily chosen to be *Pi*, as shown in the following equation.

$$\text{If } (\text{SGN}(E_i) \text{ EQ SGN}(W_i) \text{ AND } E_i \text{ LT } 0.0)$$

$$D_i = |E_i| * \arctan(0.5 * |E_i|)$$

2.2.1.3 Adaptive Threshold for Averaging

The distortion values obtained from the memory could be reduced to a scalar simply by averaging. However, if some pitch positions contain negligible values, the impact of significant adjacent narrow band distortions would be reduced. Such biasing of the average could be prevented by ignoring all values under a fixed threshold, but frames with all distortions under that threshold would then have an average distortion of zero. This also seems like an unsatisfactory bias. Instead, an adaptive threshold was chosen for ignoring relatively small values. That is, distortions in a particular pitch range are ignored if they are less than one-tenth of the maximum in that range.

The average distortion over time for each pitch range is obtained by summing the mean distortion across successive non-zero frames. A frame is classified as non-zero when the sum of the squares of the most recent 1024 input samples exceeds 8000 (i.e., more than 9 dB per sample on average).

2.2.2 Maximum Distortion Level

The maximum distortion level is obtained independently for each pitch region by finding the frame with the maximum distortion in that range. The maximum value is emphasized for this calculation by defining the adaptive threshold as one-half of the maximum value in the given pitch range instead of one-tenth that is used above to calculate the average distortion.

2.2.3 Average Reference Level

The average reference level over time is obtained by averaging the mean level in each pitch range across successive non-zero frames.

2.2.4 Reference Level at Maximum Distortion

The value of this variable in each pitch region is the reference level that corresponds to the maximum distortion level calculated as described above.

2.2.5 Coefficient of Variation of Distortion

The coefficient of variation is a descriptive statistic that is defined as the ratio of the standard deviation to the mean [10]. The coefficient of variation of the distortion over frames has a relatively large value when a brief, loud distortion occurs in an audio sequence that otherwise has a small average distortion. In this case, the standard deviation is large compared to the mean. Since listeners tend to base their quality judgments on this brief but loud event rather than the overall distortion, the coefficient of variation may be used to differentially weight the average distortion versus the maximum distortion in the audio sequence. It is calculated independently for each pitch region.

2.2.6 Similarity of reference and distortion spectra

When the peak magnitudes of the distortion coincide in pitch with the peak magnitudes of the reference signal, perceptibility of the distortion may be differentially affected. The correlation between the distortion and reference vectors should reflect this coincidence, and this is found by calculating the cosine of the angle between the vectors for each pitch region as follows:

$$C = \frac{\vec{R} \bullet \vec{E}}{\left|\vec{R}\right| \times \left|\vec{E}\right|}$$

2.2.7 Harmonic structure in distortion

Listeners may respond to some structure of the error within a frame, as well as to its magnitude. Harmonic structure in the error can result, for example, when the reference signal has strong harmonic structure, and the signal under test includes additional broadband noise. In that case, masking is more likely to be inadequate at frequencies where the level of the reference signal is low between the peaks of the harmonics. The result would be a periodic structure in the noise that corresponds to the structure in the original signal. The harmonic structure is measured as the magnitude of the largest peak in the spectrum of the log energy autocorrelation function. The correlation is calculated as the cosine between two vectors.

3 Audio Quality Measurement

3.1 Model Calibration

The function relating the above features to listener quality ratings was calibrated using data from eight different listening tests conducted according to the ITU-R BS-1116 recommendation. These experiments are known in the ITU-R TG 10/4 as MPEG90, MPEG91, ITU92CO, ITU92DI, ITU93, EIA95, MPEG95 and DB2. The main features of these experiments are described in Appendix 1. Data from CRC97, a more recent experiment described in Appendix 2, was set aside to evaluate the generalization performance of the calibrated model.

3.1.1 Network Training

As indicated above, the variables computed by the cognitive model are mapped to the corresponding SDG by adapting the weights of a neural network. Informal tests indicated that a network with six hidden units was close to optimal. Therefore, the network architecture consisted of 19 input units, six hidden units, and one output unit. Outputs of the hidden and output units were generated using the asymmetric sigmoid activation function. Each input variable as well as the desired output value was scaled to span the range from zero to one. The training set consisted of all of the available data except those from the CRC97 experiment. Weight adaptation was performed using an accelerated form of the backpropagation learning algorithm [7].

An important concern when training a neural network is to ensure that overfitting does not occur. When the training error is reduced too much, idiosyncratic variations in the training data begin to have undue influence, and generalization to a new data set usually suffers. Overtraining can be detected by periodically testing with a validation test set not used during training. Training should be stopped at the point where generalization error with the test set is at a minimum. The purpose of the preliminary training phase was to identify the critical mean square error for the training set when validation test set performance failed to improve. The final network could then be trained with all the available data to this same level of performance with some assurance that significant overtraining did not occur.

3.1.1.1 Preliminary Training

This training phase was designed to discover when overtraining is expected to occur with this particular input distribution and output mapping, so that overtraining could be avoided when training a final network using all of the available data. The procedure uses validation test sets drawn from the same distribution as the training sets. The set of 610 available audio sequences was divided into five equivalent subsets in terms of distortion severity. All items were sorted by subjective quality rating and then divided into subsets by selecting every fifth item of the sorted sequence, each subset starting at a different offset from the beginning. Then five different training and validation test sets were created by choosing each subset as a validation test set, and combining the remaining subsets to form a corresponding training set. Each training and validation test set was used to train a different network. Training was stopped when test set performance ceased to improve.

The overall generalization performance based on all 610 sequences was obtained by combining the validation test set performance from each of the five networks. Because the outcome of network learning is not predetermined due to random initialization of the network weights, this procedure was repeated five times to measure the average critical mean square error. Therefore, the procedure required training and evaluating 25 separate networks.

A linear correlation between the system's quality prediction and the mean quality rating obtained from human listeners is used as a performance indicator. The usefulness of the present feature set was judged on the basis of overall training and validation test set performance. Table 1 shows the average correlation for each of the five training sets, as well as the validation test set performance for each network. The correlation for all the validation sets combined was 0.848. Figure 3 shows the predicted qualities versus the actual mean listener ratings for the combined set. Table 2 shows the corresponding critical mean square error. The overall average of 0.011 was used as the stopping criterion for training the final network with all the training data.

Table 1. Training and validation test set performance

R-squared values								
		Training	**Training**	**Training**	**Training**	**Training**	**Training**	**Overall**
NetID	**Replicate**	**set0123**	**set0124**	**set0134**	**set0234**	**set1234**	**avg**	**val_set**
Net1	1	0.843	0.800	0.842	0.770	0.844	0.820	**0.721**
Net2	2	0.822	0.829	0.763	0.785	0.839	0.808	**0.723**
Net3	3	0.829	0.798	0.805	0.775	0.844	0.810	**0.712**
Net4	4	0.823	0.794	0.787	0.757	0.772	0.787	**0.710**
Net5	5	0.825	0.831	0.816	0.850	0.826	0.830	**0.731**
Mean		**0.828**	**0.810**	**0.803**	**0.787**	**0.825**	**0.811**	**0.719**

Table 2. Training set critical mean square error

Training msqe at the minimum test msqe							
		Training	Training	Training	Training	Training	Training
NetID	**Replicate**	**set0123**	**set0124**	**set0134**	**set0234**	**set1234**	**avg**
Net1	1	0.009	0.012	0.008	0.015	0.008	0.010
Net2	2	0.010	0.009	0.015	0.014	0.009	0.011
Net3	3	0.010	0.012	0.011	0.014	0.009	0.011
Net4	4	0.010	0.013	0.012	0.016	0.015	0.013
Net5	5	0.010	0.009	0.010	0.008	0.011	0.010
Mean		**0.010**	**0.011**	**0.011**	**0.013**	**0.010**	**0.011**

3.1.1.2 Final Network Training

The final network was trained with all 610 sequences using the critical mean square error determined above (0.011) as the stopping criterion. The resulting correlation of predicted qualities of training items with subjective qualities was 0.914.

Figure 4 shows the predicted qualities versus the actual mean listener ratings for the items used to train this network.

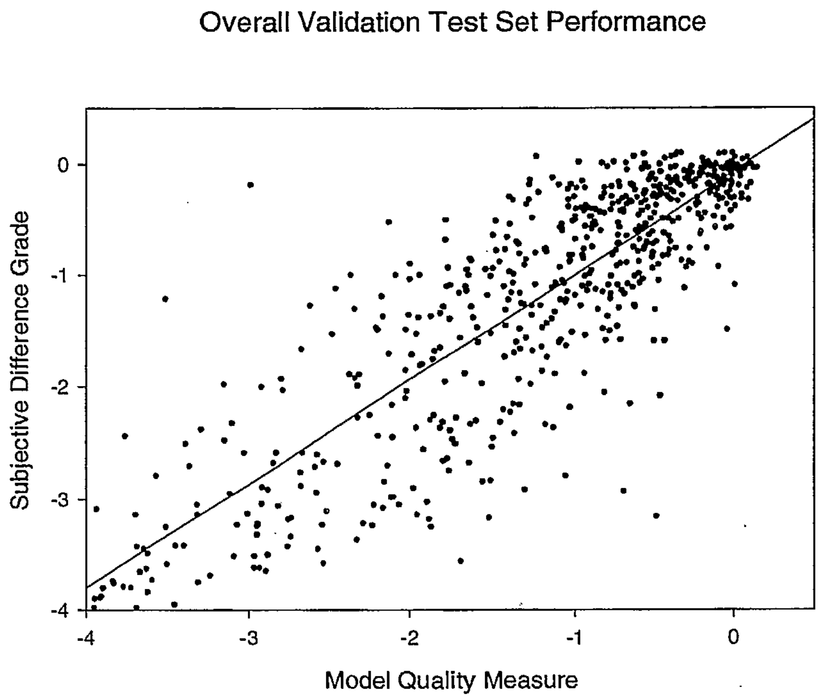


Figure 3. Model prediction vs subjective rating for validation test items

3.2 *Generalization Test*

3.2.1 Predictions of mean item quality

Because the validation test sets guarded against overtraining, they were still involved in the training process, and therefore, do not provide a true test of generalization performance. Fortunately, the ability of the final network to generalize to a truly independent data set could be evaluated by using the 136 items in the CRC97 database (Appendix 2). The correlation of predicted qualities of items in this database with subjective qualities was 0.764. This is a significant decrease compared to the correlation of 0.848 obtained from the validation test sets during the preliminary training phase.

Figure 5 shows the predicted qualities versus the actual mean listener ratings for the CRC97 data set.

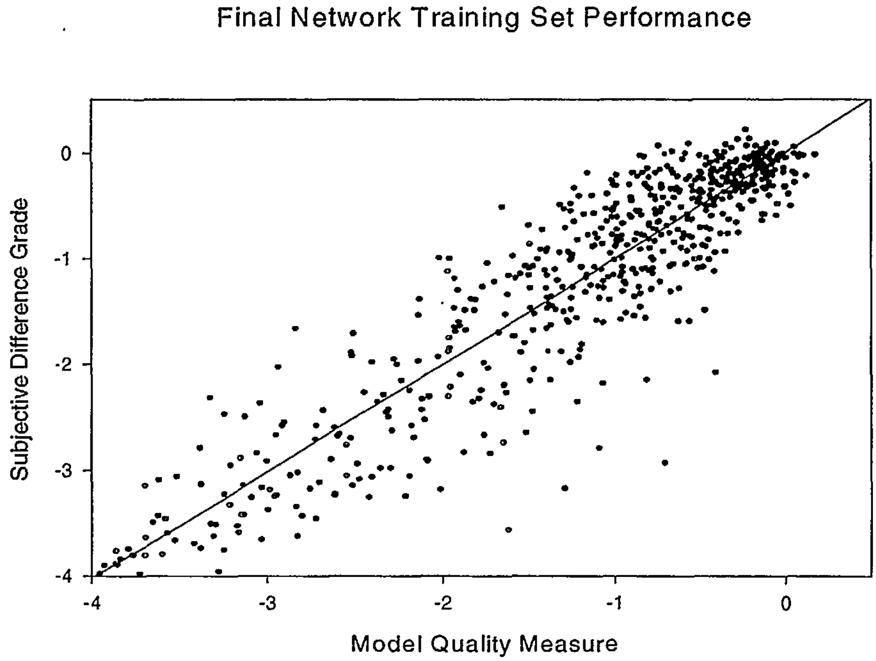


Figure 4. Model prediction vs subjective rating for training set items

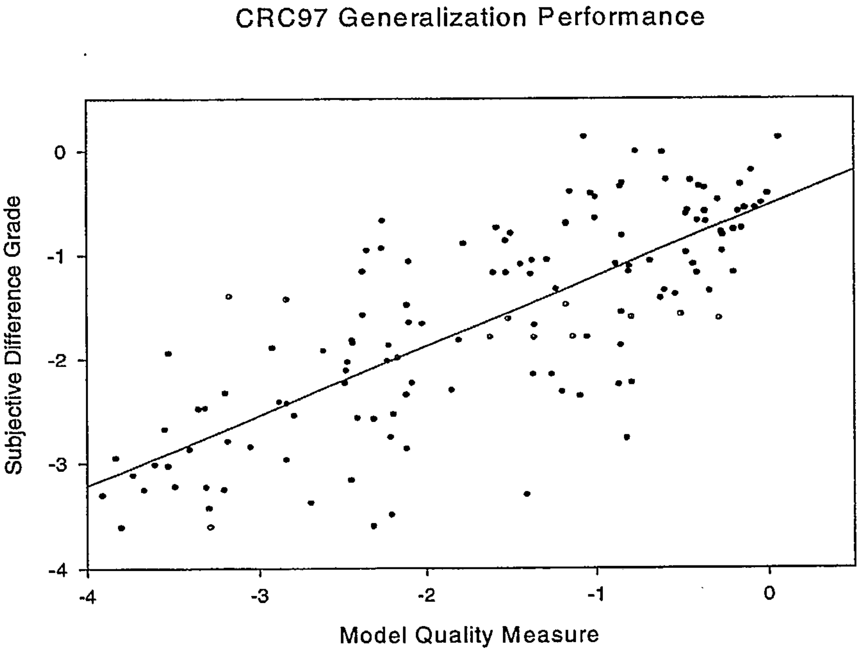


Figure 5. Model prediction vs subjective rating for CRC97 items

3.2.2 Predictions of codec performance

The CRC97 listening test (Appendix 2) was conducted to compare the performance of six state-of-the-art codecs. The perceived qualities of the codecs were compared as a function of bit rate, and the analysis permitted statements about the statistical significance of observed differences.

Since the objective quality measurement system predicts only the mean quality of an item and not the associated variance, statistical significance testing of differences is not possible. Nevertheless, it is of interest to compare the rank order of the objective means with that of the subjective means. Could the objective means have given a similar impression of the relative merits of the codecs under test?

The first comparison, shown in Figure 6, plots the 17 mean subjective ratings for the available codecs and bit rates versus the corresponding mean objective qualities. The correlation of 0.91 indicates that the overall correspondence is quite good, although there are several noticeable deviations.

The nature of these deviations becomes clearer from a more systematic comparison. Figure 7, taken from [8], shows the subjective quality of each codec at the available bit rates. The overall confidence interval obtained from the highest order analysis of variance interaction is superimposed on each point of the graph. Figure 8 shows the corresponding mean objective quality measurements.

The model did well in identifying the best performing codecs as defined by the results of the subjective listening test. The highest quality codecs were correctly found to be the AAC codec at 128 kbps and the AC3 codec at 192 kbps.

The model's largest measurement error was the excessively low quality assigned to the Layer II codec at the 160 kbps bit rate. The quality of the PAC codec at the same bit rate was also slightly lower than it should have been. Finally, the measured qualities of both the AC3 codec at 128 kbps and the Layer III codec were too high.

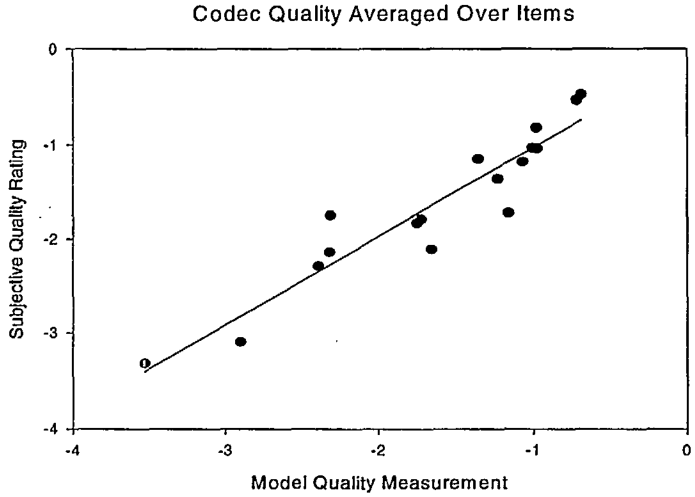


Figure 6. Model measurement of codec qualities

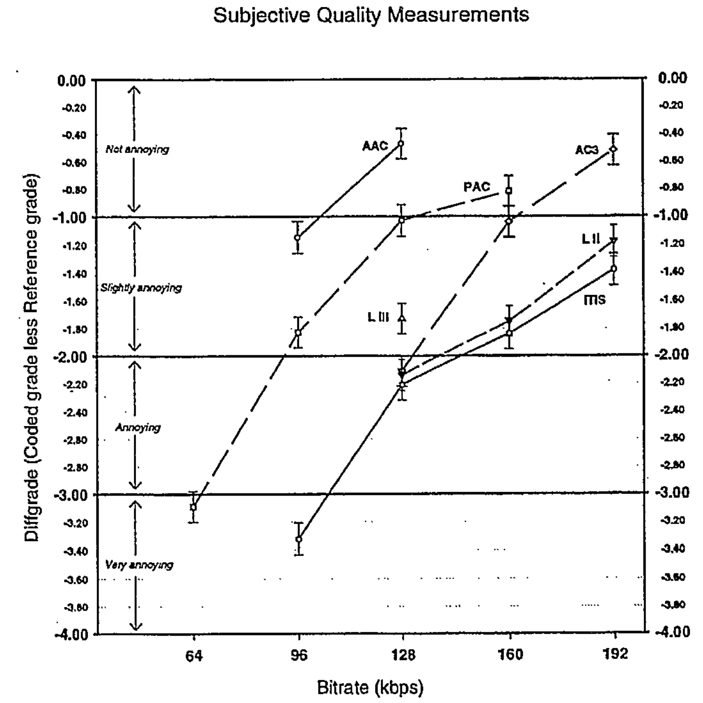


Figure 7. Subjective ratings of codecs x bit rates

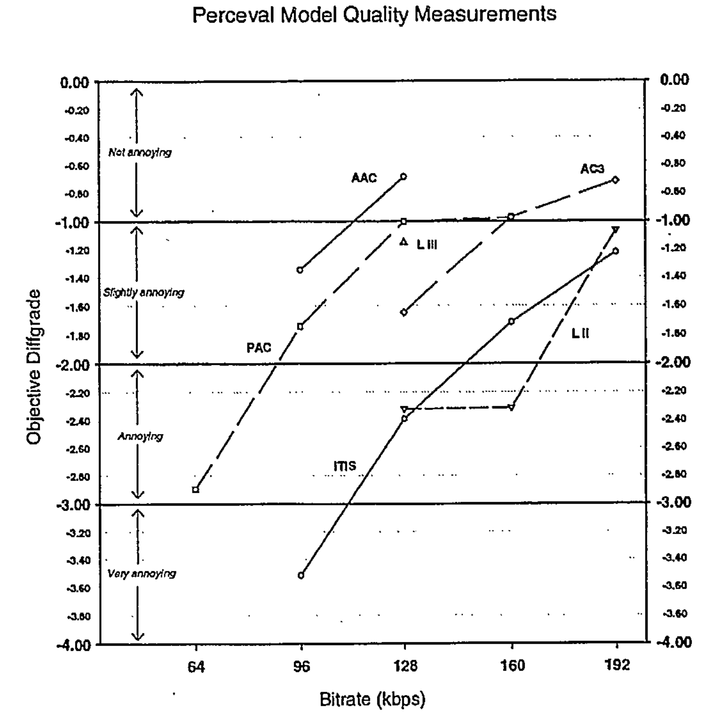


Figure 8. Model quality measurements of codecs x bit rate

4 Conclusions

The model was assessed in terms of its ability to predict the perceived quality of an individual audio item, as well as its ability to judge the average performance of a codec system.

4.1 Individual item quality measurement

The correlation between predicted and actual qualities for individual items was 0.914 for the training data and 0.848 for the validation test set drawn from the same distribution. However, the correlation decreased to 0.764 for the true generalization test data. This drop in performance might be due to inadequacies in the ear model, the cognitive model, or even the quality of the subjective test data used to calibrate the model. On the other hand, the generalization test data may contain types of distortions not present in the training data. Further work should consider all of these possibilities.

4.2 System quality measurement

Codec quality was measured by averaging the qualities of a number of individual items processed by the codec at a particular bit rate. The resulting data will have less variance, so the observed increase in the correlation with the subjective quality ratings should be expected. The pattern of results was consistent with that obtained from the subjective listening tests, and the model successfully identified the best performing codecs.

Comparison of Figure 7 with Figure 8 is useful for uncovering systematic errors made by the measurement model. For example, Figure 8 indicates that the model is overly sensitive to the kind of distortion(s) made by the Layer II codec operating at 160 kbps. Conversely, it seems somewhat insensitive to a type of distortion inserted by the Layer III codec and the AC3 codec at 128 kbps. Investigation of these anomalies could provide useful clues for improving the model.

5 References

[1] Paillard, B., Mabilleau, P., Morisette, S., and Soumagne, J. *PERCEVAL:* Perceptual evaluation of the quality of audio signals. *J. Audio Eng. Soc.*, **40**(1/2):21-31, Jan./Feb. 1992.

[2] Hearst, E. Psychology and nothing. *American Scientist*, **79**:432-443, 1979.

[3] Terhardt, E., Stoll, G., Sweeman, M. Algorithm for extraction of pitch and pitch salience from complex tonal signals. *J. Acoust. Soc. Am.* **71**(3):678-688, 1982.

[4] Treisman, M. Features and objects in visual processing. *Scientific American*, **255**[5]:114-124, 1986.

[5] Zwicker, E. and Terhardt, E. Analytical expressions for critical-band rate and critical bandwidth as a fuction of frequency. *J. Acoust. Soc. Am. 68*(5): 1523-1525, 1980.

[6] Zwislocki, J.J. Temporal summation of loudness: An analysis. *J. Acoust. Soc. Am.* **46**(2):431-441, 1969.

[7] Jacobs, R.A. Increased rates of convergence through learning rate adaptation. *Neural Networks*, **1**:295-307, 1088.

[8] Soulodre, G.A., Grusec, T., Lavoie, M., and Thibault, L. Subjective evaluation of state-of-the-art 2-channel audio codecs. *J. Audio Eng. Soc.*, **46**(3):164-177, 1998.

[9] Zwicker, E. and Fastl, H. *Psychoacoustics: Facts and Models.* Springer-Verlag, Berlin, 1990.

[10] Freund, J.E. *Modern Elementary Statistics.* Prentice-Hall, Englewood Cliffs, 1967.

Appendix 1 – Description of the Training Database

Database I consists of audio files and corresponding observer ratings from five different listening tests. They are as follows:

DataSet Name: MPEG90

- Ten stereo sequences:
  - Suzanne Vega
  - Tracy Chapman
  - Glockenspiel
  - Fireworks
  - Ornette Coleman
  - Bass Synth
  - Castanets
  - Male speech
  - Bass Guitar
  - Trumpet (Haydn)
- Processed by two codecs at three bit rates (64, 96, and 128 kbit/sec/channel)
- The mean subjective difference grade (SDG) per item quite uniformly covered the range from 0.01 to –3.98.

DataSet Name: MPEG91

- Nine stereo sequences:
  - Suzanne Vega
  - Carmen
  - Male speech
  - Ornette Coleman
  - Accordian/Triangle
  - Tambourine
  - Bass guitar
  - Glockenspiel
  - Percussion
  - George Duke
- Three bit rates: 64, 96, and 128 kbits/sec/channel
- Seven codecs:
  - MPEG Layer I
  - MPEG Layer II
  - MPEG Layer III
  - MUSICAM
  - ASPEC
  - NICAM

At least 88 percent of the mean SDG per item were above –2.0, and the range was 0.09 to -3.75.

DataSet Name: ITU92DI

Twelve stereo sequences:

Asa Jinder
Dalarnas Spelmarsforbund Trettondagsmarchen
Stravinsky: Wind Octet
Triangles
Solo harpsichord
Castanets
German male speech
Ornette Coleman
Bass guitar
Suzanne Vega
Ravel: "Feria" (Spanish Suite)
Dire Straits: "Ride Across the River"

Five distribution codecs: (120 kbits/sec/channel)
Each item was processed by the same codec three times in tandem, with a 0.1 dB drop in level before each pass. Each channel of the stereo pair was coded independently.

ISO Layer 2
ISO Layer 3
Dolby AC-2
Aware
NHK

Eighty percent of the mean SDG per item were above –2.0, and the range was 0.13 to -3.43.

DataSet Name: ITU92CO

Ten stereo sequences:

Asa Jinder
Dalarnas Spelmarsforbund Trettondagsmarchen
Stravinsky: Wind Octet
Triangles
Solo harpsichord
Castanets
German male speech
Ornette Coleman
Bass guitar
Suzanne Vega

Six contribution codecs: (180 kbits/sec/channel)
Each item was processed by the same codec three times in tandem, with a 0.1 dB drop in level before each pass. Each channel of the stereo pair was coded independently.

ISO Layer 2
ISO Layer 3
Dolby AC-2
Dolby Low-Delay
Aware
ATT DSQ 5TR620

At least 96 percent of the mean SDG per item were above three, and the range was 0.22 to –2.41.

DataSet Name: ITU93

Seven stereo sequences:

German male speech
Solo castanets
Asa Jinder
Bass clarinet arpeggio
Solo harpsichord arpeggio
“Vi salde vara hemman” (solo violin)
Bagpipes

ISO Layer II tandem codec configurations: (bit rate given for stereo pair)

Emission codec alone at 256 kbit/sec (independent channel coding)
Emission codec alone at 192 kbit/sec (joint stereo coding)
Eight contribution codecs at 360 kbit/sec followed by one emission codec at 256 kbit/sec
Eight contribution codecs at 360 kbit/sec followed by one emission codec at 192 kbit/sec
Five contribution codecs at 360 kbit/sec followed by three distribution codecs at 240 kbit/sec and one emission codec at 256 kbit/s
Five contribution codecs at 360 kbit/sec followed by three distribution codecs at 240 kbit/sec and one emission codec at 192 kbit/s

Listening tests were performed by CRC (Canada) and RAI (Italy). Most of the mean SDG per item were above –2.0, and the range was -0.08 to -2.27. There was no significant difference between the data from the two labs.

DataSet Name: EIA95

Experimental results are published in the IEEE Transactions on Broadcasting, Vol. 43, No. 3, Sept. 1997.

Nine stereo sequences:

Bass clarinet arpeggio
Dire Straits cut
Glockenspiel
Harpsichord arpeggio
Music and rain
Pearl Jam cut
Muted trumpet
Suzanne Vega with breaking glass
Water sound

Nine codecs:

Eureka 147 #1	224 kbps/2 channels
Eureka 147 #2	192 kbps/2 channels
AT&T/Lucent	160 kbps/2 channels
AT&T/Lucent/Amati #1	128 kbps/2 channels
AT&T/Lucent/Amati #2	160 kbps/2 channels
VOA/JPL	160 kbps/2 channels
USADR-FM #1	128-256 kbps/2 channels
(variable bit rate)	

USADR-FM #2 (variable bit rate)	128-256 kbps/2 channels
USADR-AM	96 kbps/2 channels

At least 93 percent of the mean SDG per item were above –2.0, and the range was 0.14 to -3.73.

DataSet Name: DB2

Eighteen stereo sequences:

- Bass clarinet
- Clarinet
- Clarinet+horn
- Horns
- Horn
- Strings
- Oboe
- Oboe+string bass
- Castanets
- Trumpet
- Tambourine
- Triangle
- Drum
- Glockenspiel
- Xylophone
- Tuba
- Speech (German, female)
- Singing (Suzanne Vega)

Types of distortions:

Tandem Layer II, 256 kbit/s	1,3,5,7 and 9 stages
Tandem Dolby AC2	1,3,5,7 and 9 stages
Layer II, 256 kbit/s	
Layer II, 192 kbit/s js	
Layer II, 64 kbit/s mono	
Layer II, 384 kbit/s	
Layer III (ASPEC3), 192 kbit/s	
Layer III ASPEC3), 128 kbit/s	
Layer III ASPEC3), 160 kbit/s	
MPEG2/L2 LSF	
APT-X, 256 kbit/s	
APT-X, 384 kbit/s	
Quantizing distortions	
Analogue distortion	
Digital errors	
Clipping	

The database consisted of 91 items. At least 83 percent of the items were given a mean SDG above –2.0, and the range was 0.0 to -3.98.

Appendix 2 – Description of the Test Database

DataSet Name: CRC97

Experimental results are expected to appear in the Journal of the Audio Engineering Society, March, 1998.

Eight stereo sequences:

Bass clarinet
Double Bass
Dire Straits
Harpsichord
Music and rain
Pitch pipe
Trumpet
Suzanne Vega

Six codecs: (bit rate given for stereo pair)

ATT PAC	64, 96, 128, and 160 kbit/sec
Dolby AC3	128, 160, and 192 kbit/sec
MPEG Layer II software	128, 160, and 192 kbit/sec
MPEG Layer II hardware (ITIS)	96, 128, 160, 192 kbit/sec
MPEG AAC	96 and 128 kbit/sec
MPEG Layer III	128 kbit/sec

The mean SDG per item quite uniformly covered the range from 0.13 to –3.60.

Industry Canada Industrie Canada

Canada