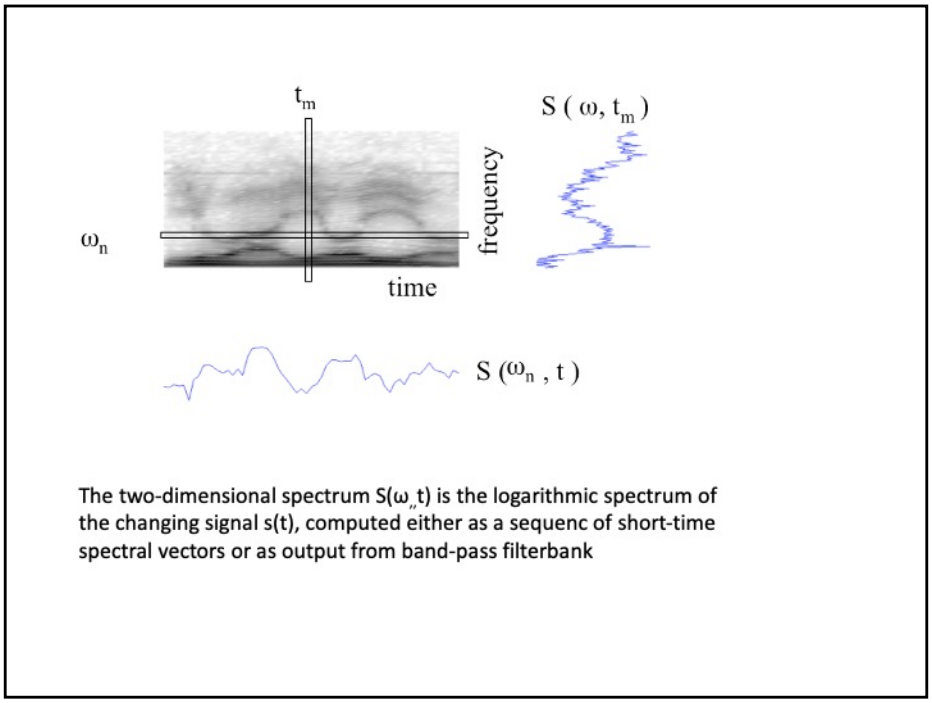


The two-dimensional spectrum S(ω,,t) is the logarithmic spectrum of the changing signal s(t), computed either as a sequenc of short-time spectral vectors or as output from band-pass filetrbank.

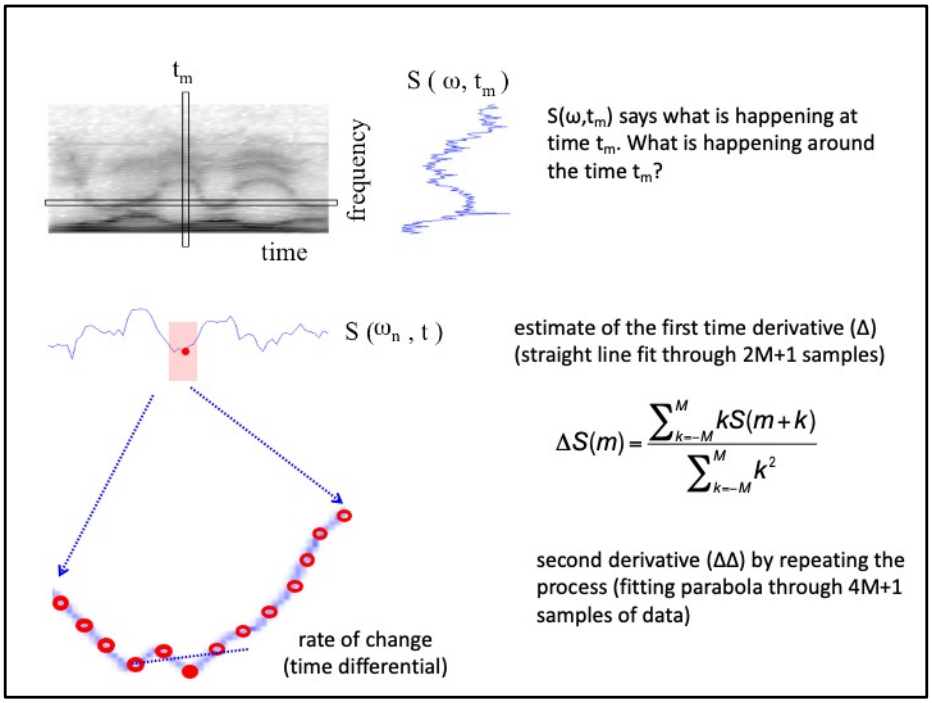
A function *f(x)* at its point *a* can be approximated by using a finite number of terms of its **Taylor series.**

$$\sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n$$

The first element of the series in the function value *f(a)* itself
The second element is the function firsts difference *f(x-a)*

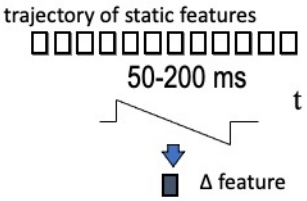
- This difference can be better estimated by a line fit to the function values in the neigborhood of *a*

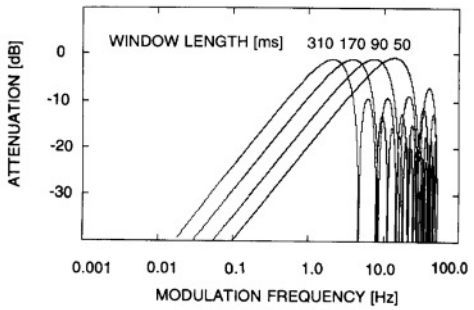
One of the earlier attempts for using spectral dynamics was computing the so called "dynamic features" of speech. In effect these are Taylor series expansions of the logarithmic spectral trajectory at the time instant t_a, computed at each time instant.



Typically, the first two Taylor series coefficients are used, representing estimates of the first and the second differencial of the logarithic spectral trajectory at a given time t_m.

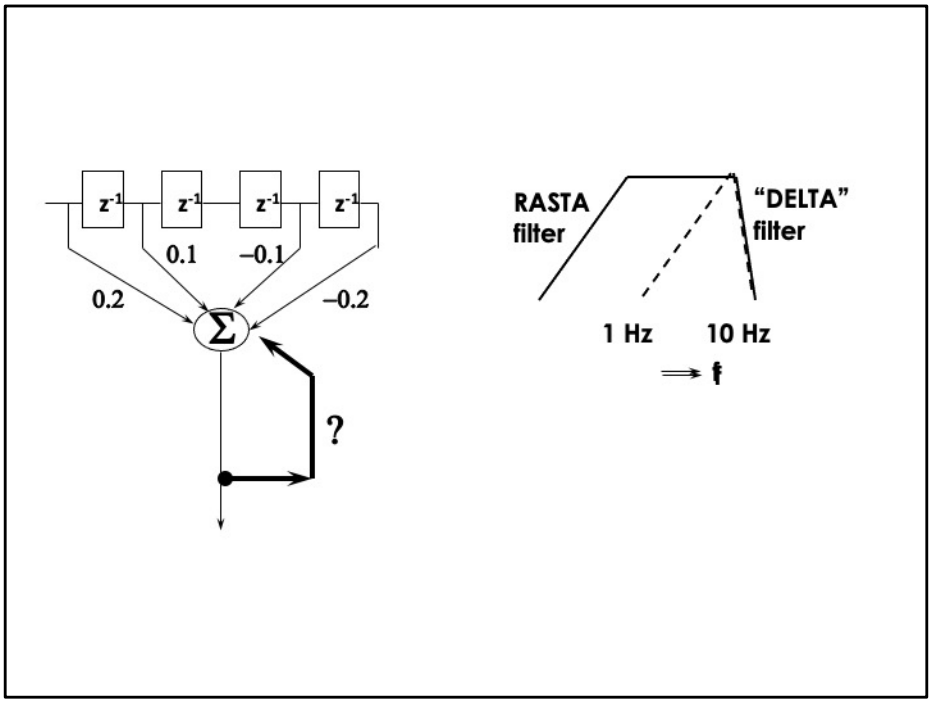
Delta Features



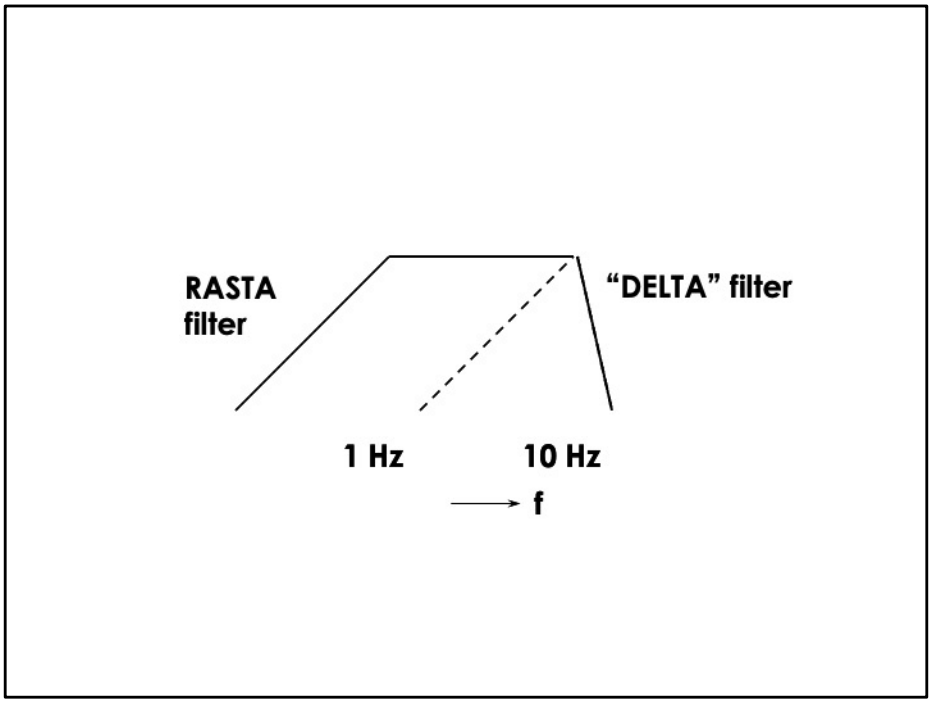


- linear combination of several short-term features
- equivalent to FIR filtering of feature trajectory
- selective band-pass with 6 dB/oct slope
- frequency response of the FIR filter depends on the length of interval of the estimation

In the modulation frequency domain, the way the dynamic features are being computed represent finite impulse response (FIR) filtering. The first differential is estimated using the filter with a rather unusual impulse response with sharp discontinuities at its edges. As a consequence, the filters generates very significant splatted in the modulation domain. Frequency responses of the delta-comuting filters depend on the filter length, i.s. on the intervalk over which the Taiyor series weights are being estimated. We can see that the delta computation strongly enhances a narrow range of modulation frequencies.

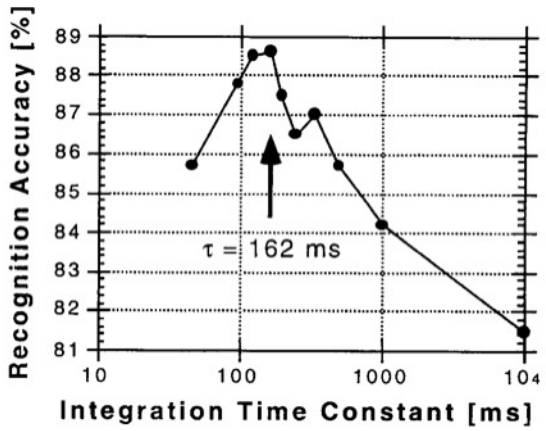


Once we understand the delta FIR filter, we can start thinking about modifying it. A simple but substantial modification is to provide a feedback from its output, thus effectively changnging the FIR filter into the infinite impulse response (IIR) filter. For those not skilled in digital signal processing, please, trust me ☺

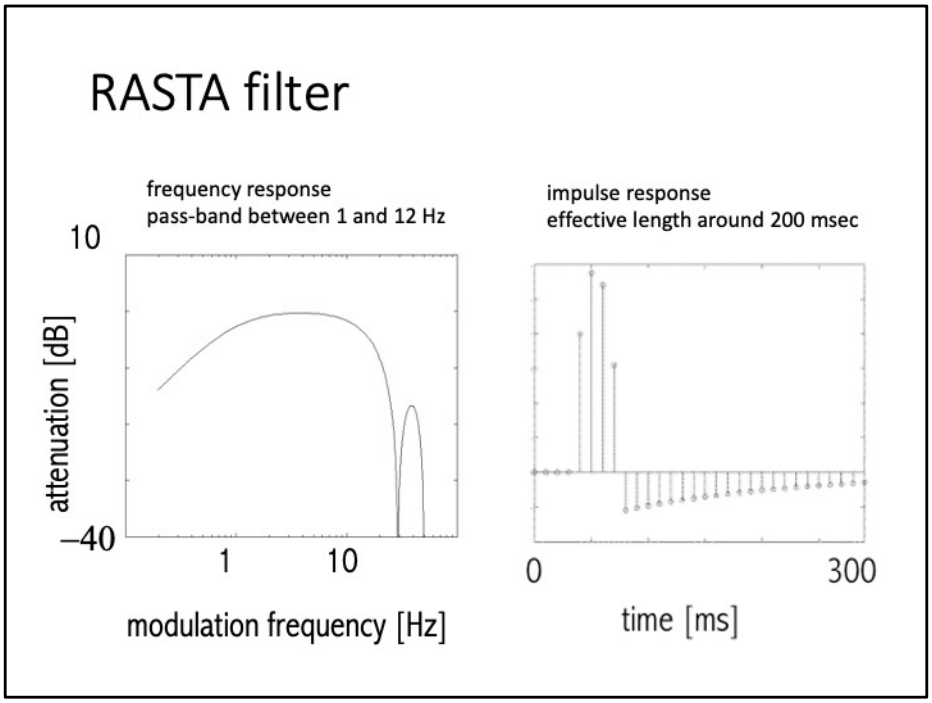


Rthis addition of the feedback loop widents the passband of the filter.

Optimizing RASTA Filter

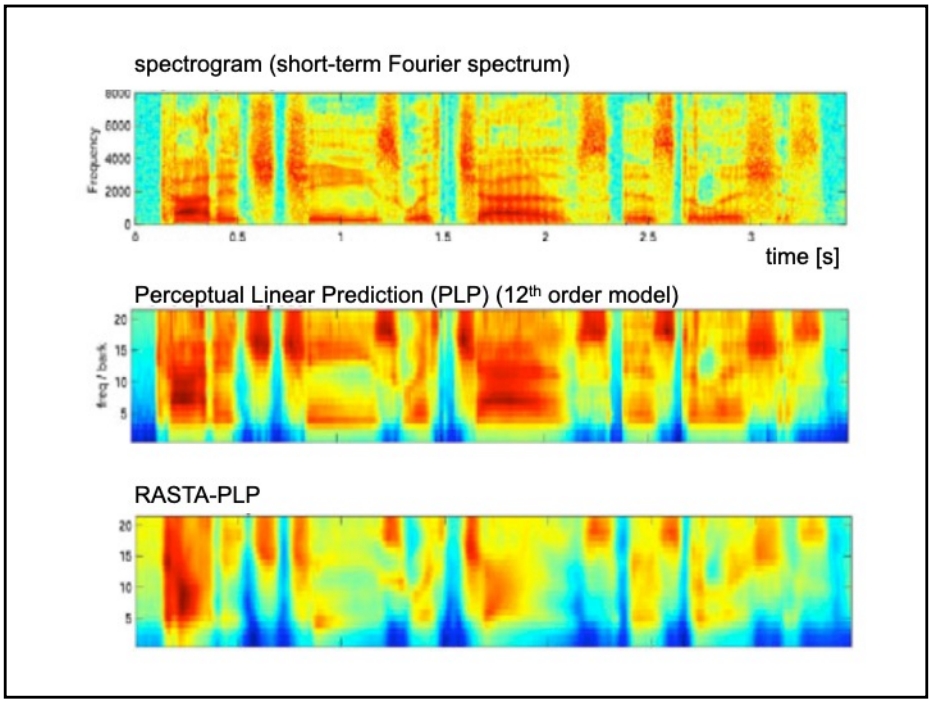


So far we do not know what should be the weight on the feedback loop. Being engineers, when we do not know something, we run experiment ☺. As one well known reseracher said: "Research is what I am doing when I do not know what am I doing." There was a broad optimum in speech recogntion performance somewhere around rhe weight 0.94.

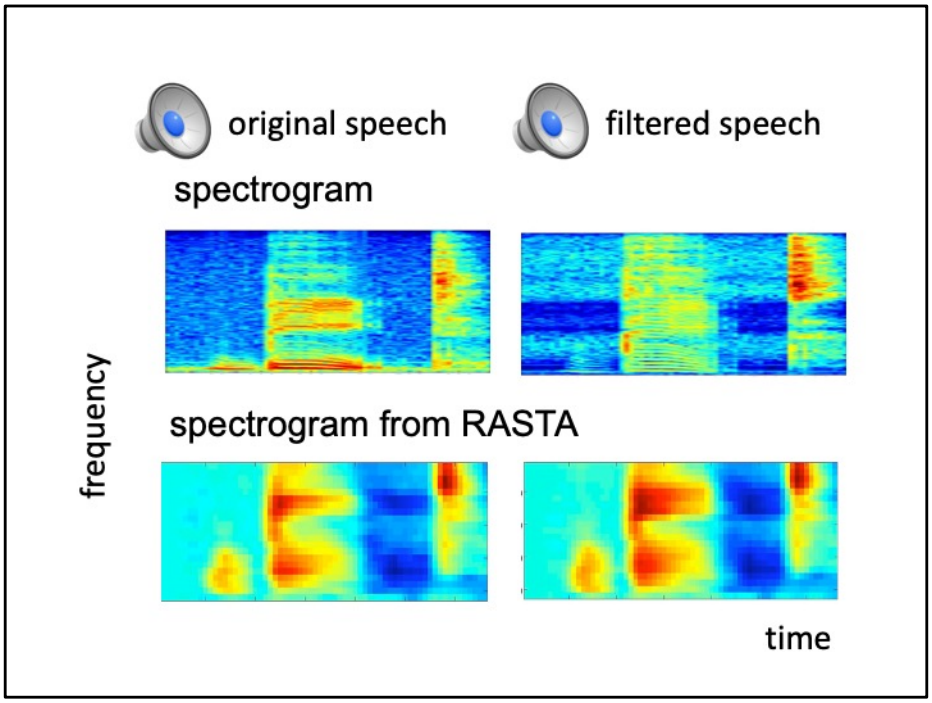


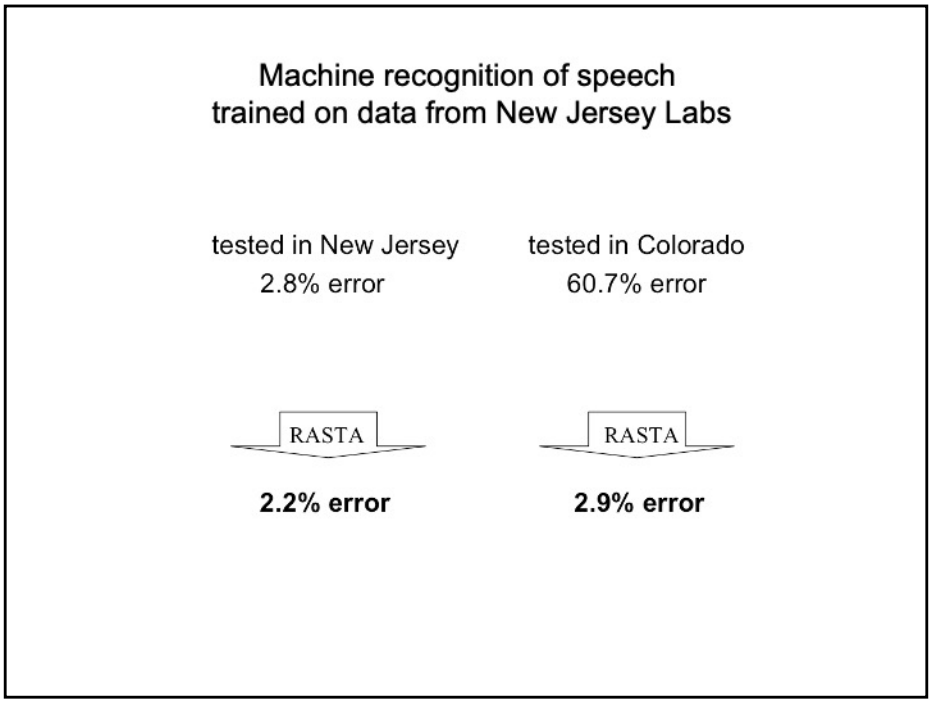
When l;ookoing at the frequency response of the optimal filter we can see that the filter passes modulation frequencies betwee 1 and 12 Hz, well where the most significant modulations frequencies of speech are. Effective length of the impulse response is ab=around 200 msec.

We did not know that at the time when we were designing the filter, maybe we did not have to run the experient.
Indeed, “Research is what I am doing when I do not know what am I doing.”

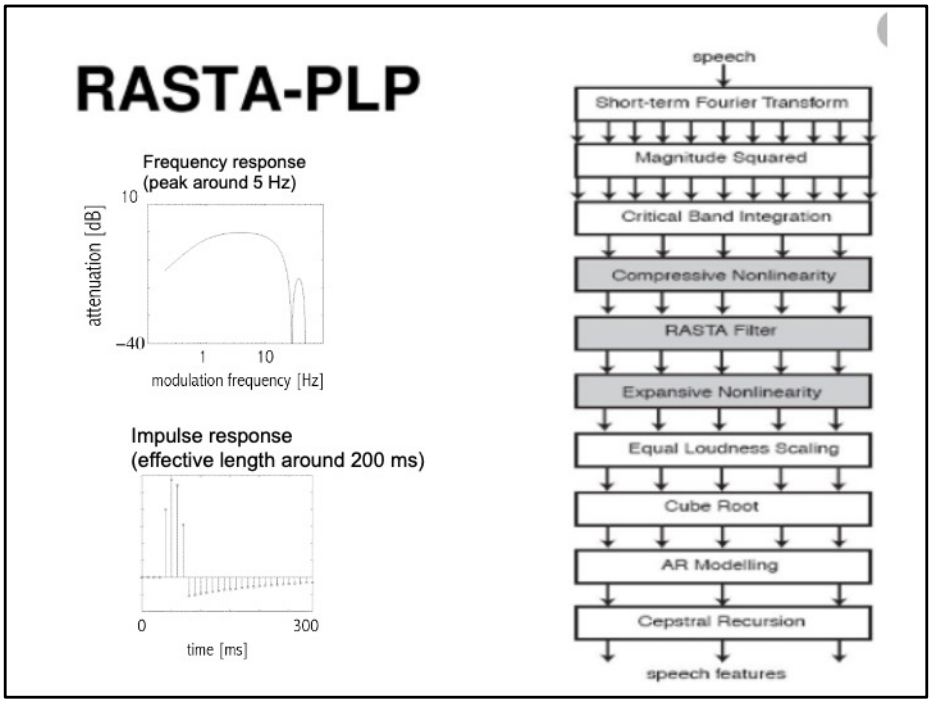


The RASTA technique supresses the steady parts of speech spectrum, effectively enhanciong the transitions (remember what perception likes to do?)

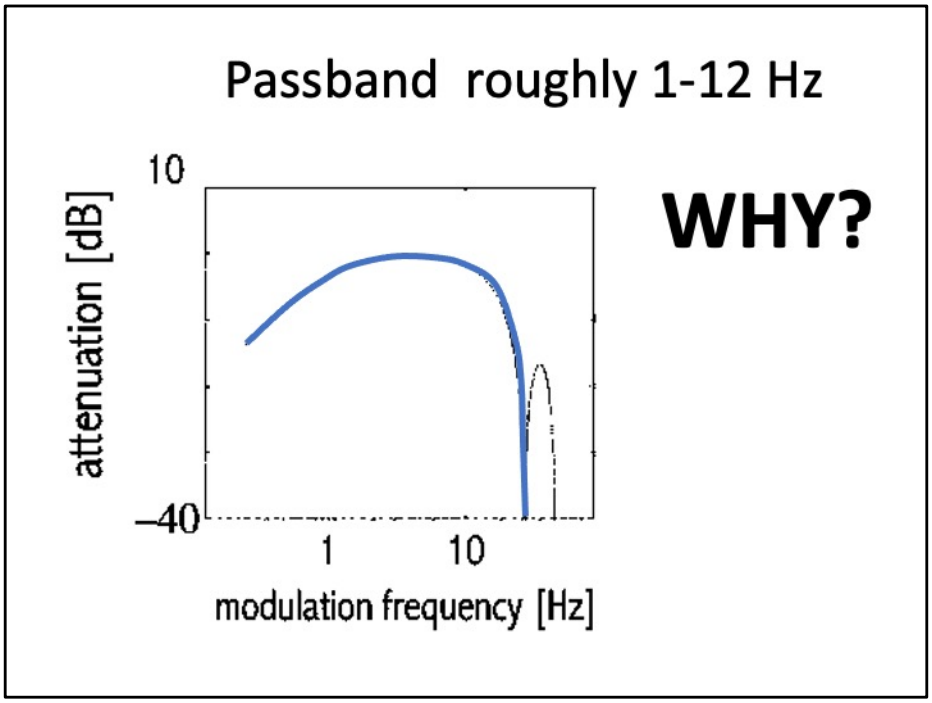

original speech
filtered speech
spectrogram
frequency
spectrogram from RASTA
time



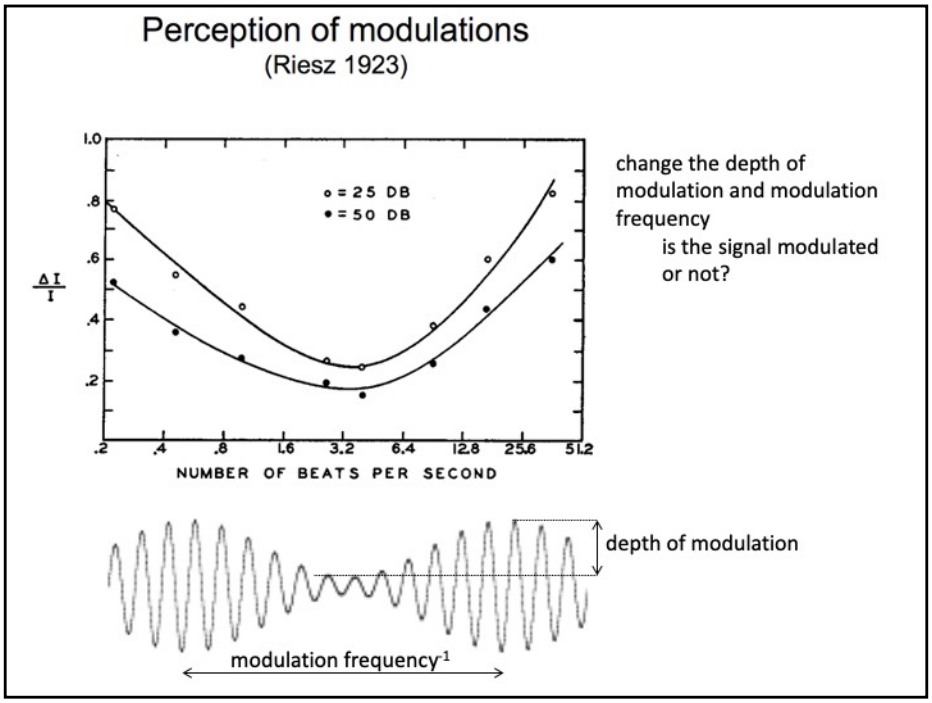
At that time we had a problem with our recognizer in our lab in Boulder, Colorado. The recognizer was trained on data from New Jersey, worked very well on the test data from the same acoustic environment in New Jersey but its performance on our data in Boulder was terrible. After using the new technique, the problem was gone, the differences in acosutic environments between New Jersey and BOulder was gone.



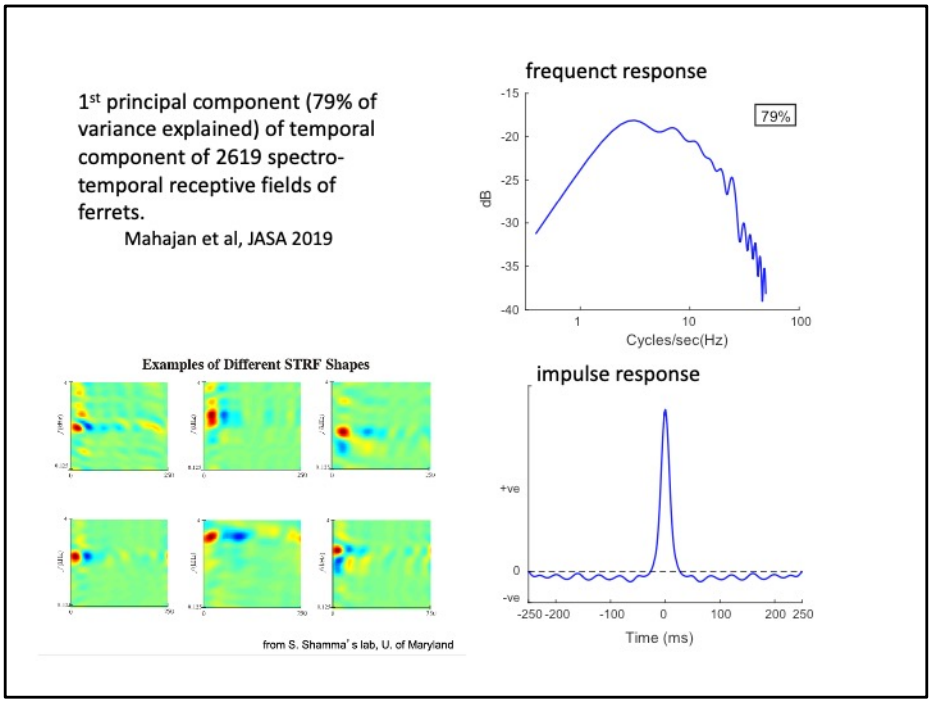
Here we see the whole system for RASRA-PLP analysis. The critical-band spectral energy trajectories are compressed by logarithmic nonlinearity and filtered by the RASTA band-pass modulation filter. FIltered trajectories are transformed back into power spectral domain and the equal-loudness with the subsequent intensity-loudness root transform yield auditory-like spectrum for the final autoregressive spectral estimation.



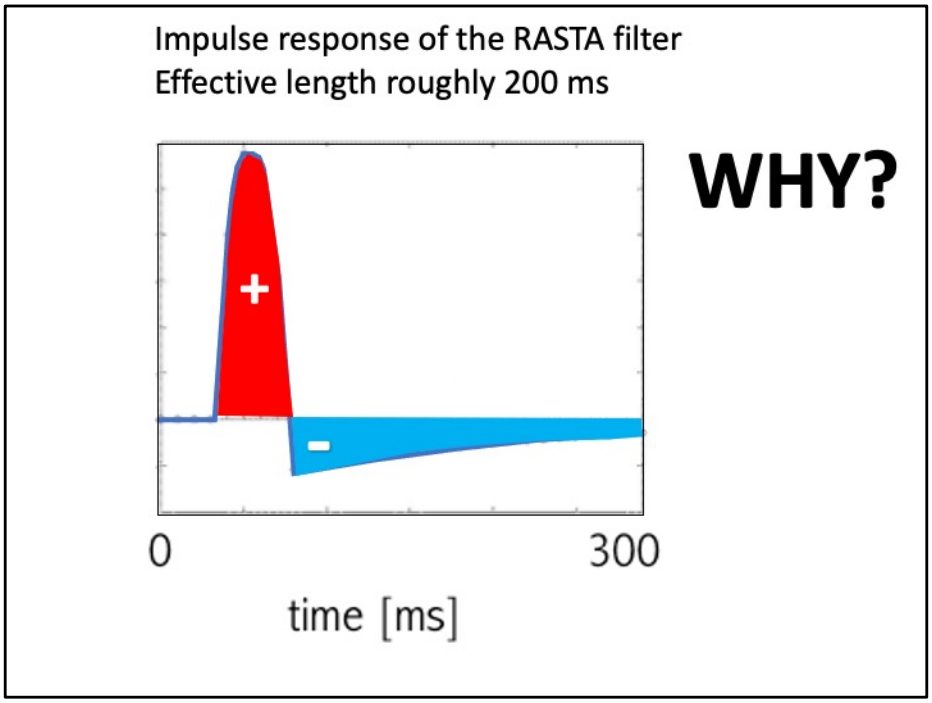
The pass band of the technques is between 1 and 12 Hz in the modulation domain.



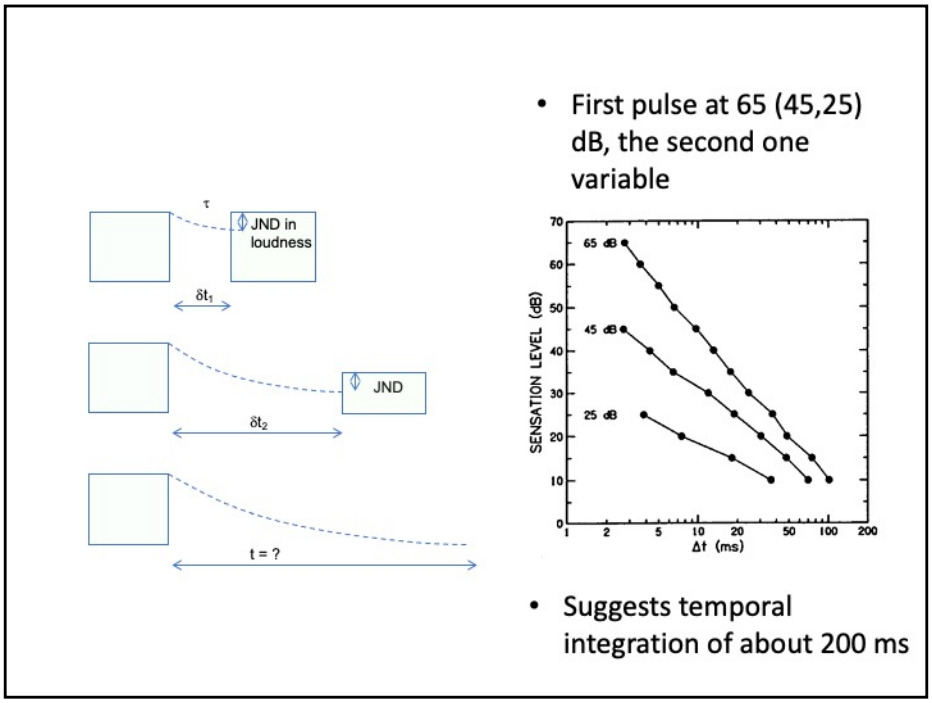
Human hearing is most sensitive to modulations around 4-5 Hz (period 200-250 ms), passing well modulations between 1 and 12 Hz.



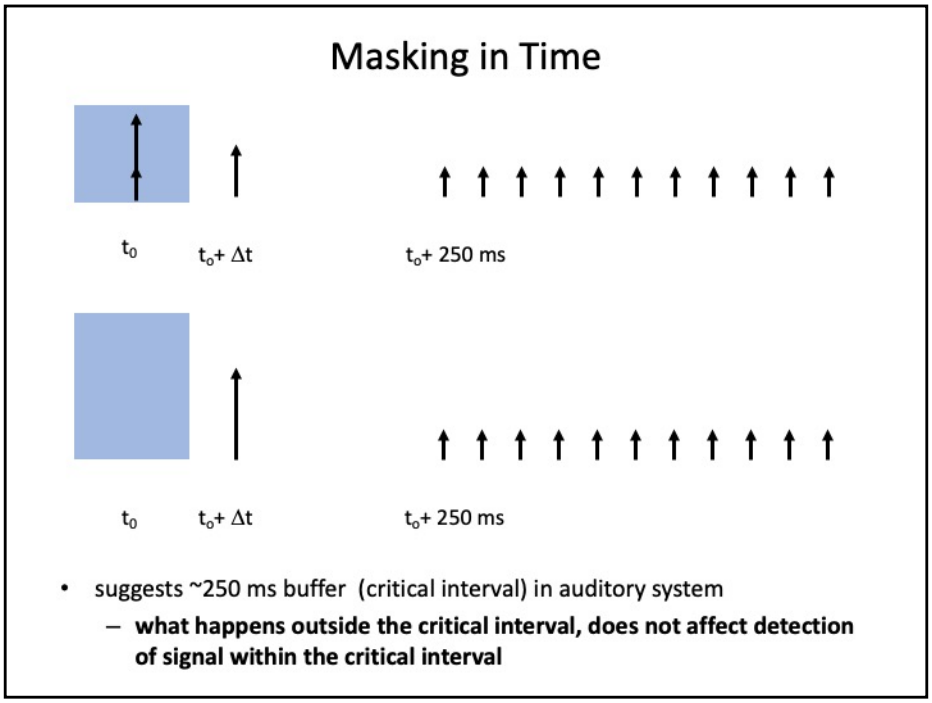
Some examples of auditory cortical receptive fields of ferrets (human receptove fields re quite similar) show that the field often expands over several hundreds of msec. There is a way of estimating their dominant temporal components. Fourier transform of the average dominant temporal component (the impulse response) shows that the higher stages of hearing are the most sensitive to modulations around 3-6 Hz. That is where the dominant rate of change of speech also lays. ”We speak in order to be heard…”



Impulse response of the RASTA filter remembers past values for about 200 ms. The long impulse response implies the necessary temporal “buffer” for the processing and induces the memory” or “slugisshness” of the processing, with an effective lengths of the order of a couple of hundreds of msec. Such a buffer is implied in several hearing experiments.

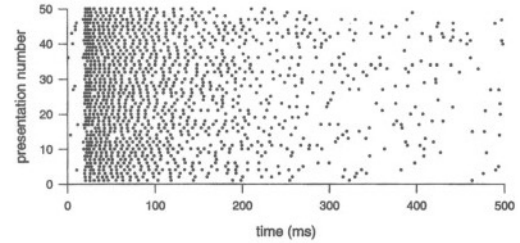


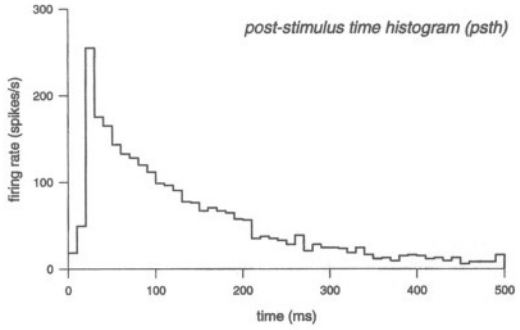
The inertia of human hearing is around 200 msec. This is well shown in the experiments with the gap detection. The gap is perceived when the decline in the intensity of the first sound (shown here as dashed curve) is such that there is perceived discontinuity when the second sound comes. This happens when the difference between the fading of the first sound and the amplitude of the second sound represents the just noticeable difference (JND) in intensities. Weaker the second sound is, longer it takes for the effect of the first sound to decay. The effect of the first sound dies away completely after some time, which can be extrapolated from the perceived gaps for different levels of the second sound. Think about similarities with the effect of forward masking which we discussed earlier.



In addition to the simultaneous masking, when both the masker and the probe occurs at the same time, there is also masking which happens when the masker and the probe are offset. So called forward masking happens when the masker precedes the probe. The effect of the masker decreases with the increasing time interval between the masker and the probe, and diminishes entirely after about 200 ms, regardless of how strong the initial masker was. For weaker maskers, the decline in masking is more gradual, for stronger masker, the masking effect declines faster. It seems fair to say that this effect is consistent with the notion of some "critical interval" (about 200-250 ms) within which the masker interact with the probe. When the masker and the probe do not share this critical interval, there is no interaction.

Post-stimulus histogram





We have seen post-stimulus historgmas sevel times before but we nebver discussed how these histograms are created. Post-stimulus histogram is obtained by counting a number of spikes in repeated stimulus presentation. Here we see the result of changing the scene at the time zero 50 times and averaging the response of optical nerve (of a fly). The visual system is the most responsive shortly after the scene change (some 20-30 ms after the change) and then the response gradually dies off

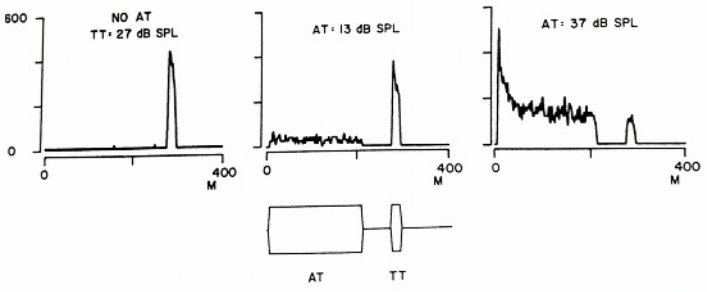
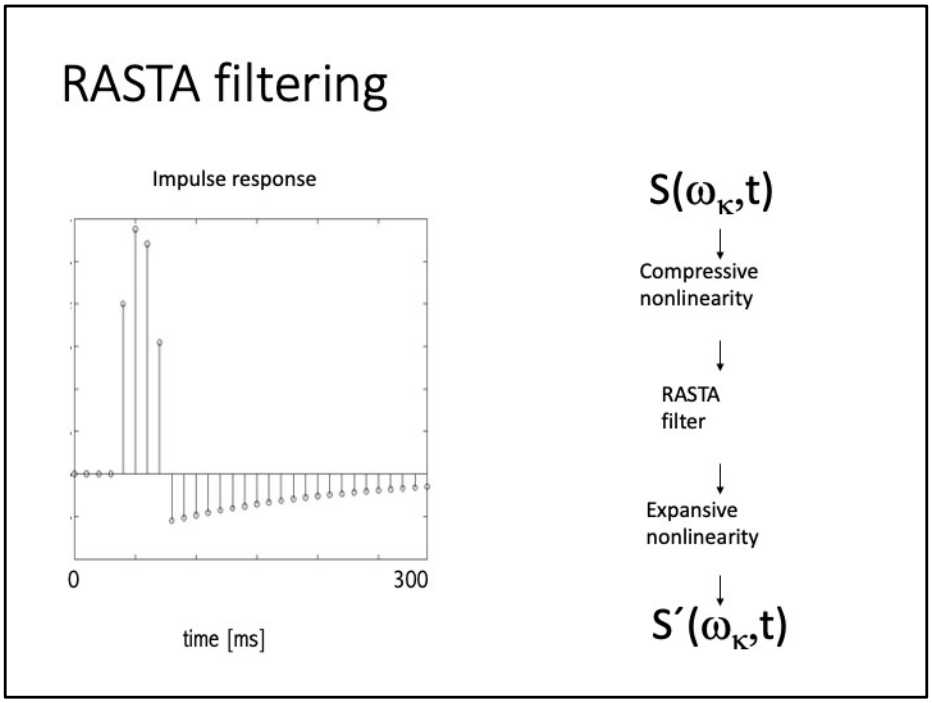
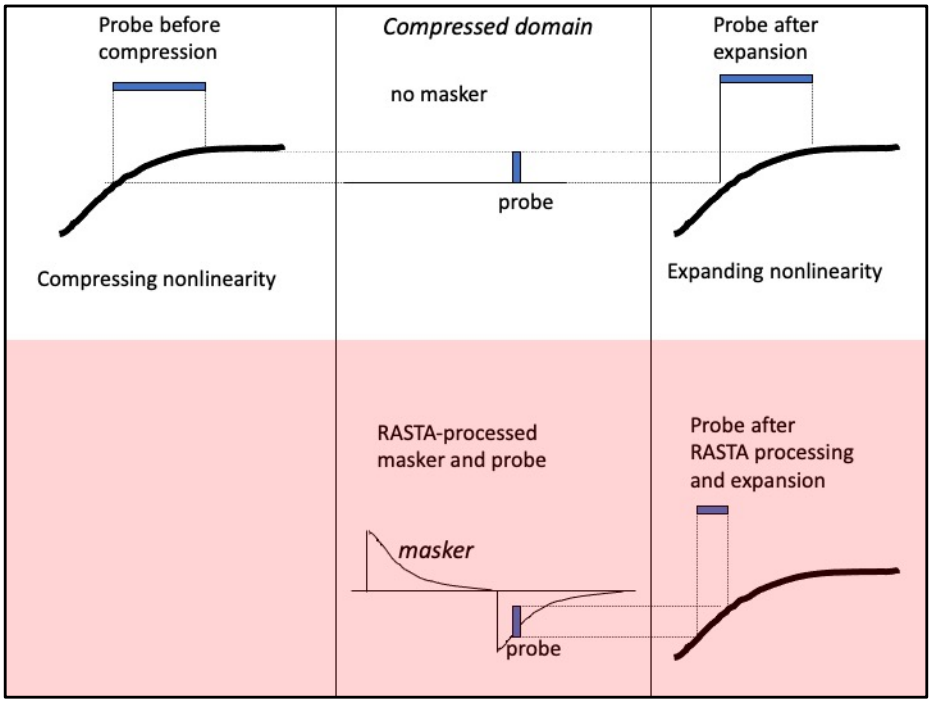


Figure 4.17 Poststimulus time histograms for an auditory nerve fiber in response to a test tone (TT) of duration 20 ms preceded by an adapting tone (AT) of duration 200 ms, with an interval between tones of 60 ms. The timing of AT and TT is shown at the bottom. The different panels show data for different levels of the adapting tone, with the test tone fixed at 27 dB SPL. (From Delgutte, 1980.)

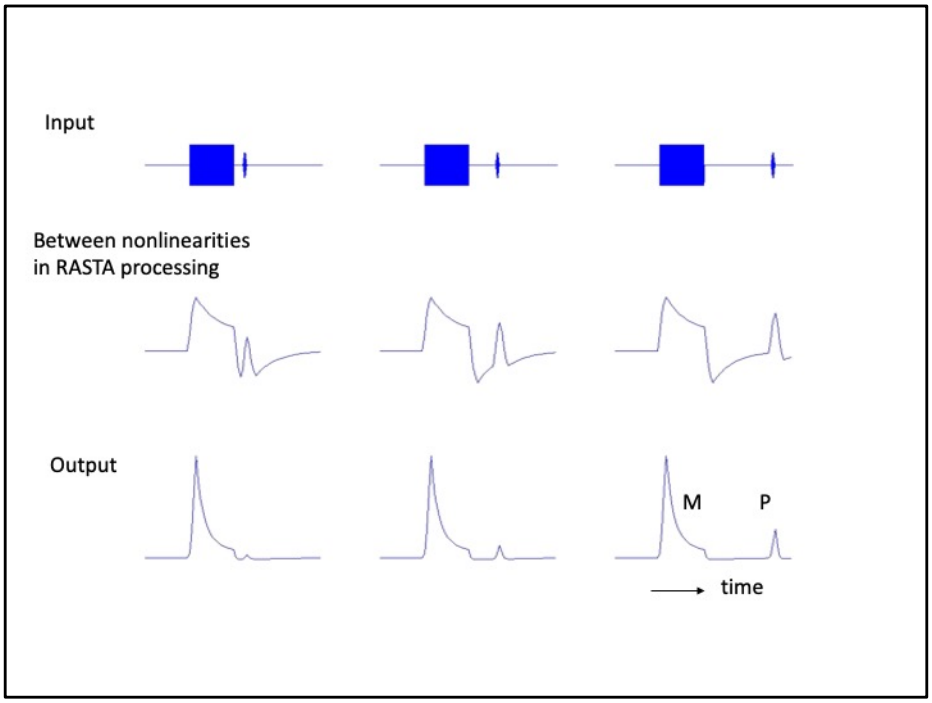
The mechanism of the forward masking was observed on measuring firing rates on the auditory nerve, supporting the peripheral origing of this phenomenon. It can be interpreted as an estimate of posterior probability of a response to a given stimulus. In the first display, there is only a short burst (test tone/probe) of sound preceeded by a slilence. In the second display, the burst is preceeded by a weak adapting tome (masker). The response to the burst is slightly diminished. In the last display, the masker is much stronger. We can see a strung response to this masker at the beginning of the masker. The response is diminished after the masker onset. The test burst (the probe) is significantly diminished.



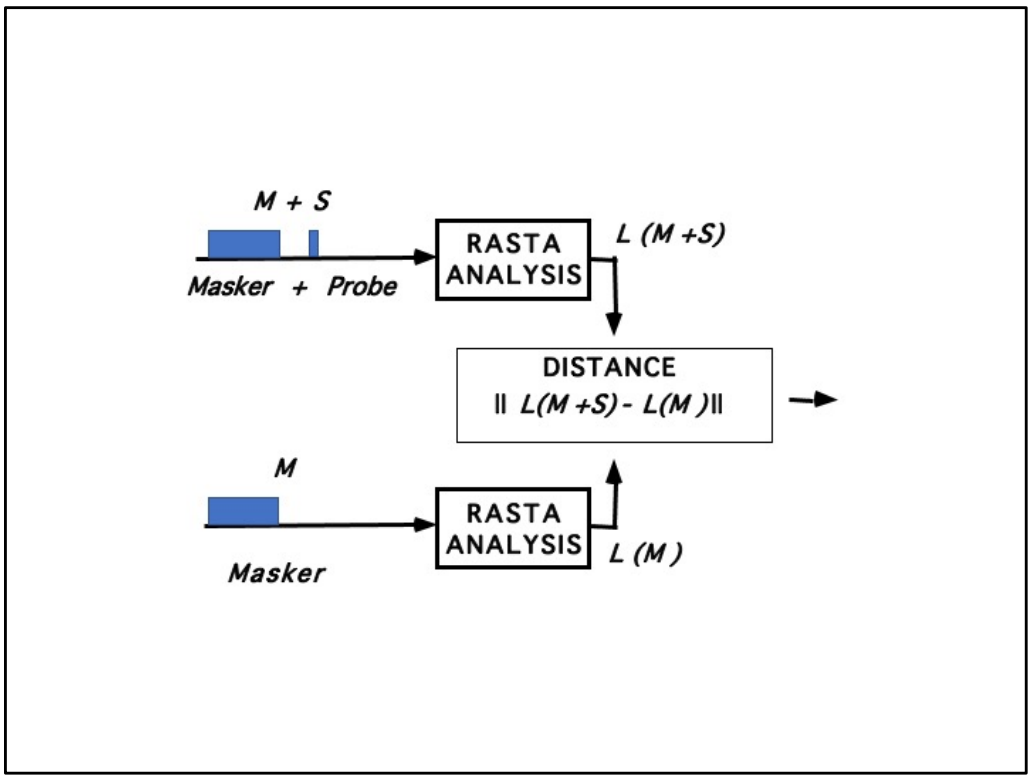
Rasta filtering done between the compressive and the expansive nonlinearity can emulate the forward masking. The time constant of the system is similar to time constant observed in forward masking. Can it model the temporal masking?



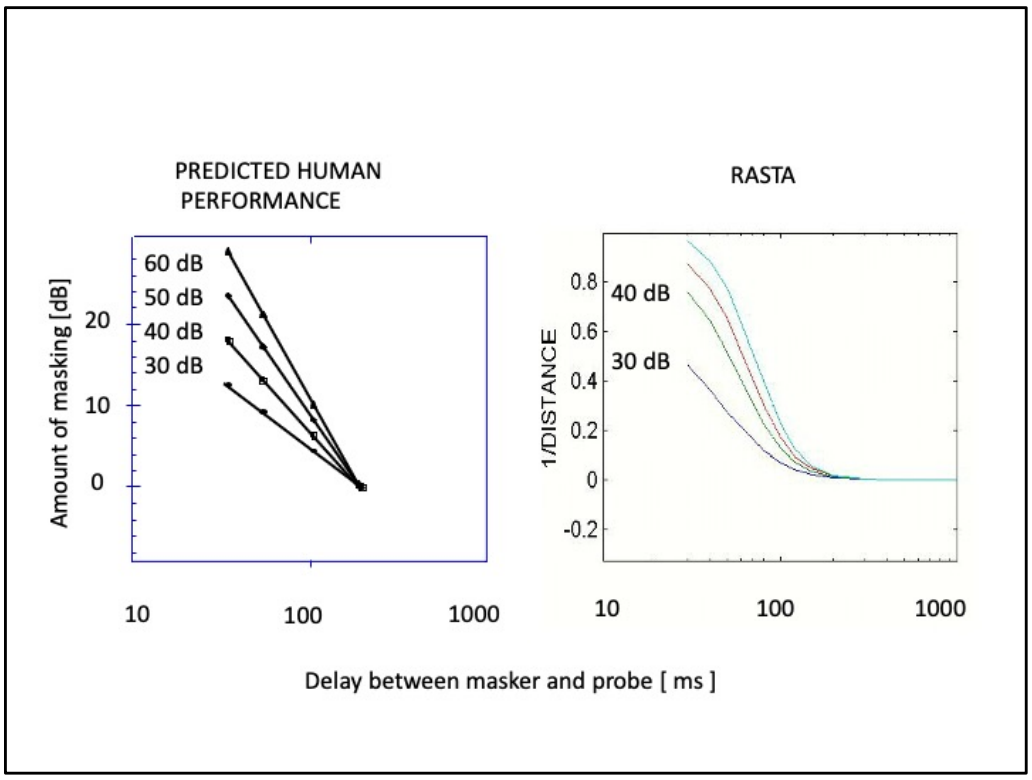
The masker-probe interaction is happening between the compressive and the expansive nonlinearity. Bpth are first compressed by the compressive nonlinearity. The switching off the masker elicits the negative response, which eventually goes to zero. If the probe comes after the response to the masker dissapears, it is being projected back through the subsequency expansive nonlinearity is its original size. However, when it comes during the negative tail of the masker, it is projected on the less expansive part of the nonlinearity so it appears on the output of the whole processing as being attenuated.



This shows the real processing of the masker and the probe at the input, between the nonlinearities and at the output of the circuit. Diminished response to the probe near the masker is clearly demonstrated.



The experimental setup measures the effect of the masker on the probe. The larger the probe appears, the larger is the measured distance.



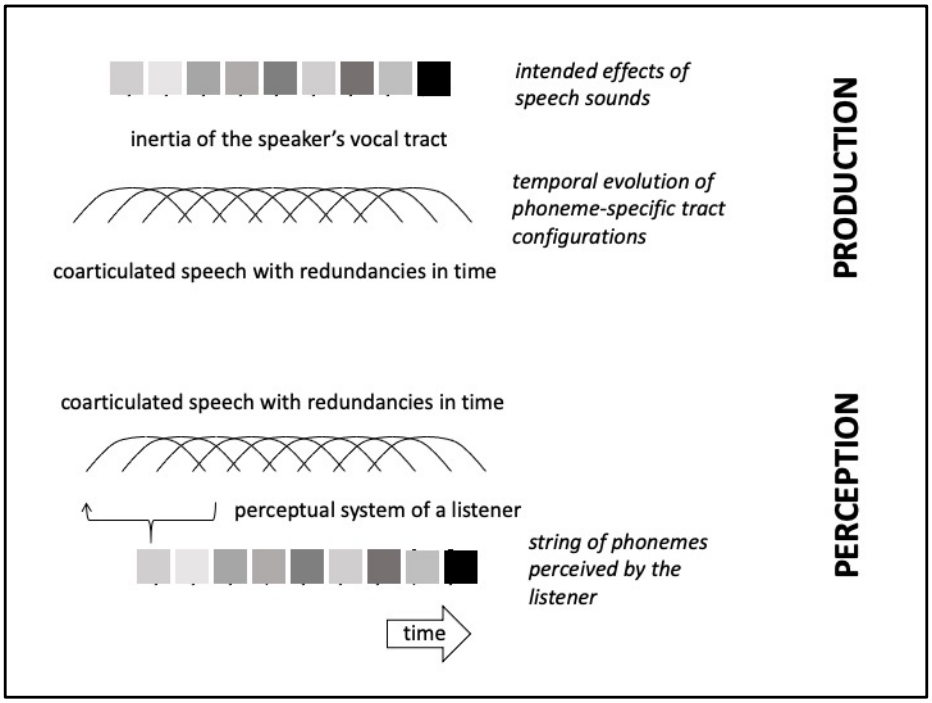
Inverse of the measured distance indicates the amount of the masking. The result of the emulation agrees reasonably well with the perceptual data.

Speech Production

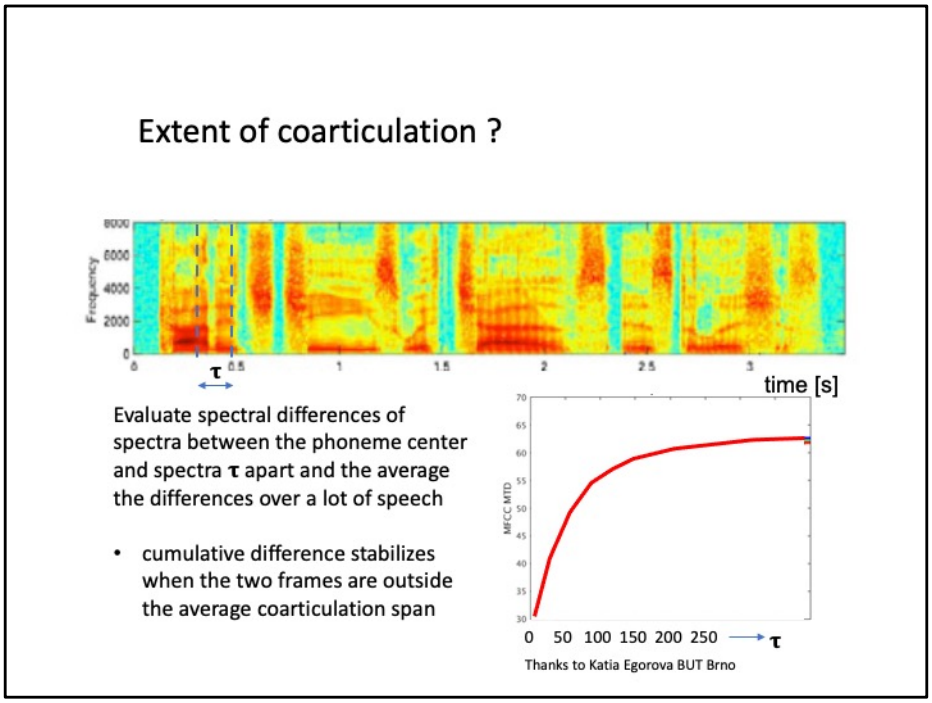
Message is carried in **changes** in vocal tract shape, which modulate spectral components of speech
Dudley 1940

26

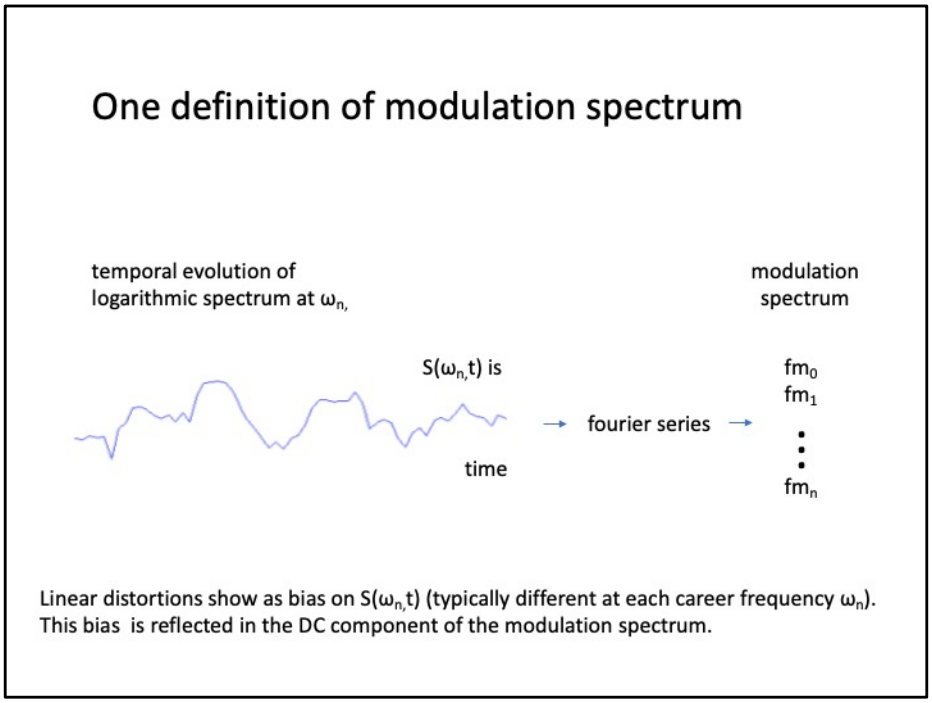
As stated earlier, is is likelu that human speech evolved to emply some properties of human hearing. It would be interesting to see what is the dominant frequency of vocal tract movements in production of speech.



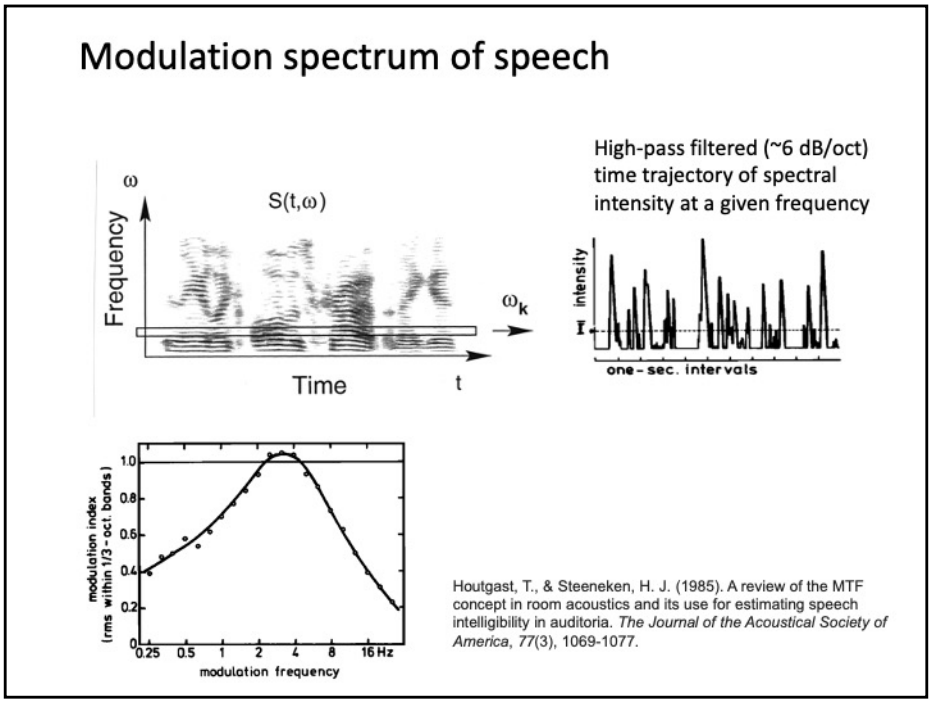
Inertia in speech production results in coarticulation of speech sounds. Coarticulation spreads lengths of the sounds into longer time spans, and can bee seen as building in temporal redundancies into the sound coding..DUe to the coarticulatiuon, individual speech sounds carry also information about the neighbouring speech sounds, making ASR based on indidependent speech sounds (most of the conventional ASR techniques) more difficult. Human hearing seem to be able to recovre the individual speech sounds from the coarticulated mixture.



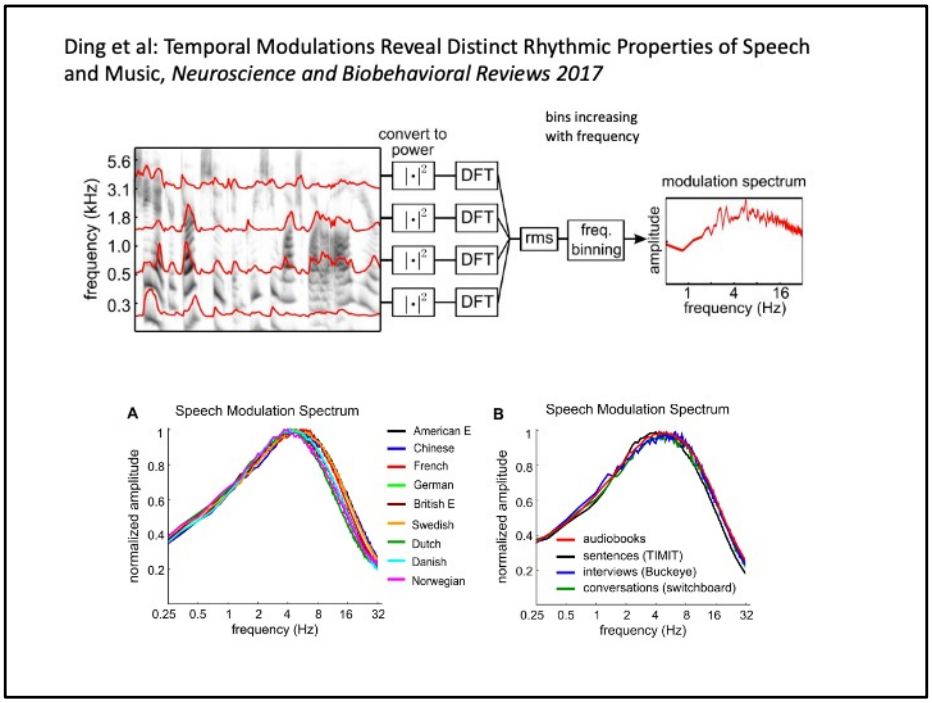
The average extent of the coarticulation can be estimated by mesuring the cummulative distances between speech features describing local acoustics, The distances typically increase with the span between the measurements and tned to saturate once the span between the measurements exceed the typical extent of the coraticulation (typically around 200-250 msec).



Spectrum of $S(\omega_n,t)$ is called the modulation spectrum. Slowly changing linear distortions in the signal show in low modulation frequency components. The dominant modulation frequencies due to speech are around 4 Hz.



Modulation spectrum of speech can be computed by computing spectrum of temporal evolution of spectral envelope at a given frequency. Modulation frequency has a typical decline at about -6 dB/oct. Therefore when computing the modulation spectrum, this decline is being compensated for by differentating the trajectories (or by weighting the modulatipn frequency components by theitr indexes). The modulation spectrum of speech peaks at 4 Hz (where the sensitivity of human hearing to modulations is highest).



This is one of the more recent results where the overal modulation frequency of speech data from different languagaes and different American English speech databases are shown. It is striking how similar are the modulation spectra from these different databases.