AUDITORY MODELS FOR SPEECH PROCESSING

Matti Karjalainen

Helsinki University of Technology
Acoustics Laboratory, Otakaari 5 A
SF-02150 Espoo FINLAND

ABSTRACT

Computational modeling of the auditory periphery has become an integral part of hearing and speech research in recent years. This reflects the importance of computers and computational models as a research tool for experimenting flexibly in the domain of complex auditory phenomena. Both our general understanding and the fragmental knowledge of details known from hearing research can be reconstructed and tested in the form of functional models.

This paper approaches the auditory models primarily from another aspect: their applications within speech processing. Although there are almost no existing practical applications where systematic modeling has proven to be superior to traditional methods, the approach as such is seen as promising and necessary. Several approaches to auditory modeling are viewed in the paper with the main emphasis on functional and psychoacoustical properties, including some principles proposed for higher-level processing. Potential areas of applications are discussed with examples taken from our own studies.

INTRODUCTION

The theories, models and applications of speech perception are without any doubt lagging behind the level of knowledge in speech production. The main reasons for this are due to the complexity of the hearing system and the difficulties in experimenting with it, the lack of basic understanding of the higher-level processes and the problems in the implementation of experimental models to simulate the auditory system.

What is (or could be) auditory modeling?

The development of computers and software-based simulation makes it more and more attractive to experiment with principles of hearing. To some extent electronic and even mechanical models have been tried but the computer has become a superior tool for the task. The concept of *auditory model* is used normally to refer to a computational model of the peripheral hearing system. The *physiological* functions of the basilar membrane and other cochlear processes up to the neural levels are considered as the primary subject to be simulated by the models.

Another theoretical and experimental basis for auditory modeling comes from *psychoacoustics*, where the correspondence to the underlying physiology is not direct anymore. Perception thresholds and psychophysical "transfer functions" are more central to the approach. Psychoacoustical concepts like pitch and loudness that are related to the peripheral hearing are well developed and exact to a high degree. They have been verified by subjective listening experiments. More abstract properties exhibit fuzziness and random behaviour but can be included in computational models if they are stable enough.

The third approach to modeling is to hypothesize functional principles that possibly could be found in the hearing system. They may not be verified by direct physiological or psychological experiments. Most auditory models concerning higher levels of hearing will probably be of this type because the physiological basis is too complex and hard to access, and even the psychological approach does not test and validate the models. The borderline between auditory modeling and general information processing principles is not very clear at these levels.

Why auditory models?

Auditory modeling is attractive as a research tool because it presents the possibility to test hypotheses and experiment with new ideas flexibly in a proper context. The hearing system consists of complex subsystems that tend to be nonlinear and contain feedback loops, which makes it practically impossible to apply analytical modeling methods except to small subproblems. Computational models are useful also in conceptualizing the signal and information processing aspects in hearing apart from the underlying physiology.

The basic research of hearing is only one of the motivations for auditory modeling. Major challenges for future work are to be found in potential applications, especially in speech recognition. The human hearing system is the best processor to recognize speech messages; why not to try to duplicate it in technical form. The results so far show that this will not be done easily. In principle, however, this approach is promising and necessary, at least to gain a deeper insight into the many problems of speech recognition.

This paper reflects the point of view of the author towards auditory modeling. Physiological models are not seen as the only, and even not the major subject of research, when it comes to applications. Especially for speech processing we need flexible functional models based on signal processing and artificial intelligence. The rest of the paper will tie together a number of subproblems in auditory modeling along with some applications and experiments performed by our own research group.

MODELS OF THE PERIPHERAL HEARING SYSTEM

External and Middle Ear Models

Computational modeling of the hearing system begins from the acoustics of the external ear. Localization of sound and

the frequency sensitivity of the ear are greatly influenced by the acoustical details of the pinna and ear canals. The experimental studies and measurements [1] that have been made have led to successful results in computationally reconstructing an authentic sound environment sensation [2]. In combination with cochlear and neural modeling this could lead to better directional selectivity and sound localization [3] e.g. in speech recognition devices. Otherwise the role of the external and middle ear is a relative simple, almost linear filter as a part of a complex auditory model, contributing to the frequency sensitivity properties.

Cochlear Modeling

The physiology of the inner ear [4] - [8] is a main source of knowledge providing a concrete basis for present auditory modeling. This area of research is fairly rich in results and approaches, see [9] - [14], but no comprehensive and systematic cochlear models exist in computational form.

Modeling of the inner ear can be divided into several subproblems. The mechanics of the basilar membrane has received considerable attention since the studies of von Békésy [7]. The mathematically elegant principle of the nonhomogeneous transmission line must be enhanced with nonlinear processes and complex interactions with hair cells and neural processes [15] - [17]. Some recent results propose interesting computable models of interaction to improve the sensitivity and selectivity of the inner ear, Zwicker [18] and Lumer [19]. The acoustic emission, see Kemp [20], [21], should also be included into a full-scale model. The mechanical to electrical and neural transduction takes place in the hair cells and can be modeled in physiological detail or functionally including the random nature of single cell firings, e.g. Schröder and Hall [22], Lyon [23].

Towards Higher Levels

As the computational modeling domain moves towards the neural levels, more functional principles must be used instead of physiological facts. Mixing of physiology, psychology and highly hypothetical ideas in the form of computer programs is an important approach. Some models are more oriented towards the study of advanced computational implementations and applications than the hearing process as such, e.g. Lyon [23] - [25]. Reaching higher abstraction levels in relation to physiology by computer programs may prove to be valuable when studying the representation of speech and complex stimuli in the hearing system.

The problem of neural representations of speech signals has become a subject of remarkable research in recent years, [10], [26] - [28]. There are different explanations of how the spectral and temporal information is coded into the neural signal. The saturation effect of a single nerve fiber in sending amplitude data must be taken into account. The computational models try to capture the essentials of this process in different ways: e.g. the synchrony model proposed by Seneff [29], [30] is based on the firing synchronism principle found in the auditory nerve to avoid the spectral structure from being flattened. Seneff also made a generalization of the principle so that it can be applied to pitch detection.

Psychoacoustical Models

Some existing computational models find their theoretical and experimental basis primarily in psychoacoustics. The concepts of Bark scale (critical band scale), loudness and loudness density spectrum, masking curves, temporal time constants etc. [31] cannot be entirely reduced to the physiology of hearing. Examples of auditory models that are closely related to psychoacoustics are given e.g. by Schröder et al. [32] and Zwicker [33]. Both of them were developed with technical applications in mind.

Including Phonetic Aspects

Some research groups have worked by experimenting and modeling the perception of speech and its phonetically relevant features. Peripheral models of hearing tend to be nonspecific in relation to speech. How should the formants and formant transitions be processed by auditory models, and how should the phonetic features and categories be reflected in them? These problems are important from the point of view of applications, especially speech recognition.

Carlson, Granström et al. have discussed these questions and proposed several models for auditory speech analysis [34] - [36]. Klatt has a similar approach and he suggests a phonetic distance measure for comparison and classification of phonemes [37] - [39]. Principles and models relating auditory concepts to higher-level perception of speech are studied by Chistovich et al., [40] - [42]. Among them is the concept of center of gravity.

Auditory Modeling and Traditional Speech Processing

Many technically oriented systems for speech processing contain features that model the human hearing to some extent but some widely used methods do not exploit auditory features at all. It has been shown that linear predictive coding (LPC) in the original form is not optimal because it is based on a linearly weighted frequency scale. With Bark and loudness scaling its performance could in principle be considerably better [43] - [45]. Hermansky et al. have presented novel modifications of LPC analysis to include many important auditory features that can be applied to speech recognition [45] - [46].

AUDITORY SPECTRUM COMPUTATION

Most auditory models analyze audio signals by returning something we could call an *auditory spectrum*. This is natural because the inner ear (basilar membrane, hair cells) also forms some kind of a spectrum analyzer, even if it is different from technical devices and algorithms for the Fourier transform.

The models for auditory spectrum analysis can be divided into two classes according to the processing of temporal dynamics. If we are not especially interested in the detailed time constants of the resulting (short-time) spectrum representation, we can first apply the Fourier transform and then warp the frequency scale to the Bark scale. Otherwise, we need a transmission-line or filter-bank type analyzer to allow more freedom in the design of temporal features.

Auditory Spectrum by Fourier Transform

The human auditory system may be seen as a spectrum analyzer that differs from Fourier analyzers in many ways. The most important differences are:

1. spectral emphasis by the inverse of the equal loudness curves,
2. use of the Bark scale (critical band scale) instead of the Hertz frequency scale,
3. frequency domain resolution of about one Bark,
4. masking effect in the frequency domain and spreading of the spectral components, and
5. time domain dynamics: temporal integration and masking effect in the time domain (forward and backward masking).

All these properties are known from psychoacoustics [31] but there have not been very many attempts to apply them in practical applications. Schröder & al. have used a computational model when evaluating signal-to-noise ratios in speech transmission [32]. We adopted their mathematical formulation with minor modifications as follows:

* Computation of the Fourier transform with a 35 ms Hamming window.
* Emphasis of the spectrum by an approximation of the frequency sensitivity curve of the ear (inverse of the equal loudness curve).
* Transformation of frequency **f** fo Bark variable **x** by:

  $\mathbf{x} = 7\ \mathrm{arsinh}(\mathbf{f}/650\mathrm{Hz})$.
* So called "excitation function" $\mathbf{E(x)}$ is found by smoothing the Bark-scaled pre-emphasized power spectrum $\mathbf{S(x)}$ with a "spreading function" $\mathbf{B(x)}$:

  $\mathbf{E(x)} = \mathbf{S(x)} * \mathbf{B(x)}$, (* indicates convolution)

  where $\mathbf{B(x)}$ in our model is a piecewise linear approximation of the Schröder et al. spreading function

  $$10\log(\mathbf{B(x)}) = 15.81+7.5(x+0.474) - 17.5\sqrt{1+(x+0.474)^2}$$

  by linear slopes (+ 25 dB/Bark, -10 dB/Bark) and power series approximation for the top of the curve (see Fig. 1b).
* dB-scaled $\mathbf{E(x)}$ is the final auditory spectrum used in the study. Two examples of such spectra of simple signals are shown in Fig.1. The spectrum of an impulse (1a) has a form which is similar to the frequency sensitivity of the ear. The auditory spectrum of a sine wave (1b) gives the masking curve and an approximate form of the spreading function $\mathbf{B(x)}$.

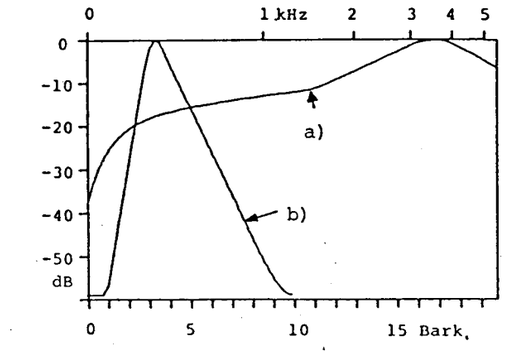


Fig. 1. Auditory spectra of simple test signals: (a) impulse spectrum and (b) sine wave spectrum.

Some examples of auditory spectra with corresponding Fourier spectra for speech sounds are plotted in Figures 2 and 3. The Finnish vowel /a/ shows clearly how the harmonic structure in a Fourier spectrum is smoothed out but the main formants are retained in an auditory spectrum (Fig. 2). In the fricative /s/ the random variation of the Fourier spectrum is also smoothed and the "fricative formant" shows up in the form of a normal vowel formant (Fig. 3).

Auditory Spectrum by Filter-bank Modeling

It was found to be difficult to include proper temporal dynamics when using the Fourier transform techniques. 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 in the biological system 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

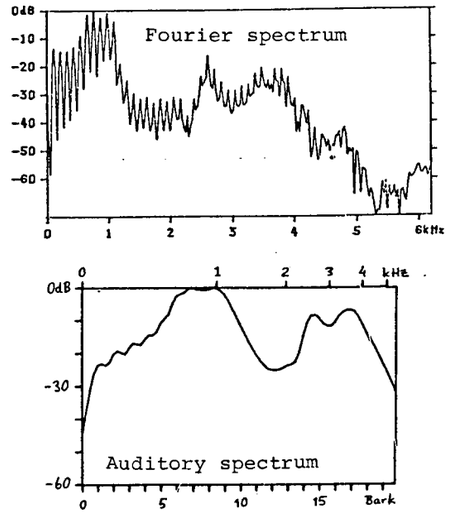


Fig. 2. Fourier spectrum and auditory spectrum for a Finnish vowel /a/.

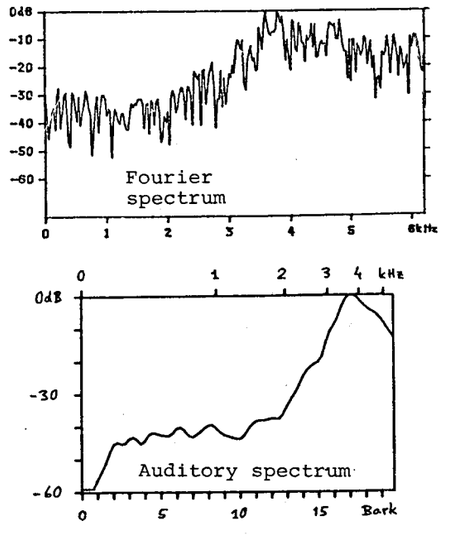


Fig. 3. Fourier spectrum and auditory spectrum for a Finnish fricative /s/ (in context /assa/).

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 a 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.4).

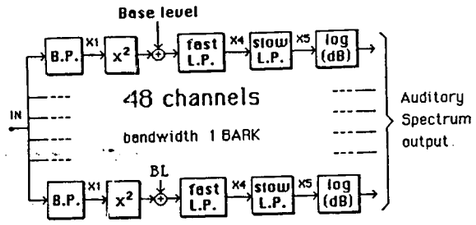


Fig. 4. A 43-channel filter-bank model for auditory spectrum computation. B.P.=bandpass, L.P.=lowpass filter, x^2= square-law detection, LOG = dB-scaling.

Bandpass filter bank

Bandpass filters with 0.5 Bark spacing and about 1.3 Bark bandwidth give the desired frequency selectivity to the model. Each bandpass is a 256th order FIR-filter, carefully designed to have a frequency response which is the mirror image of the spreading function B(x) given by Schröder et al.

This filter bank design gives a good approximation of the desired masking properties in the frequency domain. Computation of the filter bank was implemented as a matrix multiplication in an array processor (Floating Point Systems FPS 100). Even an array processor could not run it in real time. By a proper IIR-filter design the speed of the computation could be more than 10 times faster but accurate design of these 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 be proportional to the inverse of the equal loudness curve (60 dB-level).

Rectification

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

Filters of temporal integration and forward masking

The remaining two filters are for smoothing the outputs of the rectified bandpass filters. The faster one is a first-order lowpass with a time constant of about 3 ms. The second one is more important. Its purpose is to implement many effects: temporal integration and pre- and postmasking effects.

Temporal integration is realized by linear first-order lowpass filtering (time constant of about 100 ms) applied to the output of the square-law rectifier. Premasking is not a very important and critical phenomenon. No additional modeling was necessary to match it well enough.

Postmasking was found to be more difficult to implement in sufficient detail. A linear lowpass filter with a 100 ms time constant yielded postmasking effect that was many times too long. We used a nonlinear (logarithmically linear) behaviour of the filter for masking conditions (X4 < X5). The form of the temporal masking pattern is now close to the actual one found in psychoacoustical studies [31] but a delay of about 10 ms present in the real masking effect is lacking in the model. The overall response of the slow nonlinear lowpass can be stated now:

$$X5(n) = K1*X4(n) + (1-K1)*X5(n-1), \quad \text{if } X4 \geq X5,$$
$$X5(n) = X5(n-1)*\exp(K2*\log(X4(n)/X5(n-1))), \quad \text{if } X4 < X5,$$

where X4 and X5 are the input and output of the filter, K1 and K2 the filter coefficients, and n the discrete time variable. A good value for K1 was found to be 0.0005, and 0.0007 for K2 when the sampling frequency is 20 kHz.

Auditory short-time spectra computed by the model can be displayed in many forms: spectral series, spectrograms, etc. Examples of these are given later in this paper.

AUDITORY FORMANTS AND FORMANT SPECTRA

The auditory spectrum, as was analyzed by the models above, is not a speech-specific representation. Attempts to utilize it or other similar preprocessing methods in speech recognition have shown only moderate results, see e.g. [47]. It is obvious that some further processing of auditory spectra is needed to exhibit speech-specific features and more "phonetic-like" auditory representations of speech.

Some hints and guidelines can be found e.g. from the studies of Klatt [37] - [39], paying special attention to the formant peak regions in the auditory spectra. Global properties such as the slope of the spectrum have only a minor effect on the phonetic quality of a sound. Klatt suggested the use of *phonetic distance measures* [39] based on local properties of the formants in auditory spectra. Another concept that is closely related to auditory formants is the *center of gravity* by Chistovich et al. [40].

Emphasizing and Sharpening the Auditory Formants

Possible conclusions that may be drawn from the results of using short-time auditory spectra in speech recognition could be that the perceptually important formant peaks are excessively smoothed and the local properties of the formants are not prominent enough. Is it possible to compensate for these effects? There are neurophysiological principles that are candidates for the spectral sharpening effect: *lateral inhibition* is one such candidate. A strong excitation at a certain place along the basilar membrane tends to suppress the neighbouring channels.

The formant features can be sharpened or emphasized computationally in many ways. We can perform highpass or bandpass filtering of the auditory spectrum in the Bark domain to supress the global forms (e.g. spectral tilting) and to emphasize the local formant peaks. This can be realized by convolving the loudness-scaled auditory spectrum by a proper spatial (Bark domain) bandpass filter impulse response. Figures 5 and 6 show original auditory spectra for a vowel /ä/ and fricative /s/ along with the resulting *auditory formant spectra*, as we call them.

In both cases the auditory formant spectrum exhibits clearly the formant peaks so that the global spectrum structure does not have a major dominance. Serial displays of auditory spectrum and auditory formant spectrum are shown in Fig. 7 for the vowel combinations /aiai/.

The Concept of Auditory Formant

The perceptual relevance of the peaks in auditory spectra implies the usefulness of the concept *auditory formant*. It must be recognized as different from the acoustic and articulatory aspects of formants even if there is a clear correspondence between them. A useful characterization of the auditory formant is to state it as any *peak or relatively localized high-loudness region*, a kind of landmark in an auditory spectrum.

Several studies have been done on the perceptual behavior of auditory formants and formant groups, see e.g. Chistovich

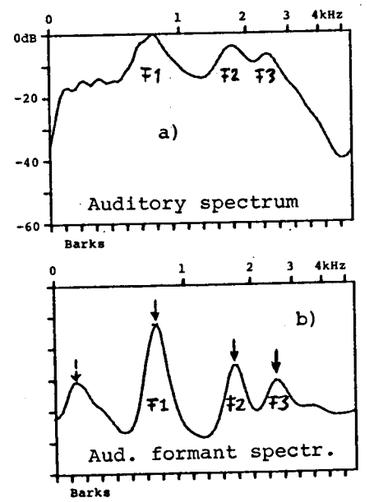


Fig. 5. Finnish vowel /ä/: (a) auditory spectrum, (b) derived auditory formant spectrum

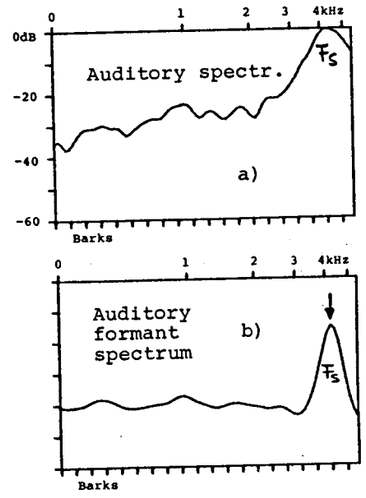


Fig. 6. Fricative /s/: (a) auditory spectrum, (b) auditory formant spectrum.

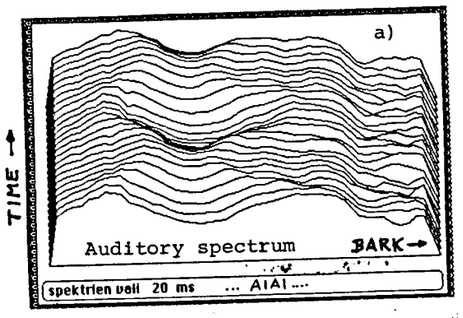
[40]. The integration of closely spaced formants, the concept of center of gravity, etc. are principles that should also be implemented in computational models.

Is it possible to extract auditory formants and to describe them as discrete units? The auditory formant spectrum above is a good data source for this extraction. In Figures 5, 6 and 7 a spatial bandpass filtering with a one Bark resolution was applied to give a proper pre-emphasis to the spectrum. A peak-picking algorithm easily finds the formants as the local maxima of the curves. Fig. 8 illustrates how the short-time auditory spectrum of the utterance /kuusi/ (Fig. 8(a) dB-scaled, Fig. 8(b) loudness-scaled) is transformed to an auditory formant spectrum, Fig. 8(c). The formant peaks are finally plotted and shown in a spectrogram-like display (Fig. 8(e)).

The one Bark resolution does not always work. Closely spaced formants may give a better response e.g. to a 2 Bark resolution filtering, Fig. 8(d). (See also the 2 Bark formant spectrogram in Fig. 8(f)). This finding shows the need for different resolutions in different contexts. According to Chistovich the auditory system can integrate neighbouring formants up to a distance of 3.5 Barks [40]. This corresponds to about a 2-3 Bark resolution in our bandpass filtering.

Since there is no single optimal resolution a better strategy is to use multiple resolutions in parallel. This means that the formant peaks are picked to form several formant lists. Later on it is possible to utilize the data that seems to be the most reliable based on the context.

We can also visualize the multiple resolution auditory formant data in a spectrogram form by using different gray levels or colors for the formant trajectories of different resolutions. Fig. 9 shows the mixed result of 1 and 2 Bark auditory formant analyses for the word /kuusi/. The general principle of multiple resolution analysis is discussed below.



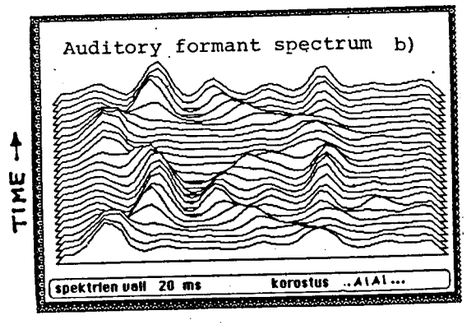


Fig. 7. Spectral series for the vowel combination /aiai/: (a) auditory spectrum, (b) auditory formant spectrum.

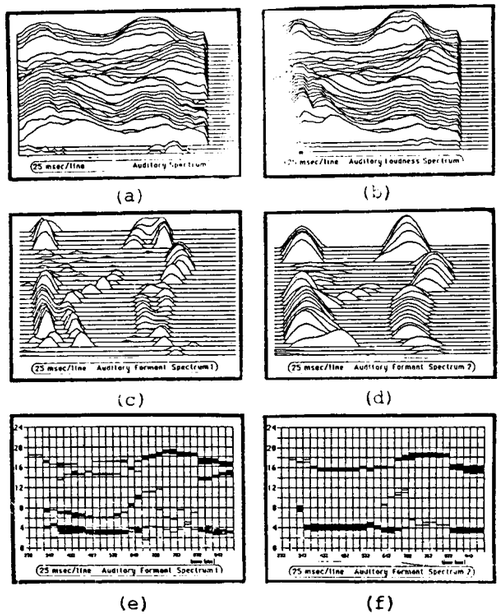


Fig. 8. Auditory spectrum presentations for the word /kuusi/: (a) original dB-scaled spectrum, (b) loudness-scaled spectrum, (c) auditory formant spectrum with 1 Bark resolution, (d) auditory formant spectrum with 2 Bark resolution, (e) formant spectrogram of 1 Bark resolution and (f) formant spectrogram of 2 Bark resolution.

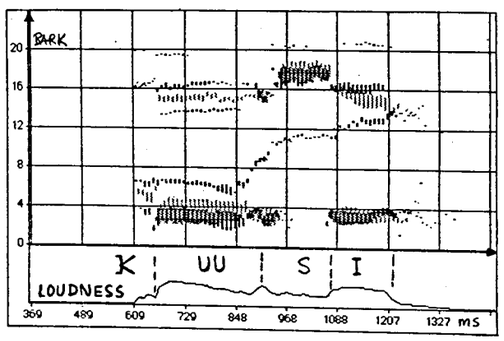


Fig. 9. An auditory formant spectrogram with two overlayed displays for resolutions of 1 Bark (black) and 2 Barks (gray).

Local vs. Global Features of Auditory Spectra

Both the auditory formant spectrum and the detection of discrete formant parameters emphasize the prominent *local* features in auditory spectra. We could also analyze and characterize the *global* properties. An average spectral slope and the center of gravity over the whole audio range are good examples of such global attributes. From the point of view of speech perception the absolute values of these parameters are not as important as the relative changes they exhibit. For instance, large static spectral tilting is allowed with only minor change in the phonetic quality of a vowel [39].

Spectral tilt or center of gravity can also be computed over any limited range in the Bark domain. An interesting special case is to analyze closely the effective movement ranges of the lowest formants, e.g. over 2 to 6 Barks for F1. The values of these parameters describe the average slope or the approximate position of the formant in the defined range. Fig. 10 shows the results of such an analysis for the full audio range, F1, F2, and F3&F4 range, along with the loudness function and a two-resolution auditory formant spectrogram for the Finnish word /viisi/.

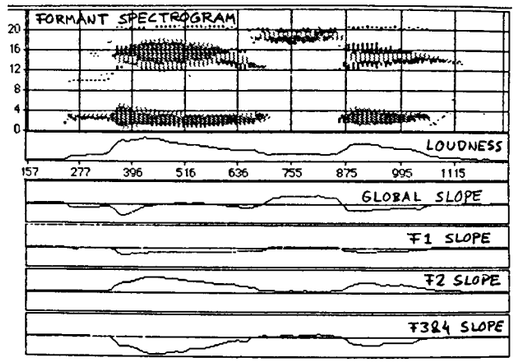


Fig. 10. Several different analyses for the Finnish word /viisi/: (a) formant spectrogram, (b) loudness function, (c) global slope and the local spectral slopes for the formant ranges F1, F2 and F3&F4.

Analysis of Formant Movements

It is known that the hearing system is especially sensitive to *changes* in sound. It is also known that the auditory system contains specialized analyzers for frequency sweeps and formant movements [48]. Such detectors may have an important role in the perception of speech signals and they should be included in computational auditory models. To some degree, the derivatives of the slope funtions above represent this kind of information. The output of an advanced detector could be a series of "formant movement events" similar to the time structure analysis method in the following section.

TIME STRUCTURES AND MULTIPLE RESOLUTION ANALYSIS

Time is one of the most difficult and least understood dimensions in speech signal analysis. The rhythm and timing in real speech varies widely according to the context and therefore straightforward methods of segmentation do not work reliably. The transformation from continuous-time representations to discrete units in time should be studied more carefully so that time resolution is seen as one parametric scale.

Let us consider a set of parametric or feature functions as analyzed from a speech signal. Fig. 11 shows the total loudness function (sum of all the filter-bank channels), nonstationarity (relative change in short-time auditory spectrum) and the global spectral slope as a function of time, along with the formant spectrogram for the utterance /yksi/. What is a flexible and reliable way to do "segmentation" based e.g. on the loudness function?

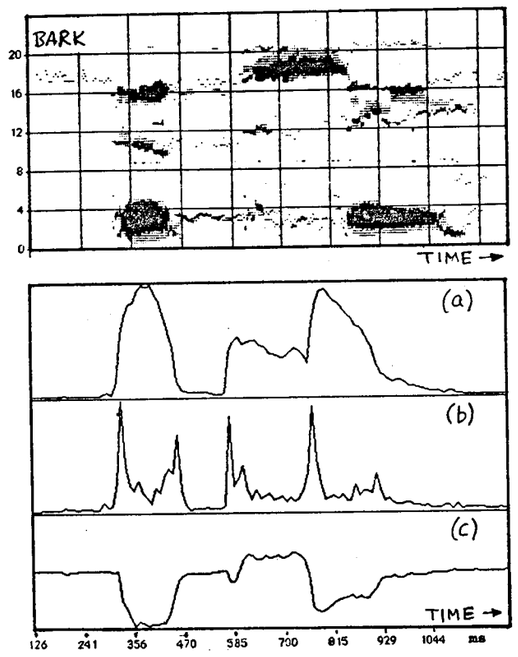


Fig. 11. Auditory formant spectrogram and multiple temporal feature functions for the Finnish word /yksi/: (a) total loudness function, (b) nonstationarity function and (c) global spectral slope.

If proper bandpass filtering is applied to the loudness function the "events" that match best to the impulse response of the filter are emphasized. The same principle is used as was for the filtering of formants in the frequency domain. An example of a useful impulse response for a *resolution filter* is shown in Fig. 12.

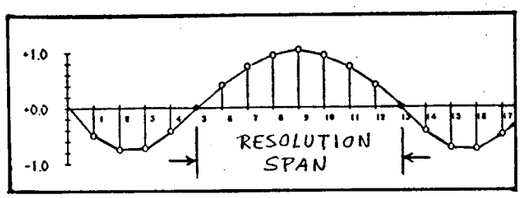


Fig. 12 An example of an impulse response for a resolution filter.

Any single filter emphasizes the events of its corresponding time resolution the most. The extrema (maxima and minima) of the response are easily picked up as prominent *events* in the time structure of the signal. To be more flexible *a set* of resolution filters can be applied to the loudness function in parallel. In this *multiple resolution analysis* each resolution filter channel produces a list of potential *loudness events*. Other parametric functions like the global spectral slope create their own event lists and list structures. The idea of multiple resolution analysis has some resemblance to the scale-space filtering proposed by Witkin [49].

As an example of using the principle, nine filters with time resolutions ranging from 10 to 320 ms were applied to the loudness function of Fig. 11. The convolution results are plotted in Fig. 13. A continuous scale of resolutions is in principle the ideal case but a series of filters with resolution ratios of about $1:\sqrt{2}$ was found to be practical. The method of multiple resolution analysis leads to an excessive amount of computation in comparison to single resolution (single window, frame, etc.). This is the cost to be paid for more flexibility. In highly parallel neural networks such computational redundancy is easily achieved but with present digital signal processing hardware it is a problem.

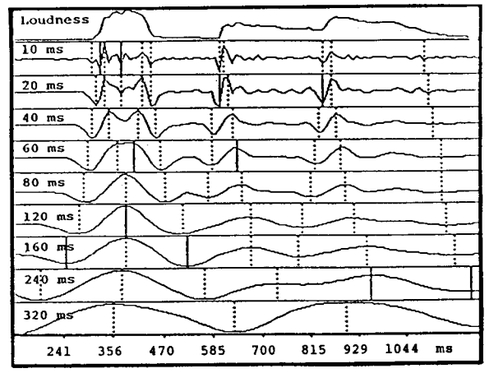


Fig. 13. Resulting curves from the multiple resolution analysis of the loudness function. Vertical lines indicate potential event positions.

Event-based Approach to Auditory Speech Analysis

Each feature to be used in a speech analysis system and each resolution of feature produces a corresponding list of events, containing much redundant information. In Fig. 13 e.g. the maxima of the neighbouring channels are closely interrelated. By a proper method we can discard many of the peaks as masked by more prominent neighbouring peaks. The potential events can be organized into the form of complex event list structures and processed further by rule-based and other artificial intelligence methods. This approach is discussed in more detail by Altosaar and Karjalainen in [50].

The event-based approach may be useful at several levels of auditory modeling. We could apply it at the auditory nerve level by picking up the most prominent peaks from the multiple resolution filtering of a single critical band channel in the model of Fig. 4, point X4. Here the range of interesting time resolutions is within a typical pitch period of speech. At the output level of the model (point X5) the resolution range corresponds to typi- cal speech segments of 10 to 300 ms. The prosodic features reveal still longer event objects. By parallel processing and concurrent programming techniquesa realization of this approach could be undertaken.

TEMPORAL FINE STRUCTURE ANALYSIS OF SPEECH SIGNALS

It is still a common belief that the temporal fine structure of speech signals within a pitch period (below 10 ms span) and the phase properties are irrelevant to speech perception. This belief is mostly due to the interpretations of the studies by Helmholtz [6]. Even if this is true as the first approximation several findings suggest a more important role for these details. For example, the success of the so-called multi-pulse LPC [51] in comparison to the simple impulse source and all-pole modeling shows how much the detailed time structure may affect the sound quality. In our Finnish speech synthesis studies we found that a careful zero-phasing (i.e. setting the phase of all harmonics to zero) of a natural utterance /illi/ changed it to be heard sometimes rather like /inni/.

Concepts Related to Temporal Fine Structure

The auditory spectrum output in our filter-bank type auditory model (Fig. 4) does not represent the temporal fine structure of speech signals at all. It is therefore possible to obtain identical auditory spectra for a voiced and an unvoiced sound from this model. The *degree of voicing* and *pitch for voiced sounds* are certainly concepts with close relationship to the temporal fine structure. This information should be analyzed from the fast response of the auditory filter bank, i.e. the point X4 in Fig.4, corresponding to the first neural stages of the hearing system.

Lyon [23] has presented computational auditory models to analyze the periodicity properties of speech signals by following the principles proposed by Licklider [52]. The models rely on correlation and coincidence functions from neural firings.

The *phase properties* of speech signals are difficult both to analyze and to interpret meaningfully. A traditional way of looking at phase has been by the Fourier transform of the signal. The phase in this sense is, however, very sensitive to noise, reverberation and other disturbances in speech. If any concept of *auditory phase* can be formed, it must be defined in a totally different way.

A step towards auditory phase could be to interpret it from the point of view of the modulation envelope in different critical band channels (Fig. 4). The fast response corresponding to the output of the hair cells in the hearing system carry this information. In the case of voiced speech sound these outputs exhibit the fundamental frequency of the speech. The relative phase shift of these pitch modulations between the neighbouring channels could be useful as the auditory phase function. The same data could be expressed also in the form of auditory group delay.

Sound Separation

The role of auditory analysis of temporal fine structure takes on a new appearance when we set the goal of modeling to be the phenomenon of *sound separation*. People can easily follow a single speaker in a high noise environment (e.g. the cocktail party effect). To make machines recognize speech at or below the level of background noise, a successful modeling of sound separation is needed.

This is a fairly new subject for serious computational modeling. Weintraub [53] has made remarkable contributions by studying some peripheral processes of auditory analysis in sound separation. He has emphasized the need for a multilevel strategy to solve this problem.

Some special cases of sound separation can be experimented with easily. We used the fast response outputs of our filter-bank model (Fig. 4) and computed the cepstrum for each of them. If there was a voiced speech sound that dominated a Bark channel, the corresponding pitch period was easily found as a dominant peak in its cepstrum. By summing the cepstra of all the channels the pitch periods of the individual sounds in a mixed signal were possible to be separated. After this it is feasible to estimate the spectrum of each speech sound in those frequency areas where the signal is not totally masked by other sounds. Such methods tend to be computationally so excessively heavy that it prevents their use in any real-time applications now or in the near future.

APPLICATIONS OF AUDITORY MODELS

There are very few if any practical applications of computational auditory models. This is quite natural because the subject of research is complex and relatively new. Some preliminary results of using them e.g. in speech recognition have shown poor or at best only marginal results. This has been discouraging but it has not stopped either basic research or application-oriented work in the field. If the human hearing system performs as the best as a speech recognizer, why couldn't a good model of it be the best speech recognizer as well.

Speech recognition has been an explicit motivation in the development of several practically oriented auditory models [33], [36], [46], [54] and implicitly in many other cases. One of the earliest works was documented by Zagoruiko and Lebedjev [55]. Only recently has there been signs of obtaining better results than with traditional methodology, see e.g. Cohen [54]. It is premature to draw conclusions about the real status of auditory models in speech recognition. It seems to be evident that no fundamental problems can be solved without models covering many or even all essential levels from acoustic signals to liguistic processing. One area of preprocessing where auditory models could help is in high background noise conditions.

Speech analysis in phonetics and basic speech research can gain from applying computational models of hearing. For example, auditory spectra give a picture of the important spectral features in speech signals from the point of view of perception. Carlson and Granström [35] and Klatt [38] among others have discussed the development of an auditory spectrograph. Traditionally the articulatory and acoustic aspects have been more dominant in speech analysis because of better instrumentation and tools for experimental work. With modern signal processors, personal computers and new programming techniques it was possible to develop a speech research workstation called ISA [56] that utilizes many of the representations described in this paper.

Speech synthesis is a process where auditory models can not be used directly. In the development of speech synthesis, however, we have been succesful in applying them. The microphonemic method by Lukaszewicz and Karjalainen [58] showed how the auditory formant spectrogram exhibited just the information needed for extracting pitch period prototypes from real speech. The use of the Bark scale was essential.

In *speech coding* the auditory models could be applied in the analysis phase, as a design tool, and in performance analysis, see Schröder et al. [32]. It is shown that the compactness of LPC-analysis could be impoved if auditory features could be integrated into the coding [43] - [45]. Unfortunately the auditory analysis tends not to preserve the properties that are needed for easy resynthesis of speech.

Measurement of sound quality (especially nonlinear distortion) in speech transmission was studied by us to find a better correlation between subjective and objective measures, Karjalainen [58], [59] and Helle and Karjalainen [60]. We have shown that an auditory spectrum distance of 2 dB corresponds to the just noticeable level of nonlinear distortion in speech signals. A prototype of a microprocessor-based distortion measurement system was also developed.

Technical audiology and phoniatrics are other areas of potential applications. In the delopment of hearing aids and cochlear implants it is evident that some kind of auditory models will be used in the future. In phoniatrics the properties of pathological voice can be analyzed based not only on articulatory and acoustic measurents but also on advanced auditory models [56].

CONCLUSION

This paper has presented an overview of auditory modeling from the point of view of speech processing research and applications. Both physiological, psychoacoustical and higher-level functional models are needed to gain a deeper understanding of the underlying phenomena and to be able to apply this knowledge to speech technology. Auditory modeling is a difficult area of research where progress is not always rapid. It also takes time to transfer the results into practice. Within the last ten years the interdisciplinary studies of computational auditory modeling have shown trends to expand and grow. Without any doubt this tendency will continue as the computational capabilities of modeling rapidly develop.

ACKNOWLEDGEMENTS

The paper presents some results from a research project on the modeling of auditory and speech communication that was financed by the Academy of Finland. The Posts and Telecommunications of Finland supported the development of some of the applications. I am grateful to all my co-workers who have made a contribution to the results.

REFERENCES

[1] **Blauert J.**, Spatial Hearing. The MIT Press, Cambridge 1983.

[2] **Pösselt C., Schröter J., Opitz M., Divenyi P.L.**, Generation of Binaural Signals for Research and Home Entertainment. Proc. of ICA-86, pp. B1-6, Toronto 1986.

[3] **Lyon R.**, A Computational Model of Binaural Localization and Separation. Proc. IEEE ICASSP-83, Boston 1983.

[4] **Pickles J. O.**, An Introduction to the Physiology of Hearing. Academic Press, London, 1982.

[5] **Möller A.R.**, Auditory Physiology. Academic Press, New York 1983.

[6] **Helmholtz H.**, On the Sensations of Tone. Dover Publications 1954.

[7] **von Békésy G.**, Experiments in Hearing. McGraw-Hill, New York, 1960.

[8] **Zwislocki J.J.**, Five Decades of Research on Cochlear Mechanics. J. Acoust. Soc.of Am, 67 (1980), pp. 1679-1685.

[9] **Allen J.B.**, Cochlear Modelling. IEEE ASSP Magazine, Vol.2 number 1, Jan 1985.

[10] **Carlson R., Granström B.** (ed.), The Representation of Speech in the Peripheral Auditory System. Elsevier Biomedical Press, Amsterdam, 1982.

[11] **Schröder M.R.**, Models of Hearing. Proc. of IEEE 63 (1975), pp. 1332-1350.

[12] **Allen J.B.**, Cochlear Micromechanics -- A Physical Model of Transduction. J. Acoust. Soc. Am. 68 (1980), pp. 1660 - 1670.

[13] **Delgutte B.**, Some Correlates of Phonetic Distinctions at the Level of the Auditory Nerve. In ref. 10.

[14] **Dolmazon J.**, Representation of Speech-Like Sounds in the Peripheral Auditory System in the Light of a Model. In ref. 10.

[15] **Neely S.T., Kim D.O.**, An Active Cochlea Model Showing Sharp Tuning and High Sensitivity. Hearing Research 9 (1983), pp. 123-130.

[16] **Khanna S. H., Leonard D.G.B.**, Basilar Membrane Tuning in the Cat Cochlea. Science, vol. 215, pp. 305-306, Jan. 1982.

[17] **Davis H.**, An Active Process in Cochlear Mechanics. Hearing Research 9 (1983), pp. 79-90.

[18] **Zwicker E.**, Peripheral Preprocessing in Hearing and Psychoacoustics as Guidelines for Speech Recognition. Proc. of Montreal Symposium on Speech Recognition, Montreal, 1986, pp. 1-4.

[19] **Lumer G.**, Computer Model of Cochlear Preprocessing (Steady State Condition) I. Basics and Results for one Sinusoidal Input Signal. Acustica Vol. 62 (1987), pp. 282-290.

[20] **Kemp D.T.**, Stimulated Acoustic Emission from within the Human Auditory System. J. Acoust. Soc. Am. 64 (1978), pp. 1386 - 1391.

[21] **Kemp D.T., Anderson S.D.**, Proc. of the Internat. Symposium on Nonlinear and Active Mechanical Processes in the Cochlea. Hearing Science, 2, no. 3 and 4, 1980.

[22] **Schröder M.R., Hall J.L.**, Model for mechanical to Neural Transduction in the Auditory Receptor. J. Acoust. Soc. Am. 55 (1974), pp. 1055-1060.

[23] **Lyon R.**, Computational Models of Neural Auditory Processing. Proc. of IEEE ICASSP-84, Tampa 1984.

[24] **Lyon R.**, A Computational Model of Filtering, Detection and Compression in the Cochlea. Proc. of IEEE ICASSP-82, Paris 1982.

[25] **Lyon R.**, Experiments with a Computational Model of a Cochlea. Proc. IEEE ICASSP-86, Tokyo 1986.

[26] **Evans E.F.**, Representation of Complex Sounds at the Cochlear Nerve and Cochlear Nucleus Levels. In ref. 10, pp. 27-42.

[27] **Möller A.R.**, Neurophysiological Basis for Perception of Complex Sounds. In ref. 10, pp. 43-60.

[28] **Sachs M.B., Young E.D., Miller M.I.**, Encoding of Speech Features in the Auditory Nerve. In ref. 10, pp. 115-130.

[29] **Seneff S.**, Pitch and Spectral Estimation of Speech Based on Auditory Synchrony Model. Proc. IEEE ICASSP-84, pp. 36.2.1-4, San Diego 1984.

[30] **Seneff S.**, A Computational Model for the Peripheral Auditory System: Application to Speech Recognition Research. Proc. of IEEE ICASSP-86, pp. 37.8.1-4, Tokyo 1986.

[31] **Zwicker E., Feldtkeller R.**, Das Ohr als Nachrichtenempfänger. S. Hirzel Verlag, Stuttgart, 1967.

[32] **Schröder M.R., Atal B.S., Hall J.L.**, Objective Measures of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception. In: Frontiers of Speech Communication Research, pp. 217-229, (ed. Lindblom & Öhman). Academic Press 1979.

[33] **Zwicker E, Terhardt E., Paulus E.**, Automatic Speech Recognition Using Psychoacoustic Models. J. Acoust. Soc. Am. 65 (1979), pp. 487-498.

[34] **Carlson R., Granström.**, Model Predictions of Vowel Dissimilarity. STL-QPSR 3-4/1979, pp. 84-104.

[35] **Carlson R., Granström B.**, Towards an Auditory Spectrograph. In ref. 10, pp. 109-114.

[36] **Blomberg M., Carlson R., Elenius K., Granström B.**, Auditory Models and Isolated Word

Recognition. STL-QPSRL 4/1983, 1-15.
[37] **Klatt D.H.**, Speech Perception: A Model of Acoustic - Phonetic Analysis and Lexical Access. J. Phonetics 7 (1979), pp. 279 - 312.
[38] **Klatt D.H.**, Speech Processing Strategies Based on Auditory Models. In ref. 10, pp. 181-196.
[39] **Klatt D.H.**, Prediction of Perceived Phonetic Distance from Critical-Band Spectra: A First Step. Proc. of IEEE ICASSP-82, pp. 1278-1281, Paris 1982.
[40] **Chistovich L.A, Sheikin R.L., Lublinskaja V.V.**, "Centres of gravity" and Spectral Peaks as the Determinants of Vowel Quality. In: Frontiers of Speech Comm. Research, pp. 143-157 (ed. Lindblom & Öhman). Academic Press 1979.
[41] **Chistowicz L.A., Lublinskaya V.V., Malinnikova E.A., Ogorodnikova E.A., Stoljarova E.I., Zhukov S.JA.**, Temporal Processing of Peripheral Auditory Patterns of Speech. In ref. 10, pp. 165-180.
[42] **Chistowicz L.A.**, Central Auditory Processing of Peripheral Vowel Spectra. J. Acoust. Soc. Am. 77 (3), March 1985, pp. 789-805.
[43] **Makhoul J., Cosell L.**, LPCW: an LPC vocoder with linear spectral warping. Proc. IEEE ICASSP-76.
[44] **Koljonen J., Karjalainen M.**, Use of Computational Psychoacoustical Models in Speech Processing: Coding and Objective Performance Evaluation. Proc. of IEEE ICASSP-84, San Diego 1984.
[45] **Hermansky H., Fujisaki H., Sato Y.**, Spectral Envelope Sampling and Interpolation in Linear Predictive Analysis of Speech. Proc. of IEEE ICASSP-84, San Diego 1984.
[46] **Hermansky H., Tsuga K., Makino S., Wakita H.**, Perceptually Based Processing in Automatic Speech Recognition. Proc. of IEEE ICASSP-86, pp. 37.5.1-4, Tokyo 1986.
[47] **Blomberg M., Carlson R., Elenius K., Granström R.**, Experiments with Auditory Models in Speech Recognition. In ref. 10, pp. 197-201.
[48] **Lacerda F., Moreira H.O.**, How Does the Peripheral Auditory System Represent Formant Transitions? / Psychophysical Approach. In ref. 10.
[49] **Witkin A.P.**, Scale-Space Filtering: A New Approach to Multi-Scale Description. Proc. of IEEE ICASSP-84, pp 39A.1.1-4, San Diego 1984.
[50] **Altosaar T., Karjalainen M.**, An Event-Based Approach to Auditory Modeling of Speech Perception. In this proceedings.
[51] **Atal B.S., Remde R.**, A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates. Proc. of IEEE ICASSP-82, Paris 1982.
[52] **Licklider J.C.R.**, A Duplex Theory of Pitch Perception. Experimentia, 7 (1951), pp. 128-133.
[53] **Weintraub M.**, A Theory and Computational Model of Auditory Monaural Sound Separation. Doctoral Thesis, Stanford University, August 1985.
[54] **Cohen J.R.**, Application of an Adaptive Auditory Model to Speech Recognition. In Proceedings of the Montreal Symposium on Speech Recognition. McGill University, Montreal 1986.
[55] **Zagoruiko N.G., Lebedjev V.G.**, Models of Speech Signal Analysis Taking into Account the Effect of Masking. Acoustica 31 (1975), pp. 346-348.
[56] ISA, Intelligent Speech Analyser. Instruction manual, Vocal Systems, Finland.
[57] **Lukaszewicz K., Karjalainen M.**, Microphonemics - High-Quality Speech Synthesis by Waveform Concatenation. In this proceedings.
[58] **Karjalainen M.**, Sound Quality Measurements of Audio Systems Based on Models of Auditory Perception. Proc. of IEEE ICASSP-84, San Diego 1984.
[59] **Karjalainen M.**, A New Auditory Model for the Evaluation of Sound Quality of Audio Systems. Proc. of IEEE ICASSP-85, Tampa 1985.
[60] **Helle S., Karjalainen M.**, Perception and Measurement of Distortion in Speech Signals - An Auditory Modelling Approach. In this proceedings.