A NEW AUDITORY MODEL FOR THE EVALUATION OF SOUND QUALITY OF AUDIO SYSTEMS

Matti Karjalainen

Helsinki University of Technology
Acoustics Lab., Otakaari 5 A
02150 Espoo 15, Finland

ABSTRACT

A new computational model of auditory spectrum analysis is presented and its validity in sound quality measurements is shown. The model is based on the most important features of peripheral hearing in time and frequency known from the psychoacoustic theory. The methodology of sound quality measurement is based on the use of real signals (e.g. speech) as test signals and auditory spectrum distance as a measure of sound quality degradation. This paper describes the new auditory model and the results of testing its validity. A good correlation between subjective and our objective measures of distortion is shown. Auditory spectrum distance far outperforms the traditional distortion measures in this sense. Some limitations and further ideas concerning the model are also discussed.

MODELLING OF AUDITORY SPECTRUM ANALYSIS

In our previous studies we have shown the applicability of psychoacoustically oriented spectrum analysis to the measurement of distortion level of speech and audio signals; /1/, /2/, /3/. The method was based on FFT and some further transformations and it was found to work well in the case of steady-state signals. If the spectrum is changing rapidly, however, we should carefully implement the time domain properties of the human auditory system, too. Such model should include the following phenomena in sufficient detail:

- **Frequency selectivity** of about 1 Bark and masking effect in frequency domain (spreading function).
- **Frequency sensitivity** of the human ear according to the loudness curves (60 dB-level, e.g.).
- **Temporal integration**; time response of any 1 Bark channel should be its power lowpass-filtered by a time constant of 100 - 200 ms.
- **Temporal masking**; pre- and postmasking effects (forward and backward masking).

FILTER-BANK MODEL FOR AUDITORY SPECTRUM ANALYSIS

It was found to be difficult to include proper temporal dynamics and Bark scale when using FFT-based transformations. The filter-bank principle is well suited to auditory spectrum analysis because the human auditory system - basilar membrane and hair cells - also consists of a multi-channel analyzer. The bandwidth of the overlapping channels is about one critical band or 1 Bark. Instead of thousands of hair cells it is enough to have 1 - 4 channels per one Bark in a computational model. This means 24 - 96 channels covering the 24 Bark audio range. With 0.5 Bark spacing our model has 48 channels, which seems to be a practical compromise between good resolution of spectral representation and low amount of computation.

Each channel consists of a bandpass filter, a square-law rectifier, a fast linear and a slower nonlinear lowpass filter, and a dB-scaling stage, see Fig. 1.

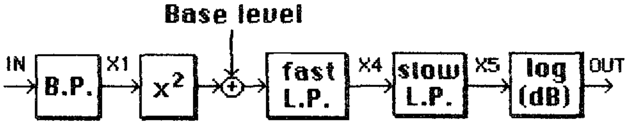

FIG 1. One channel of the 48-channel filter bank used for auditory spectrum computation. B.P = bandpass, L.P. = lowpass filter, x^2 = square-law detection, LOG = dB-scaling.

Bandpass filter bank

Bandpass filters with 0.5 Bark spacing and a little more than 1 Bark bandwidth give the desired frequency selectivity to the model. Each bandpass is a 256-order FIR-filter carefully designed to have frequency response which is the mirror image of the spreading function B(x) given by Schröder et al /4/.

This filter gives a good approximation of the desired masking properties in frequency domain. Computation of the filter bank has been implemented as a matrix multiplication in an array processor (Floating Point Systems, FPS 100). A 48-element vector filled with a new input signal sample is multiplied by a 256 x 48 -sized matrix to get the outputs of the channels as a vector. This is repeated 20 000 times per second, which is the sample rate in our system. Even an array processor cannot run it in real time. By a proper IIR-filter design the speed of the computation could be ten times faster but accurate design of the filters is a difficult task.

Not only frequency selectivity but also the frequency response (sensitivity) of the ear must be built into the filter bank. The simple way we used is to let the relative gains of the channels vary according to the inverse of the equal loudness curve (60 dB-level).

CH2118-8/85/0000-0608 $1.00 © 1985 IEEE

Rectification

The rectification effect in hair cells of the inner ear is primarily of half-wave type. Because in our model we needed a square-law element, we ended up to use it without any half-wave rectifier. We found that in auditory spectrum analysis of speech this makes no remarkable difference. A a constant level is added after the rectification to simulate the threshold of hearing.

Filters of temporal integration

The remaining two filters are for smoothing the outputs of the selective channels. The faster one is a first-order lowpass with time constant of about 3 ms. Its role is not important here. The second one is more fundamental. Its purpose is to implement many effects; temporal integration and pre- and postmasking effects.

Temporal integration is realized simply by linear first-order filtering (time constant about 100 ms) applied to the output of square-law rectification. Premasking is not a very important and critical phenomenon. No extra tricks were needed to match it well enough.

Postmasking was more difficult to be implemented in sufficient detail. Linear lowpass of 100 ms time constant gives many times too long overall masking. We used nonlinear (logarithmically linear) behaviour of the filter for masking situations. The form of the masking pattern and its duration are now close to the real ones but a delay of about 10 ms in the real masking pattern /5/ is lacking. In our application this is not important, however. The overall response of the slow nonlinear lowpass can be stated now:

For $X4 \geq X5$ (temporal integration)

$X5(n) = K1 * X4(n) + (1-K1) * X5(n-1)$, and

For $X4 < X5$ (postmasking)

$X5(n) = X5(n-1) * \exp(K2 * \log(X4(n)/X5(n-1)))$,

where X4 and X5 are the input and the output of the filter, K1 and K2 filter coefficients and n discrete time index. A good value for K1 is 0.0005 and 0.0007 for K2 corresponding to a sampling frequency of 20 kHz.

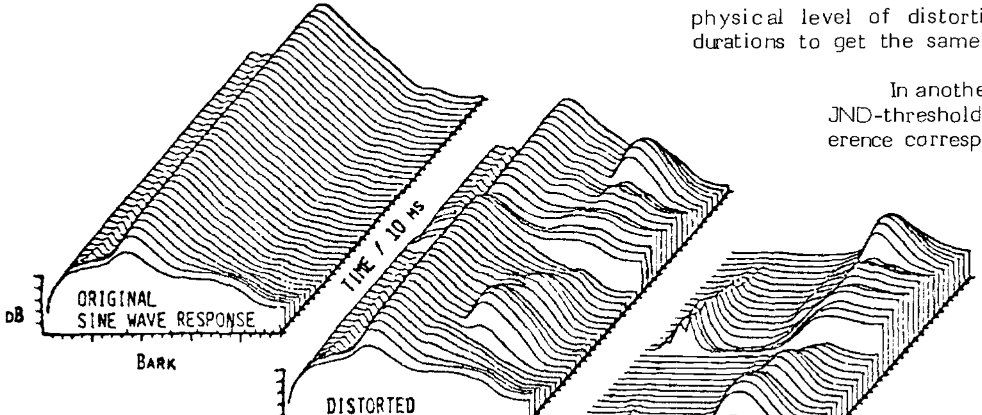

FIG 2.
Examples of auditory spectrum representation; pure sine wave (left), the same with nonlinear distortion (middle) and spectral difference (right)

AUDITORY SPECTRUM DISTANCE AS A MEASURE OF DISTORTION

If the auditory spectrum of a sine wave starting at t=0 is plotted in three dimensions of time (ms), pitch (Bark) and amplitude (dB), we get the leftmost picture of Fig. 2. If the same sine wave is distorted in different nonlinear ways in various parts of the signal the spectral representation in the middle of Fig. 2 is resulted. By subtracting this from the first one we finally get the auditory spectrum difference (right hand side). The maximum value of spectral deviation evaluated over the time and Bark scales is our first candidate for the measure of distortion.

PERCEPTION THRESHOLD OF NONLINEAR DISTORTION

One of the most useful rules of the psychoacoustic theory is the 2-dB-rule of just perceivable change. This means that any variation in a sound, resulting in about 2 dB level change in any Bark channel, will be noticeable in subjective listening tests. Now we can test if this rule is valid also for JND-threshold of nonlinear distortion.

We tested the hypothesis by distorting three Finnish speech sounds /a/, /i/, /s/ with three nonlinear distortions (square-law, crossover and clipping). Duration of the distorted sound was the third variable. Three persons were asked to find the just noticeable levels of distortions in an anechoic chamber (direct comparison of distorted and undistorted signals). The corresponding maximal distances in auditory spectra were computed next. Fig. 3 shows the results as a function of distortion duration.

It was found that the types of distortion and speech sound have no essential effect on the auditory spectrum distance of JND-distortion. Duration also has only a minor effect. The 2-dB rule is valid or, more exactly, distortion is just perceivable when the maximum value of auditory spectrum distance is about 1.5 - 2.5 dB (undistorted reference was available to the listener).

An interesting detail is that the temporal integration must really be present in the model. This means also that if the duration of distortion is less than about 100 ms, the physical level of distortion must be higher for short durations to get the same threshold of perception.

In another experiment we found that the JND-threshold of distortion without pure reference corresponds to 1.5 - 13 dB distances depending on types of distortion and speech sound. We can conclude that if the distance is less than 1.5 dB, the distortion is practically never perceivable.

We have shown the validity of the 2-dB-rule in other cases, too. Any linear distortion (frequency response error) exceeding this level can be heard. The test signal may be periodic (harmonic) or random, but also a nonharmonic complex of voiced sounds. Spectral distance can be used also as a measure of signal-to-noise ratio (SNR). The subjective threshold of noise-on-voiced-sound perception corresponds to about 0.5 dB maximum auditory spectrum distance, instead of the 2 dB rule.

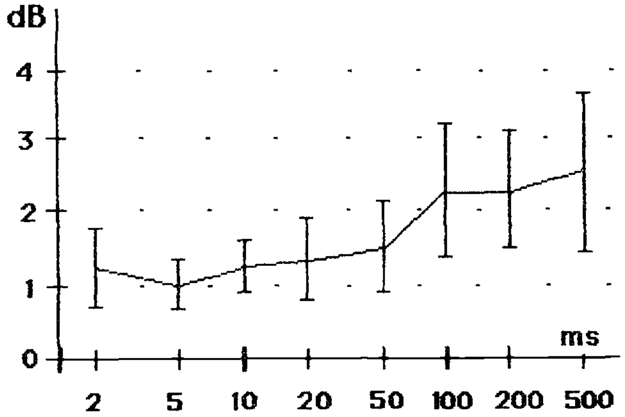

FIG. 3. Auditory spectrum distances corresponding to the JND-thresholds of different distortions applied to three speech sounds (see text) as a function of distortion duration.

SUBJECTIVE SCALE OF DISTORTION VS. AUDITORY SPECTRUM DISTANCE

The next step of our study was to test if equal subjective levels of different distortions on different sounds are related to equal distances of auditory spectra. In our experiments we found evidence strongly supporting this hypothesis. We used three persons, two types of distortion and many levels of spectral distance over the JND threshold. If two cases were perceived equally distorted, the spectral distances caused by the distortion were equal within about +/- 2 dB with zero mean value. Each experiment was arranged so that the listener heard a pure example, a reference distortion (e.g. 10 dB spectral distance), another pure example and the test distortion to be compared with the reference distortion. This complex was repeated until the subject had decided if the distortion levels were equal or not.

The results were interpreted to confirm a systematic psychoacoustic basis for the perception of distortion. The final step of our present study was to search for a subjective scale of distortion. This can be done by relating the spectral distance, an objective measure, to a subjective distortion index.

We used two kinds of subjective scales, rating from 0 to 10. The first one was fixed only by a single reference sound prior to test signals; point 5 of the scale, corresponding to a distortion of 5 dB spectral distance. (Another natural reference point is 0 vs. no distortion.) The resulting curves (averages of three listeners) for three speech sounds and the total average are plotted in Fig. 4a to show the relation of subjective and objective scales. Differences between individual listeners were in order of +/- 1 step on the subjective scale.

Another fixation of the subjective scale was given to the values of 0 to 10 by descriptive definitions in the following manner:

(0) No audible distortion.
(1) The listener supposes to have heard something like distortion but is not sure.
(2) Distortion is on the just noticeable threshold.
(3) Distortion is always perceived when concentrating on listening.
(4) Distortion can be heard easily as "soft" distortion.
(5) Distortion is not "soft" anymore, but not yet disturbing.
(6) Distortion is now disturbing.
(7) Listener feels some uncomfort because of distortion but the sounds are still easily recognized.
(8) Distortion is increased to the level where some problems of correct recognition exist.
(9) Recognition of the sounds is like guessing.
(10) Recognition of the sounds is impossible.

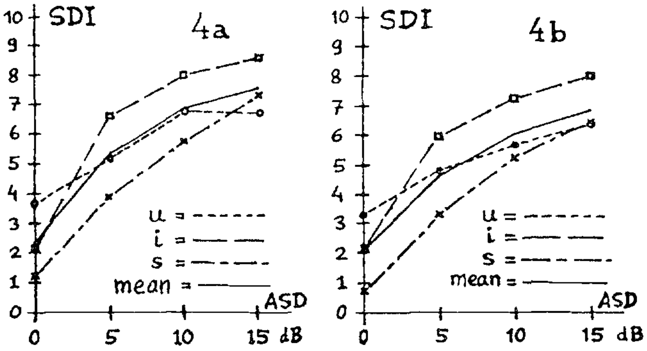

FIG. 4. Subjective distortion index as a function of objective measure (auditory spectrum distance) for three speech sounds and the average curve when two kinds of subjective scale fixation are used (see text).

Fig. 4b illustrates the results of the tests. Only small differences are seen between cases 4a and 4b; both pictures describe equally how the relation of spectral distance in dB and subjective distortion index is formed.

Subjective index is a monotonic function of spectral distance showing some saturation when distortion increases, and values in order of 2 for undistorted signals. (Undistorted references were not available to the listeners.) Differences between speech sounds exist, but negligible variations among listeners showed useful stability of the approach.

All cases in Fig. 4b where the subjective index is equal to or more than 6 were defined to have a "disturbing" amount of distortion. The level of disturbance is important in the evaluation of any audio systems, speech transmission in this case. The threshold of disturbance was also studied in another test. Differences between listeners were found to be about +/- 1 degree of distortion index. The thresholds were; 6.5 dB (i), 9.7 dB (u), 11.7 dB (s), and 9.3 dB average value. Distortion type (crossover and square-law here) showed less effect on the results than the type of sound.

In some preliminary experiments we have used words of real speech as test signals instead of single sounds. The threshold of disturbance was found be in the same range as for steady-state sounds (6 - 13 dB).

DISCUSSION

Our psychoacoustical approach to auditory modelling had two goals; one to develop the models as such and the other to apply them to practical purposes. The present model of auditory spectrum analysis was tested here by a series of experiments in order to see its limits of performance and, at the same time, to find a new and natural methodology to measure sound quality in audio and speech systems. This first application of the model was found very promising; the new method of measuring nonlinear distortion outperforms all traditional measures (e.g. harmonic and intermodulation) by having a good correlation to the results of subjective tests.

The methodology and approach is believed to be of general nature. Auditory spectra could be used in many cases of speech and audio signal processing to represent human auditory processing. We can find many limitations in the present model, too. These are of three types, each one with a corresponding way to develop the models further.

The first and the simplest way is to make the present model more accurate and more efficient computationally without including any essentially new features. Another way is to add further processing and transformations after the auditory spectra. This means segmentation, parametric, structural, and symbolic representations, "phonetic" interpretation of the raw auditory spectra, etc.

The third type of development which is needed is to analyze the periodicity properties and fine time-structure of auditory signals. It should be done parallel to the auditory spectrum analysis. We have gathered evidence that e.g. the phase in many cases is much more important feature than is traditionally recognized. We have shown that without any change of (amplitude) spectrum it is possible to add some subjective "distortion" by manipulating the phase only.

What is needed here is a relevant concept of auditory phase, different from the Fourier-transform-related phase. It could be closely connected to the phase and periodicity found in envelope signals of the rectified Bark-channel outputs, see Fig. 1.

An immediate indication of the importance of periodicity analysis is the fact that the auditory spectrum does not tell anything about periodic or non-periodic nature of signals. By proper auto- and cross-correlations of the Bark-filterbank outputs we have made preliminary experiments on such analysis and, what seems to be one of the most difficult questions, sound separation. This means the extraction of a sound object (one source) from those of other sources (e.g. speakers) mixed into a single channel. Such new models will grow in complexity, but modern signal processors and computer technology will give more and more powerful tools to realize this complexity in practical form.

As a practical step further we have started to design a measurement system based on the present model and a signal processor (TMS 320). The system should preferably have full audio range A/D- and D/A-converters and mass memory (hard disk) for test signal data base. The procedure for a typical sound quality measurement could be; (1) frequency response by using a test signal having a flat spectrum (e.g. white noise), (2) nonlinear distortion with a number of "real" signals (e.g. speech) with different spectra and amplitudes, (3) optionally S/N-ratio with the same "real" test signals. The static frequency response (1) must be compensated before using the auditory spectrum distance as a measure of distortion.

ACKNOWLEDGEMENTS

The auditory model was developed in a project sponsored by the Academy of Finland. The study of applicability of the model to the measurement of distortion in speech transmission was financed by the Finnish PTT.

REFERENCES

/1/ KARJALAINEN M., Measurement of Distortion in an Audio Signal Channel Based on Psychoacoustic Models. Proc. of NAS-82, Stockholm 1982.

/2/ KARJALAINEN M., Objective Measurements of Distortion in Speech Signal Channels by Computational Models of Speech Perception. Proc. of 11th ICA, Paris 1983.

/3/ KARJALAINEN M., Sound Quality Measurements of Audio Systems Based on Models of Auditory Perception. Proc. of IEEE ICASSP-84, San Diego 1984.

/4/ SCHRÖDER M. et al., Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception. In the book: Frontiers of Speech Communication Research (ed. Lindblom & Öhman), Academic Press 1979.

/5/ ZWICKER E., FELDTKELLER R., Das Ohr als Nachricten-empfänger. S. Hirzel Verlag, Stuttgart 1967.