PERCEPTION AND MEASUREMENT OF DISTORTION IN SPEECH SIGNALS – AN AUDITORY MODELLING APPROACH

Seppo Helle Matti Karjalainen

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

ABSTRACT

The perception of nonlinear distortion in speech signals was studied. Subjective listening tests were carried out using Finnish speech sounds as test material. A computational model was used to obtain auditory spectra from the undistorted and distorted sounds, and the spectral difference was compared to subjective sound quality evaluation.

Our studies showed the so-called 2-dB deviation rule to be a useful measure for the just noticeable level of nonlinear distortion. This rule implies that if the changes in auditory spectrum exceed 2 dB, the difference between the original and distorted sound can be perceived. This result also verifies the applicability of the psychoacoustic approach to distortion perception. For distortions exceeding the perception threshold, a more sophisticated objective measure than the maximum spectral deviation is needed. A distortion measurement system based on an auditory model has also been constructed.

INTRODUCTION

The work with auditory models has been active in our laboratory since 1981 /1/ - /4/. One aim of the research has been a psychoacoustical model imitating the human hearing process. A mathematical model that performs this is not a physical simulation of the hearing system. Instead, it attempts to imitate the functional properties of subjective perception of the sound, no matter what kind of physical processes there exist. This is our approach to auditory modelling.

Auditory models can help us, for example, to create better measuring techniques of nonlinear distortion. Conventional techniques, like harmonic distortion measurement, don't take into account how we actually perceive the distortion. This might lead to incorrect results and not to what we want – the sound quality in terms of perceived distortion. If the important properties of the auditory system are built into the measurement method, results can be improved.

Application areas include speech recognition and speech analysis for phonetic speech research. These auditory models can provide some new insights to how we perceive speech.

Some important phenomena of the human auditory system that should be implemented in auditory models are:

- Frequency selectivity of about 1 Bark and masking effect in frequency domain (excitation 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.

FILTERBANK MODEL

The filterbank 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 /6/. The bandwidth of the overlapping channels is about one critical band or one 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 (fig.1).

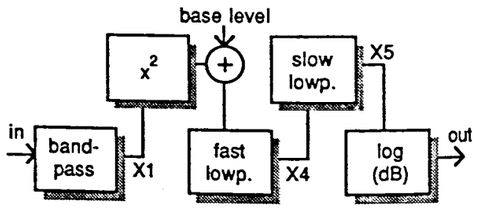

Fig. 1. One channel of the 48-channel filterbank used for auditory spectrum configuration. x^2 = square law detection, log = dB-scaling.

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 designed to have a frequency response which is the mirror image of the spreading function B(x) given by Schröder et al /5/.

Not only frequency selectivity but also frequency response (sensitivity) of the ear must be built into the filterbank. 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).

The rectification effect in hair cells of the inner ear is primarily of half-way type. Our model did not have a half-wave rectifier, because a square-law element was included. We found out that in auditory spectrum analysis of speech this makes no remarkable difference. A constant level is added after the rectification to simulate the threshold of hearing.

The remaining two filters are for smoothing the outputs of the selective channels. The faster one is a first-order low-pass 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 as well as pre- and postmasking.

Temporal integration is realized 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, and this simple solution was quite sufficient.

Postmasking was more difficult to be implemented. A linear lowpass filter with a 100 ms time constant gives an overall masking that is several times too long. To make a better match we used nonlinear (logarithmically linear) behaviour of the filter for masking situations /3/.

PERCEPTION THRESHOLD OF NONLINEAR DISTORTION

One of the most useful rules of the psychoacoustic theory is the 2-dB rule of just perceivable difference. This means that any variation in a sound, resulting at least in about 2 dB level change in any Bark channel, will be noticeable in subjective listening tests. The hypothesis was tested by distorting three Finnish speech sounds /a/, /i/ and /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 (JND). The test was made by direct comparison of distorted and undistorted signals from a loudspeaker in an anechoic chamber. The corresponding maximal distances in auditory spectra were then computed. The results are shown in fig. 2.

It was found that the types of distortion and speech sound have no essential effect on the auditory spectrum distance of JND-threshold. 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).

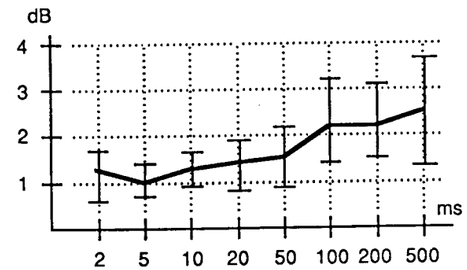

Fig. 2. Auditory spectrum distances corresponding to the JND-thresholds of different distortions applied to three speech sounds (see text) as a function of distortion duration.

An interesting detail is that the the temporal integration must really be present in the model. This also means that if the duration of distortion is less than 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 perception threshold of distortion without pure reference correponds 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.

SUBJECTIVE DISTORTION EVALUATION VS. AUDITORY SPECTRUM DEVIATION

Another series of experiments was carried out later to investigate further the correlation between maximum auditory spectrum distance and subjective distortion evaluations, this time especially for higher than JND levels. Test sounds were Finnish vowels /a/, /i/ and /u/ spoken by two male speakers. Test samples were about 200 ms long and they were distorted artificially with four types of distortions: zerocrossing, clipping, square-law and angle distortions (angle distortion: a piecewise linear input-output relation having an angle discontinuity at the origin). In each test, one of the test vowels was played to the listeners with different distortions in a random order. A test series contained 6 - 8 distortion levels for each distortion type plus clean signals. The undistorted reference could be listened to before the series, but not between the test signals. Each test signal could be repeated as many times as required before making the evaluation using a scale from 0 to 10. Definitions for the values on the scale were:

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 sound is still easily recognized.
8 Distortion is increased to the level where some problems of correct recognition exist.
9 Recognition of sounds is like guessing.
10 Recognition of the sounds is impossible.

There were three test subjects, all of which listened to each series five times. Figures 3 - 5 show the results from three vowels (/a/, /i/ and /u/) of one speaker. The figures present subjective evaluations of distortion as a function of maximal auditory spectrum distance over time and full 24 Bark range. On the y-axis is the evaluation scale that was used in the test. (Presented are only three of the six test sounds, but the results from the other speaker's sounds were roughly of the same type.)

The plots show immediately that the vowel /i/ is the most sensitive of the sounds: that is, distortion is easiest to detect. The other sounds /a/ and /u/ are less sensitive to distortion.

From the plots it is seen that the vowel /i/ exhibits the least variation between the four types of distortion while /u/ exhibits the most. If we look at fig. 5, we see that for the vowel /u/ the spectral difference corresponding to the "disturbing threshold" (value 6) is over 20 dB for square-law distortion, but only about

10 dB for crossover distortion. For the other speaker's /u/, however, the characteristics of the four distortion type curves were different (variations were again large, but the order was different).

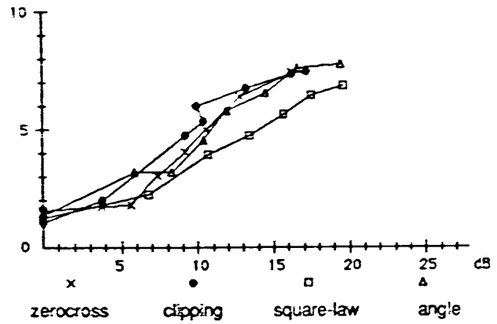

Fig. 3. Subjective distortion evaluation vs. maximal auditory spectral deviation. Vowel: /a/. Average from 15 evaluations for each point.

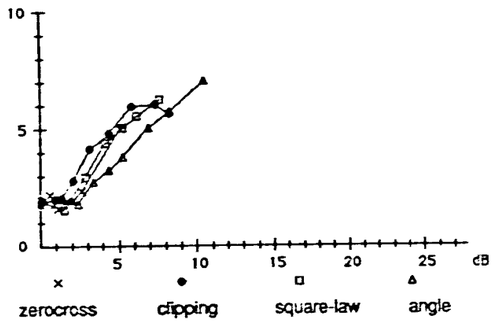

Fig. 4. Subjective distortion evaluation vs. maximal auditory spectral deviation. Vowel: /i/. Average from 15 evaluations for each point.

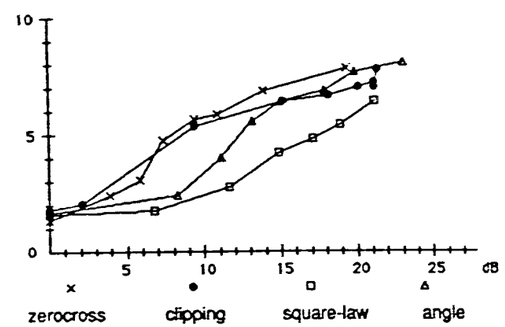

Fig. 5. Subjective distortion evaluation vs. maximal auditory spectral deviation. Vowel: /u/. Average from 15 evaluations for each point.

Considering the results we can say that although the auditory spectrum method is good at JND threshold, it has only moderately good correlation to subjective distortion evaluation at higher distortion levels. Therefore the method needs further refinements. Possible ways of doing this are: (1) to define a better distortion measure than maximal spectral deviation, and, (2) to improve the auditory model itself.

Improving the distortion measure

Some possible ways of changing the distortion measure are:

- Frequency weighting. The current measure handles all the 48 channels in the model equally, but it could be advantageous to give more weight to the highest channels, since high-frequency components are usually more disturbing than lower ones.
- Area and level weighting. The distortion measure could be made a function of the geometrical area of the spectral deviation, which would give a measure related to the total amount of distortion.

Changing the auditory model

Our model does not take into account what happens inside one pitch period of speech sound but rather only the long-term perception phenomena are considered. However, it is known that the temporal fine structure of sound has some effect on the perception. If the time constants of the model were shortened so that the fine structure of the signal would have an effect on the auditory spectra, this could give some extra information about the signal. In the case of distortion perception this information could be important: for example, if one distortion mechanism distorts only the peaks of the signal (say, clipping), it may have a different subjective effect than another type which has more effect on the low-level parts (crossover).

AUDITORY MODELLING APPROACH IN DISTORTION MEASUREMENT

Since the 2-dB rule is found to correlate well with distortion perception threshold, the auditory spectrum analysis can be used to measure distortion in audio and speech transmission equipment. This method enables the use of actual speech (or other sounds) as measurement signals. The results correspond to subjective sound quality better than results obtained with traditional methods like total harmonic distortion measurement.

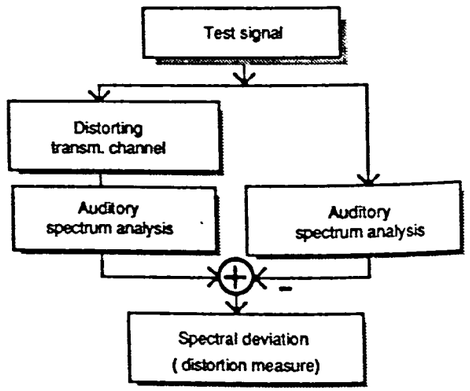

Fig. 6. Block diagram of auditory distortion measurement. By subtracting the auditory spectrum of the original test signal from the distorted signal we obtain the auditory spectral deviation, from which the distortion measure can be derived.

We have realized an auditory model based measuring system. The auditory model is implemented in a Texas Instruments TMS 32010 signal processor. An Apple Macintosh personal computer is used for system control and user interface, and a slightly modified Sony PCM-F1 pulse code modulator acts as the DA- and AD-converter. Figure 6 presents the nonlinear distortion measurement principle as a block diagram. Our system can handle the entire audio range (20Hz - 20 kHz) with a dynamic range of over 90 dB. The Posts and Telecommunications of Finland is testing the applicability of the method in telephone equipment measurements.

CONCLUSIONS

The auditory models have proven to be a useful means of determining perceived nonlinear distortion in speech. Already the relatively simple method of maximal spectral deviation is a good measure for the JND threshold (2-dB rule). More severe distortion levels need a more sophisticated measure. Practical applications of auditory methods are under development – possible areas are the evaluation of telephones and audio equipment as well as research systems for phonetic science.

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 Posts and Telecommunications of Finland. The implementation of the auditory model for the TMS 320 signal processor was realized by Vocom Ky.

REFERENCES

/1/ **Karjalainen M.,** Objective Measurements of Distortion in Speech Signal Channels by Computational Models of Speech Perception. Proc. of 11th ICA, Paris 1983.

/2/ **Karjalainen M.,** Sound Quality Measurements of Audio Systems Based on Models Of Auditory Perception. Proc. of IEEE ICASSP-84, San Diego 1984.

/3/ **Karjalainen M.,** A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of ICASSP-85, Tampa 1985.

/4/ **Karjalainen M., Helle S. & Altosaar T.,** Spectral Representations Based on an Auditory Model: Experiments and Applications. Proc. of Nordic Conference on Speech Processing, Trondheim, Norway 1986.

/5/ **Schröder M. et al.,** Objective Measure of Certain Speech Signal Deteriorations Based on Masking Properties of Human Auditory Perception. In the book: Frontiers of Speech Communication Research (ed. Lindblom & Öhman), Academic Press 1979.

/6/ **Zwicker E. & Feldtkeller R.,** Das Ohr als Nachrichten- empfänger. S. Hirzel Verlag, Stuttgart 1967.