EECS225d Audio Signal Processing in Humans and Machines **Spring 1995**

Lecture 40: April 28, 1995

Lecturer: Professor Oded Ghitza *Scribe: Angela K. Wang*

40.1 Auditory Models: Part 2

In the second lecture given by Dr. Ghitza, three main topics are covered:

- The EIH Model, which simulates and processes auditory nerve firings
- Evaluation of the EIH Model in speech recognition and speech coding
- Interpretation of the results of the evaluation

The underlying goal is to achieve human performance in tasks related to speech perception. Therefore, models which mimic the human auditory system are of great interest.

The EIH (ensemble interval histogram) model uses detailed modeling of the auditory periphery to simulate nerve firings. To process the firings, the EIH estimates the coherence of activity across the nerve fibers by counting the number of threshold crossings associated with the characteristic frequency of each cochlear channel. This differs markedly from the traditional speech representation methods (such as Fourier) which use power analysis.

The EIH and Fourier methods were compared against each other and against human performance using suitable diagnostic tests. In the task of discrimination between word pairs, it was found that EIH performs better than Fourier, but not nearly as well as humans. In the task of predicting human assessment of coded speech, EIH does well for 'high-quality' schemes. However, in CELP low-bit coding which smears the error over frequency, humans judge the speech quality to be acceptable, but EIH judges the speech quality to be bad.

An analysis of the experimental results suggests ways to improve the auditory models. Current auditory models use 'snap-shot' (local in time and frequency) representations of the speech signal and use Euclidean distance to compare two speech segments. It is hypothesized that humans use a representation which integrates information over time *intervals* (50-100 msec in length). This would explain the differences between human and EIH/Fourier performance in the diagnostic tests. The hypothesis is then confirmed by the psychophysical tiling experiment. The results also imply a need for a better distance measure which looks over time intervals. Observations from the tiling experiment should serve as a guideline for formulating a proper distance measure, so that diphones may be compared in a way that resembles the way humans compare them.

Note: Since [G94] was handed out in class, these scribe notes freely refer to the figures which appear there. The last two pages of the notes contain huge scanned files of graphs from [G94] which take a long time to ghostview and print.

40.1.1 The EIH Model

Using principles based on properties of the auditory system, the EIH model transforms a speech signal into a representation that is useful for recognition. The model is divided into two stages: (1)peripheral auditory system modeling and (2)post-auditory processing.

EIH Stage 1: Peripheral Auditory System Modeling

Figure 2 of [G94] shows the periphery of the auditory system. Sound waves at the outer ear are transformed by the middle ear into mechanical vibrations in the inner ear. Then the basilar membrane vibrates at places associated with the frequencies of the input acoustic wave. The continuous motion of the basilar membrane is sampled by the inner hair cells (IHC). Each IHC is innervated by approximately 10 auditory nerve fibers which fire signals to the higher levels of the auditory system.

The EIH model, shown in Figure 3 of [G94] uses a high-pass filter to represent the middle ear. The mechanical motion in the cochlea is modeled as a multi-bandpass non-linear filter. This continuous motion is sampled by 190 IHC channels. The channels are distributed over 200 to 7000 Hz in a logarithmic fashion and are highly overlapping for reasons to be discussed. The array of nerve fibers are modeled as five positive-going level crossing detectors per IHC channel. The levels are random Gaussian variables with means equally spaced in dB. The outputs of the level-crossing detectors represent the discharge activity of the auditory nerve fibers in terms of a 950-dimensional point process (190 channels x 5 levels/channel). This is the input to the second stage of the EIH model.

EIH Stage 2: Post-Auditory Processing

The second stage is based on observed properties of cats' auditory nerve response since post-auditory nerve properties are not available. To understand human post-auditory nerve processing, consider the behavior of the auditory nerve fibers in response to a single tone of frequency (f_0) and amplitude A, shown in Figure A. For small amplitudes, R (the total number of spikes firing per second) is constant and there is no order to the firings. As A increases beyond a threshold value, R increases accordingly, and the firings begin to have some order. That is, as A increases in the 20-40 dB range, more and more fibers fire with interval $1/f_0$. They lock onto the input frequency and synchronize with it, for frequencies up to 1-2 kHz. Then, as A increases past an upper threshold, R saturates and does not increase anymore.

To gain further understanding of how the amplitude A is encoded, suppose we could record the firing rate (R) as a function of the characteristic frequency of the fibers in response to the same pure tone. As shown in Figure B, for low amplitudes, only the nerves with CF equal to f_0 fire. The others do not fire yet since they are not tuned precisely. But, as before, for increasing amplitudes, more and more fibers fire with the same interval statistic f_0. The basic assumption of the EIH model is that the intensity (amplitude) is encoded by the *width* of the region on the basilar membrane in which fibers fire synchronously to the incoming sound wave. Thus the EIH model estimates the coherence of activity across the fiber array, not the energy content of the signal, in order to determine the amplitude.

Now that we understand the underlying principle of the model, let us examine the output of the first stage: the response of the simulated fiber array to a speech sound. Figure 7 of [G94] shows a plot of time vs. CF for the five level detectors. Darker areas mean that several levels were crossed. Note how the length of the analysis window in each channel is related to its CF. The time-varying window moves by 5 msec intervals. Since we want to use this information to measure the coherence of fiber activity, the model determines the *similarity* in the short-term interval pdf of the individual level-crossing detectors. An estimate of the interval pdf can be obtained by computing a histogram of all level crossings across all channels. The histogram is scaled in units of frequency. The channels were chosen to be highly overlapping to get a better count of the level crossings.

Example of EIH

Now that Stage 1 and Stage 2 have been described, consider a specific example of how the EIH model works. For simplicity, we examine one channel with CF=fo. This is shown in Figure 8a in [G94] for an input sinusoid of amplitude A and frequency fo. In this case, the cochlear filter behaves linearly, so the output will be a sinusoid of the same frequency. See how the lower 4 level crossing detectors contribute non-zero histograms to the ensemble. Now, since the EIH estimates coherence across the entire fiber array, consider the contribution of the other channels to this interval. As shown in Figure 8b of [G94], channels 2,3, and 4 contribute to the histogram, depending on the amplitude of the cochlear filter at f_0.

Comparison of the EIH model and Fourier spectrum based models

What is the output of the second stage of the EIH model? The EIH representation of the vowel (it) is shown in the lower plot of Figure 15 in [G94]. Notice that the first few harmonics are represented in detail. Compared to the upper plot of the Fourier magnitude spectrum, the EIH does not resolve the second and third formants which are closely spaced. This resembles the upward spread masking which originates in the cochlea.

Another way to compare the EIH with Fourier is to add noise to a clean speech signal and run the models. It turns out that as the noise increases, the Fourier (LPC) method deteriorates much faster than for the EIH model. The first row of Figure C shows the representation of a clean speech signal. The additive noise "flattens" out both curves, as seen in the second row of Figure C. However, after cepstral smoothing, the Fourier method has a much larger mismatch between the clean and noisy signals, as seen in the third row of Figure C. Therefore, the EIH model is better as a front-end to a recognizer.

Figure D summarizes the conceptual difference between the Fourier power spectrum measurement and the EIH measurement, respectively. Both consist of two stages: a filter bank and an analysis stage. The Hamming BP filter bank contains identically shaped filters equally spaced on a linear scale. The cochlear BP filter bank contains uniquely shaped filters equally spaced on a logarithmic scale. The Fourier analyzer estimates the power spectrum by computing the short-term power at the output of each filter. The EIH analyzer estimates coherence across simulated fibers.

40.1.2 Evaluation of the EIH Model

Speech Recognition

The EIH model was evaluated in a speech recognition task of discriminating between work pairs in the Diagnostic Rhyme Test (DRT). The DRT procedure was described in detail in the previous lecture.

To simulate the DRT using the EIH and Fourier front ends, an HMM recognizer was built for each word in the test. Suppose the task is to distinguish *peak/teak*. The diphone-state model of both words are loaded into the recognizer. *Peak* is then run through the model and recognizer. The recognizer decides if the word was *peak* or *teak* and the result is marked as correct or incorrect. Figure 12 of [G94] shows the results, which are plotted against the twelve phonemic categories. In summary, EIH did well against Fourier, but not that well with respect to humans.

Speech Coding

In the second test, the EIH is used to predict human judgement of the quality of synthetic speech produced by different speech coding systems. The motivation for this evaluation is: If the EIH representation is good, machine judgement should be the same as human judgement. The MOS (Mean-Opinion Score) test was used. Human subjects listen to a speech sentence and rate the quality from excellent(5) to bad(1). The results are averaged over the database and subjects.

To simulate the MOS test, the original and coded time waveform are first aligned in time. Then the EIH is run on both the original and coded speech. The Euclidean distance between two corresponding "frames" of EIH representations was measured. After weighting the distances by the loudness of the original speech at that time instant, the mean value of the weighted frame distances is computed to produce the "EIH distance". This procedure is repeated for each coder. The result is shown in Figure 13 of [G94] in the left-hand column graphs. There is a linear regression between EIH and MOS for the coders with a bit-rate 16 kbits/sec and above. This means that EIH successfully predicts the MOS of these high-quality coders. However, the MOS of the CELP coders were poorly predicted. That is, the quality of the CELP coded speech segments were incorrectly predicted to be worse than what humans (MOS) judged it to be, as seen by the cluster of solid black dots *above* the linear regression line.

40.1.3 Interpretation of Results/Future Work

Why does this discrepancy between humans, EIH, and Fourier exist?

The speech recognition evaluation results suggest that the Euclidean type of measure between two EIH observations is inappropriate. For recognition, a better measure reflects phonemic distance between speech observations. The exact location of the individual harmonics is not relevant to the identity of a phoneme and large errors due to misalignment of the harmonics mask relevant error information.

For the quality of speech coding task, first of all, note that humans (MOS) don't compare the coded speech with respect to the original. This is why coders just aim to make the speech sound "ok". However, the EIH distance compares coded-speech directly with the original speech. Second, note that the Euclidean norm captures local distance between two EIH representations. For the high-quality coders, where the quantization noise is localized, EIH distance is a good predictor of MOS (human judgement). For CELP however, the difference between the original speech segment and synthesized segment is spread smoothly over the entire segment. Consequently, a comparison between two single observations taken at the same time instant, one from the original waveform and one from the CELP coded waveform, will result in an irrelevant large error due to misalignment of the harmonics. This explains why EIH judges the quality of the CELP coded speech to be worse than what humans (MOS) judge it to be.

These results also lead to the hypothesis that humans use a representation which integrates information over time intervals of 50-100 msec.

To test this hypothesis, the tiling experiment was developed. The results of the tiling experiment confirmed human usage of perceptually related integration rules over durations of 50-100 msec, and showed different integration rules for different time-frequency regions depending on the phoneme-discrimination task. For details, see [G94].

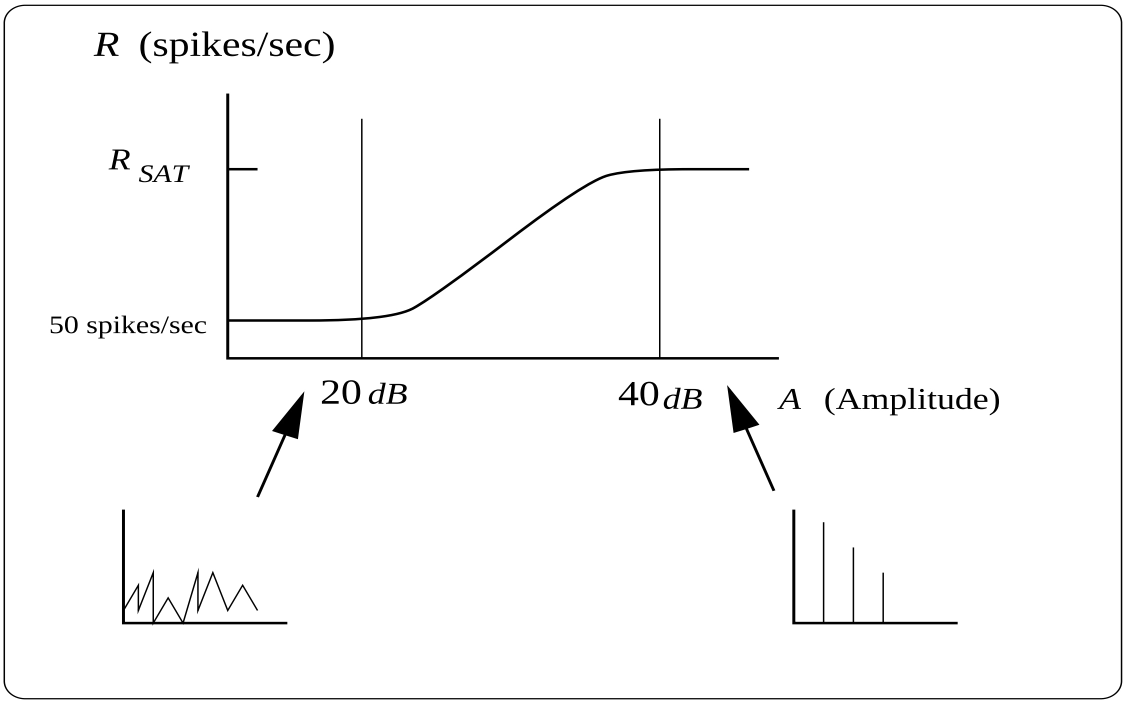
Conclusion

Current auditory models provide "snap-shot" representations (i.e., representations which are local in time and frequency) A representation based on integration over time intervals of 50-100 msec is needed. Also needed is a formulation of a proper measure of the distance between two speech segments of that duration.

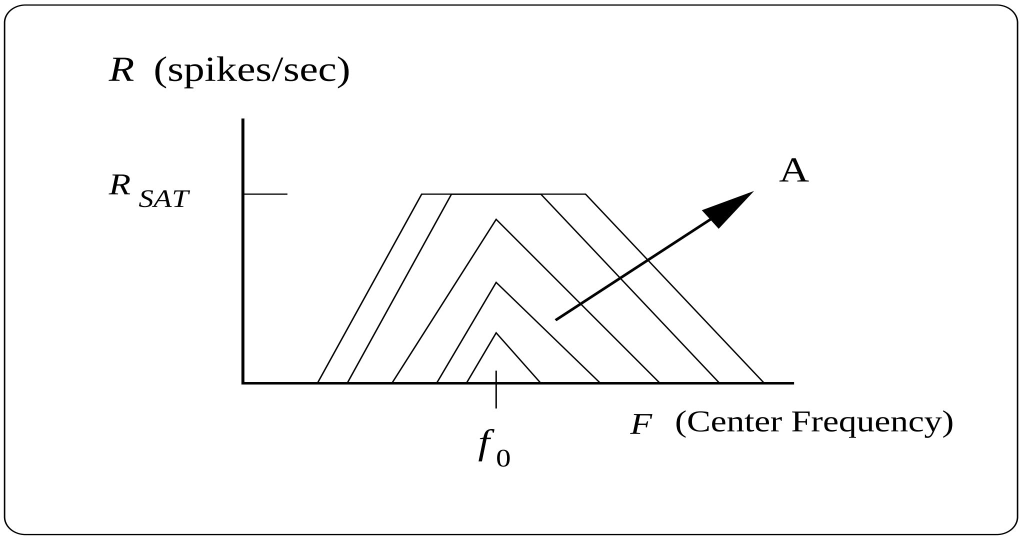
References

[G94] O. Ghitza *Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition*, IEEE Tr. Speech and Audio Processing, 1994, pp. 115–132.

Fiber Response to a Simple Tone (Ghitza) - Figure A



Fiber Response to a Simple Tone (Ghitza) - Figure B



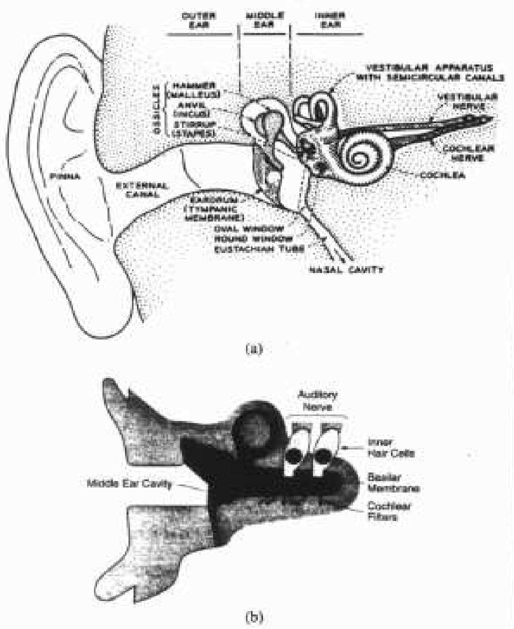


Fig. 2. (a) A physiological model of the pre-auditory-nerve parts of the auditory periphery. (b) A simplified view of (a). The 2 1/2 turn, snail-like shape of the cochlea is shown as a straight tube.

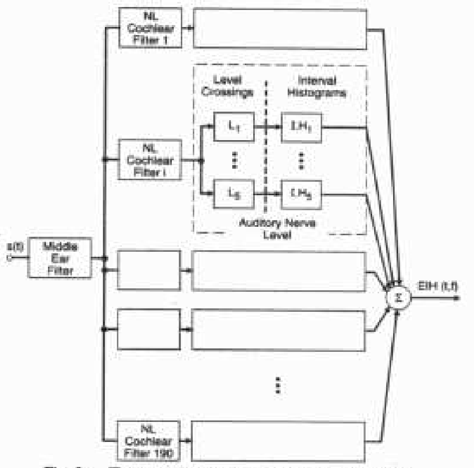


Fig. 3. The ensemble interval histogram (EIH) model.

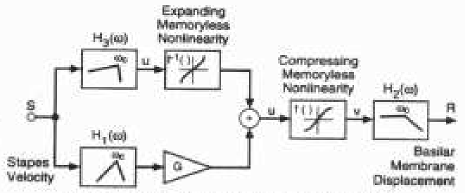


Fig. 4. The MBPNL cochlear model (after [13]).

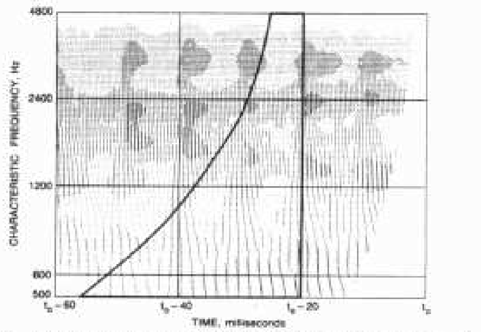


Fig. 7. Simulated auditory-nerve activity for the first 60 ms inside the vowel [o] in the word "gob." The abscissa represents time and the ordinate represents the characteristic frequency of the IHC channels. Note the logarithmic scale of the characteristic frequency, which represents the place-to-frequency mapping at the basilar membrane. In the figure, a level-crossing occurrence is marked as a single dot, and the output activity of each level-crossing detector is plotted as a separate trace. Each IHC channel contributes five parallel traces, with the lower trace representing the lower-threshold level-crossing detector. If the magnitude of the filter's output-signal is low, only one level will be crossed, as is the case for the very top channels of the figure. However, for large signal magnitudes, several levels will be activated, creating a "darker" area of activity. The figure also illustrates how the length of the analysis window in each channel is related to its CF.

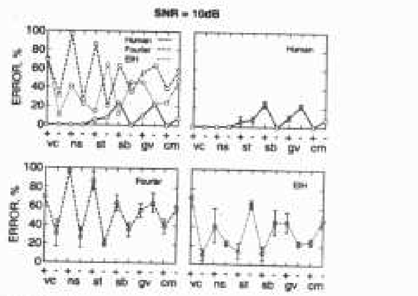


Fig. 12. Distribution of errors made by the human listener, an EIH with an ERB bin allocation and a Fourier power spectrum, in a 10 dB signal-to-noise ratio. The left-upper plot is a superposition of the other three plots, excluding the confidence intervals. The abscissa of every plot indicates the six phonetic features: "vc" is for *Voicing*, "ns" for *Nasality*, "st" for *Sustention*, "sb" for *Sibilation*, "gv" for *Graveness*, and "cm" for *Compactness*. The "+" sign stands for attribute present and the "−" sign for attribute absent. The line connecting the measurements is only for display purposes, to enable the reader to distinguish between error patterns that belong to a particular parameter value. The noise is additive and white, and the signal-to-noise ratio is defined using global measurements (see text). Note that an error value of 100% in a given phonemic category means that all 16 words in the category were mistakenly identified as their DRT counterparts.

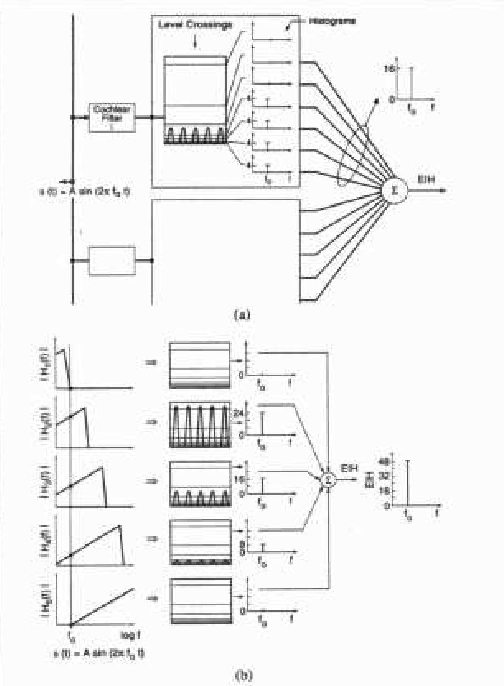


Fig. 8. (a) The contribution of the ith channel to the Ensemble Interval Histogram for an input signal $s(t) = A\sin(2\pi f_o t)$. Only the lowest four level-crossing detectors are contributing nonzero histograms to the ensemble. (b) The contribution of five successive channels $|H_i(f)|$,i=1,2, $\cdots$,5, to the EIH, for an input signal $s(t) = A\sin(2\pi f_o t)$. Channel i contributes to the f_o bin of the EIH provided that $A|H_i(f_o)|$ exceeds any of the level-crossing thresholds.

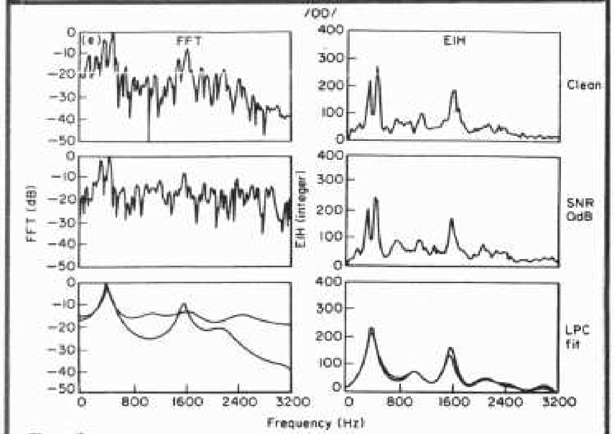


Figure C

Figure C Illustration of the robustness of the Ensemble Interval Histogram (EIH) in the presence of additive white noise, with global signal to noise ratio of 0 dB (see section 3.2 for our global-SNR definition). The left column represents the FFT spectra of a vowel /o produced by a male speaker and the right column represents the EIH spectra. The all-pole polynomial fit is a 16th order LPC for the FFT spectrum and a 19th order LPC for the EIH spectrum.

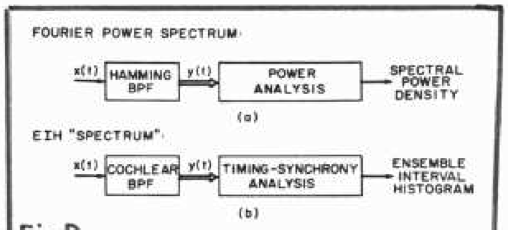


Fig D

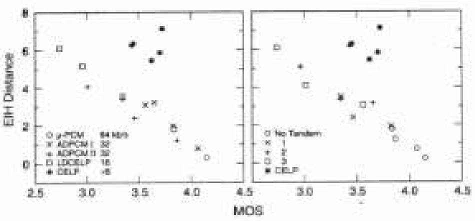


Fig. 13. EIH-distance versus MOS. The results are grouped according to coder identity (left) or according to tandem configuration (right). A particular coding condition can be located by a cross-alignment of the left and the right plots.

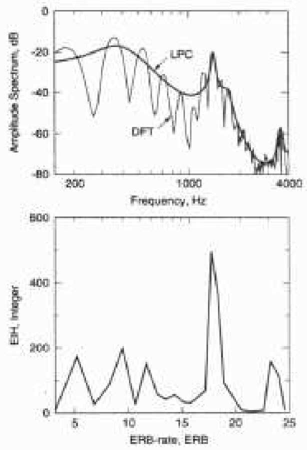


Fig. 15. Amplitude spectrum and EIH representations of the vowel [*ir*] (like in bit) produced by a male speaker. The upper plot contains the high-resolution Fourier magnitude spectrum computed by a 128 point discrete Fourier transform, using a 20-ms Hamming window. It also contains the envelope fit achieved by an tenth-order all-pole polynomial-fit. The plot uses a log-frequency/decibel scale. The lower plot is the EIH representation of the same signal. The EIH plot is in a linear-ERB/linear-EIH scale. Note that the abscissa spans the frequency range [250, 4000] Hz. In this frequency range, the ERB-rate scale overlaps with the log-frequency scale.