input ***x*** → **text-to-speech** → *symbols w*

w – string of labels (letters, words, Chinese characters,….)

need P(w!x))

Hidden Markov Model

P(w|x)) = $argmax_w$(p(x|s)P(s|w)P(w))

P(s|w) – probabiity of hidden states (speech sounds) given symbols

P(w) – prior probability of symbols

end-to-end

estimate P(w|x) directly

x – input describing the information of interest

signal → ***x*** → text-to-speech → *symbols w*

what should be the *x*

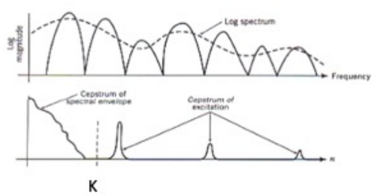
Features ***x*** (early decision making)

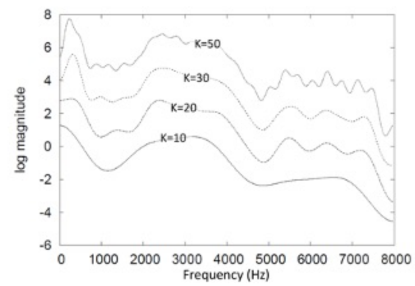
Alleviated relevant info is lost forever, irrelevant info that is left in may create problems during the inference.

~

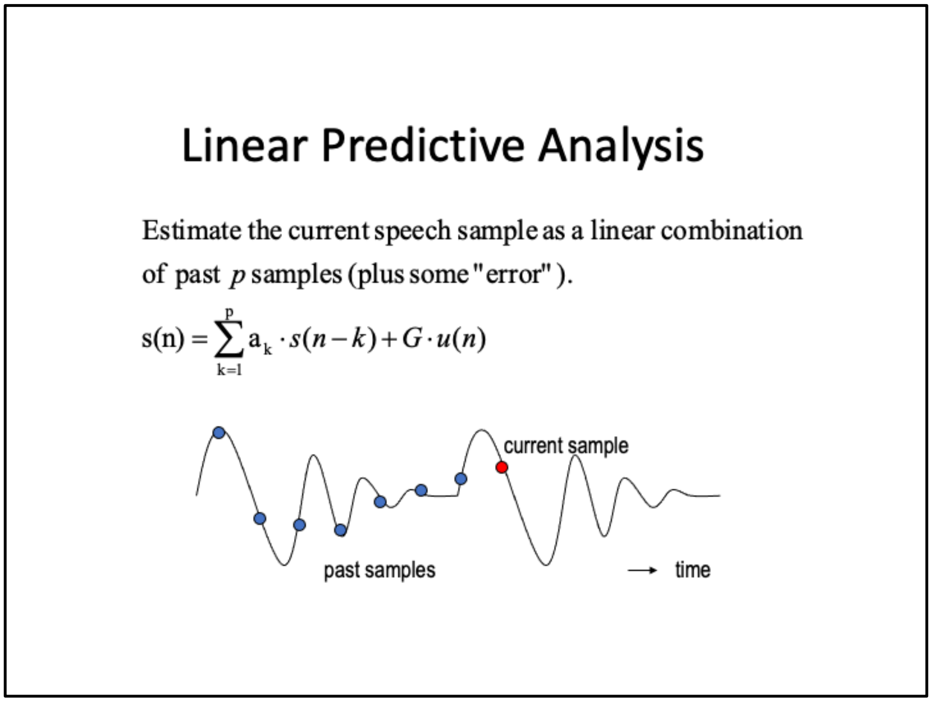
x – sequence of vectors describing evolution of envelopes of short-time spectra of speech

cepstrum
envelope by truncating Fourier expansion of logarithmic short time spectra





LPC without z-transform (for CS students)



On of the most useful techniques in speech signal processing is the linear prediction coefficients (LPC) analysis.
Its staring point is in time domain where we attempt to predict the current speech sample from the ***p*** past time samples . This involves weights of the past samples $\mathbf{a_k}$.
Sinec the perdiction is not entirely accurate, we also involve the erro of the prediction Gu(n) in the model.

PREDICTION ERROR

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k)$$

MINIMIZE ERROR OVER „SOME" INTERVAL

$$E_n = \sum_m \left[s(m) - \sum_{k=1}^{p} a_k \cdot s(m-k) \right]^2$$

$$\frac{\partial E_n}{\partial a_i} = 2 \sum_m \left[s(m) - \sum_{k=1}^{p} a_k \cdot s(m-k) \right] \cdot s(m-i) = 0$$

$$\sum_m s(m) \cdot s(m-i) = \sum_{k=1}^{p} a_k \cdot \sum_m s(m-k) \cdot s(m-i)$$

$$\sum_m s(m-i) \cdot s(m-k) = \phi(i,k)$$

$$\phi(i,0) = \sum_{k=1}^{p} a_k \cdot \phi(i,k) \quad , i=1,2,\ldots,p$$

The eror of the prediction $\mathbf{E_n}$ needs to be minimized over some particular time interval **m.** The minimzation yields a ser in linear equations which involve the prediction coefficients $\mathbf{a_k}$ and (so far unspecified functions **Φ(i.k)** with paramemets being the prediction coefficiend indexes **k** and the time signal indexes **i.**

AUTOCORRELATION METHOD

$$\phi(i,k) = \sum_m s(m-i) \cdot s(m-k)$$

ASSUME $\sum_m \rightarrow \sum_{m=-\infty}^{\infty}$

SIGNAL MUST BE FINITE
– WINDOWING !

THEN $\phi(i,k) = R(|i-k|)$

AUTOCORRELATION FUNCTION

$$R(i) = \sum_{k=1}^{p} a_k \cdot R(|i-k|), \quad i=1,2,\dots,p$$

$$R(1) = a_1 \cdot R(0) + a_2 \cdot R(1) + a_3 \cdot R(2) + \dots + a_p R(p-1)$$
$$R(2) = a_1 \cdot R(1) + a_2 \cdot R(0) + \dots + a_p \cdot R(p-2)$$
$$\vdots$$
$$R(p) = a_1 \cdot R(p-1) + a_2 \cdot R(p-2) + \dots + a_p \cdot R(0)$$

COVARIANCE METHOD

$$\phi(i,k) = \sum_m s(m-i) \cdot s(m-k)$$

ASSUME $\sum_m \rightarrow \sum_{m=0}^{N-1}$

THEN

$$C(i,k) = \sum_{m=0}^{N-1} s(m-k) \cdot s(m-i)$$

COVARIANCE FUNCTION

$$C(i,0) = \sum_{k=1}^{p} a_k \cdot C(i,k), \quad i=1,2,\dots,p$$

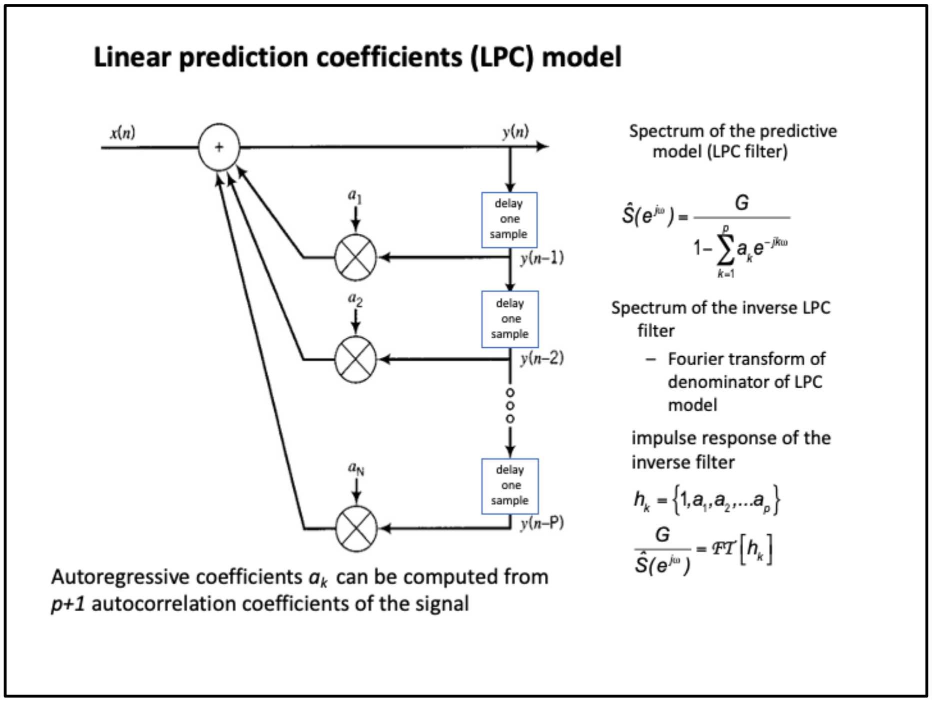
$$C(1,0) = a_1 \cdot C(1,1) + a_2 \cdot C(1,2) + \dots + a_p \cdot C(1,p)$$
$$C(2,0) = a_1 \cdot C(2,1) + a_2 \cdot C(2,2) + \dots + a_p \cdot C(2,p)$$
$$\vdots$$
$$C(p,0) = a_1 \cdot C(p,1) + a_2 \cdot C(p,2) + \dots + a_p \cdot C(p,p)$$

When the window over which the minimization of the error is done is infinite, the **Φ(i.k)** turns into the signal autocorrelation **R(|i-k|).**

This method is then called to the autocorrelation LPC method. , clearly, minimizing the error over the infinite time span is not practical (actually impossible), we need to window the signal first over the finite time window (then we not have to worry about monimizing the error over the signal which is set to zero by the window. The method yields a set on p linear equations for p unknown a_k autoregressive coefficients which specify the predictor. Methods for solving this set of equations easily do exist. The autocorrelation method yields predictors which are theoretically guaranteed to be stable (unles some quantization error causes instability). Because the windows are typically emphasizing only the center of the signal within the window, the signal which is used for the predictor design is effectively shorter.

When the window over which the minimization of the error is done is N point long, the function **Φ(i.k)** turns into the signal covariance C(i,k). **).** This method is then called to the covariance LPC method. The method again yields a set on p linear equations for p unknown a_k autoregressive coefficients which specify the predictor.

The methos direcl=tly implies the error minimizatiopn over the finite time span so it uses the signal available for the predictor design better but the stability of the predictor is not guaranteed.



LPC predictor is an infinite impulse response (IIR) recursive digital filter. Its frequency response is known. An easy way of computing the frequency response is using the fact that the inverse LPC filter, obtained by interchanging the filter numerator and denominator is the finite impulse response filter (FIR) filter. Its impulse response $\mathbf{h_k}$ is given by the autoregressive coefficients and the frequency response of the inverse filter can be computed through fourier transform. The frequency response of the original IIR filter i=can be obtained from the inverse of the frequency response of this inverse filter.

energy of prediction error :

sum errors in all (windowed) signal samples over all times

$$E = \sum_{n=-\infty}^{\infty} \varepsilon^2(n) = \sum_{n=-\infty}^{\infty} [s(n) - \sum_{k=1}^{p} a_k \cdot s(n-k)]^2$$

The error energy in the analysis window is obtained by summing the error over the finite window *m* (which is zero beyond its boundaries)

$$E = \sum_{i=0}^{m} \varepsilon^2[n]$$

Parseval's theorem: energy in time and frequency domain are equal, therefore the error energy can be also obtained by summation of spectral energy of the error in frequency domain

$$E = \sum_{-\infty}^{\infty} \varepsilon^2(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} |E(e^{j\omega})|^2 \, d\omega$$

$$E(e^{j\omega}) = S(e^{j\omega}) / \hat{S}(e^{j\omega}) = S(\omega) / \hat{S}(\omega)$$

$S(\omega)$ – spectrum of the signal

$\hat{S}(\omega)$ – spectrum of the LPC filter

Therefore, equivalently – integrate product of signal power spectrum and inverse filter power spectrum over all frequencies

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \left[\frac{S(\omega)}{\hat{S}(\omega)} \right]^2 d\omega = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{P(\omega)}{\hat{P}(\omega)} d\omega$$

where $P(\omega)$ is power spectrum of the signal and $\widehat{P(\omega)}$ is power spectrum of the LPC model

Wiener-Khinichin theorem says that
autocorrelation of the signal R(n) is given by Fourier transform of its powers spectrum P(ω)

Autocorrelation LPC method can be fully specified in frequency domain (no need to ever see the signal)

Energy in the prediction errror is theoretically computed by summing the local prediction error over all times. However, since the signal window in the autocorrelation method is finite, the infine summation turns into the finitine one. Parseval theorem allows to equal the error enerhy in the time domain and the total error energy in the frequency domain. Energy of LPC error in spectral domain is abtained by dividing the spectrum of the signal by the spectrum of the LPC model, this shows how well the model spectrum approximated the signal spectrum at all frequencies. The total energy in the frequency domain is given by integrating the spectral energy contributions in different frequency points over the whole frequency range. The local contributions in the frequency domain can be obtained as a ratio of the signal spectrum and the error spectrum at a given frequency. Inseriting the expression of the local error to the integral used in the global energu computation shows how different parts of the signal are bing approximated by the all-pole model. Due to the spectral ratio in the integrand, the peaks of the signal spectrum are fitted by the model spectrum better than the dips.

Since the autocorrelation of random signals is given by fourier transform of its power spectrum, the autocorrelation can be computed even without any access to the time domain signal. DO the LPC model which fits the power spectrum of the signal can be computed directly from this signal power spectrum.

SPECTRAL FORMULATION OF LPC

AUTOCORRELATION

– INVERSE FOURIER TRANSFORM OF POWER SPECTRUM

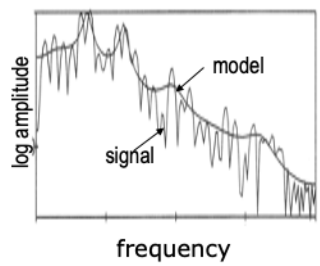
FIND ALL-POLE MODEL WITH SPECTRUM $\hat{S}(e^{i\omega})$ WHICH MINIMISES

$$E = \frac{G^2}{2\pi}\int_{\omega=-\pi}^{\pi} \frac{|S(e^{i\omega})|^2}{|\hat{S}(e^{i\omega})|^2}\, d\omega$$

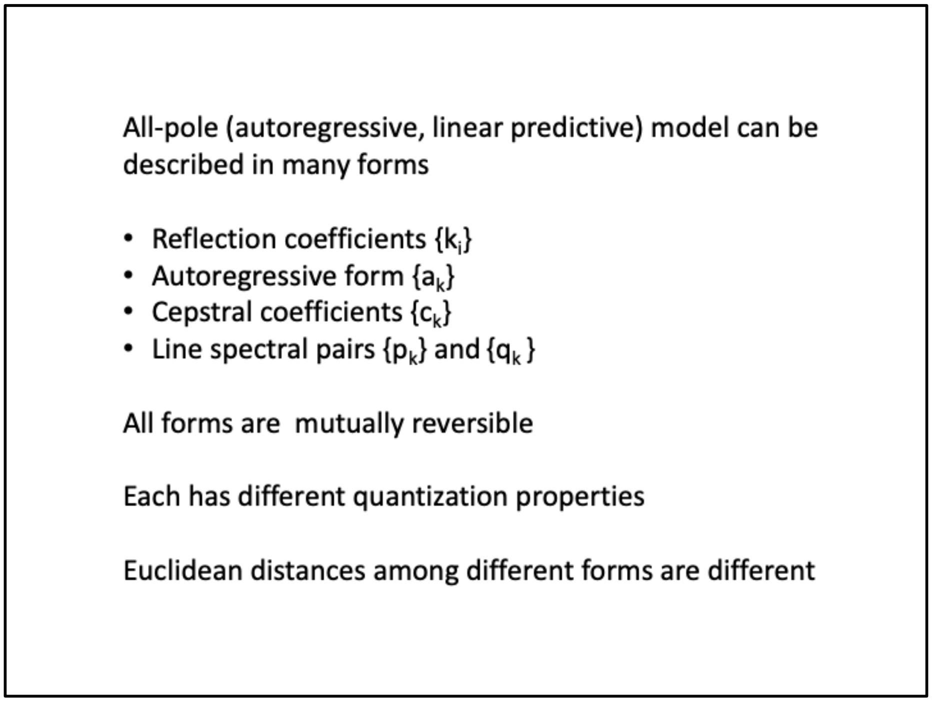
$S(e^{i\omega})$ – SIGNAL SPECTRUM

$\hat{S}(e^{i\omega})$ – MODEL SPECTRUM

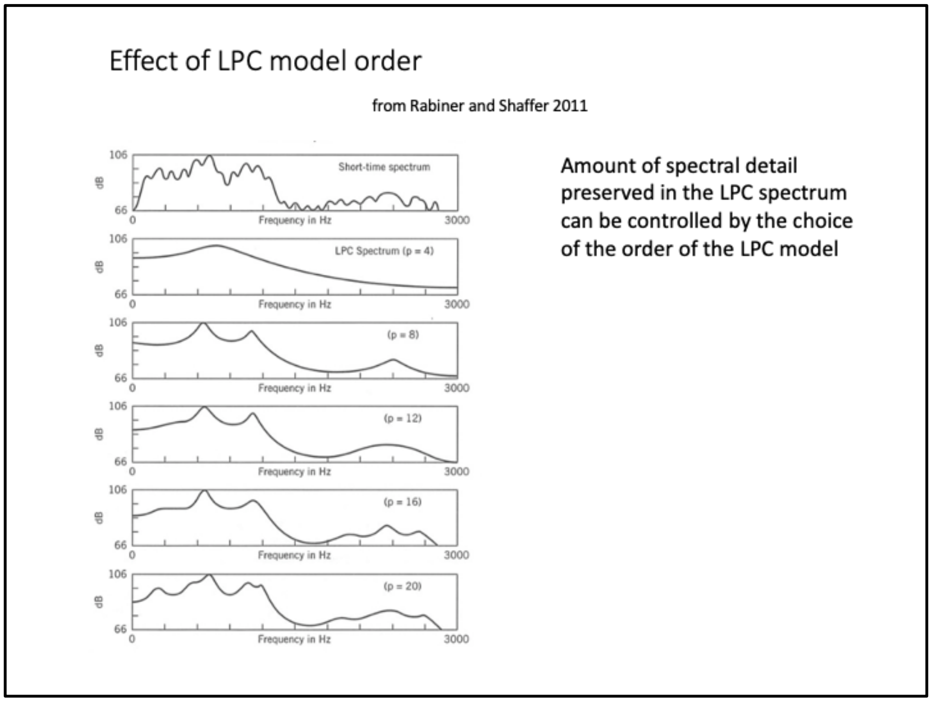
- at frequencies where the signal spectrum is larger than the model spectrum, the contribution to the total error is larger
- models fits spectral peaks rather than spectral dips



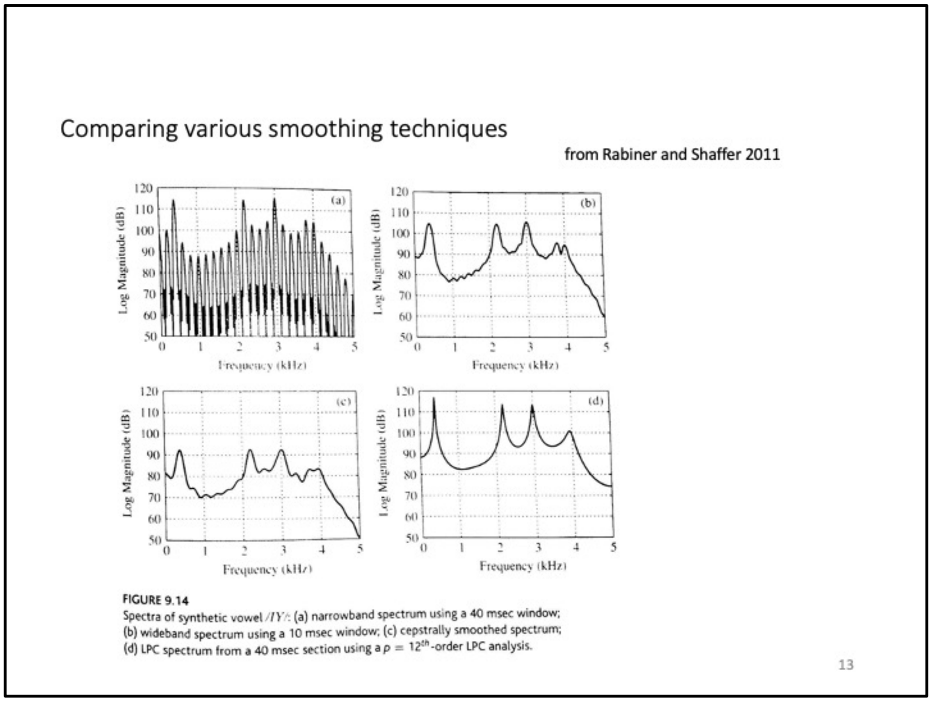
Since the autocorrelation of random signals is given by fourier transform of its power spectrum, the autocorrelation can be computed even without any access to the time domain signal. DO the LPC model which fits the power spectrum of the signal can be computed directly from this signal power spectrum.



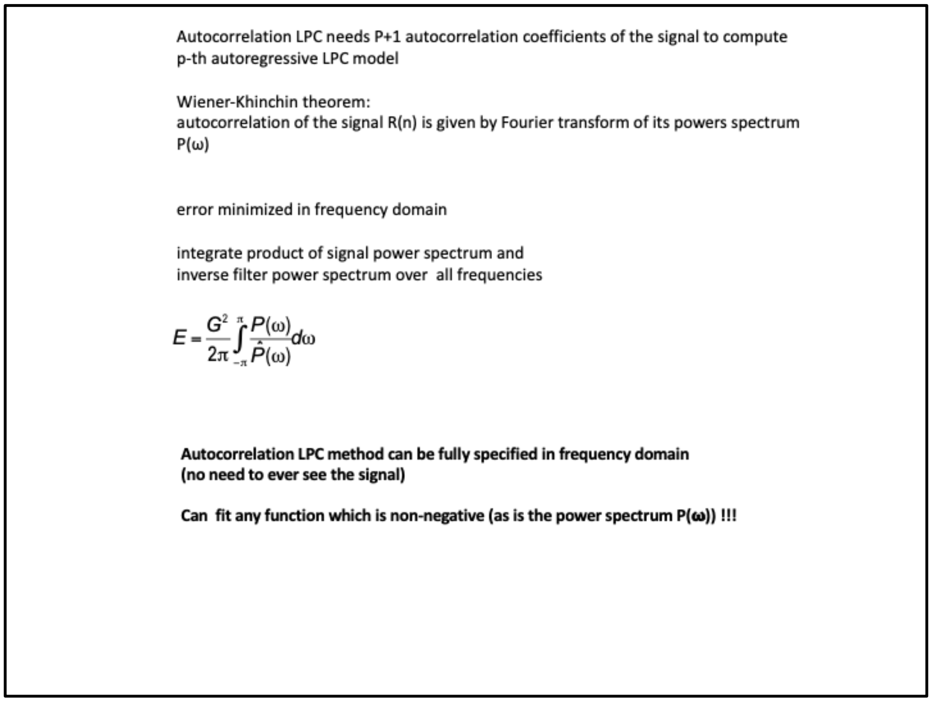
WIthout any further proofs (which can be found in any speech signal processin literature) we mention that many alternative ways of describing the LPC filter exist. The reflection coefficients are convenient since they allow for an easy check of the filter stability (they need to all be less than one for the stable filter). Cepstral coefficienns are less correlated that the other reprentations. Line spectral pairs relate to frequencies and bandwidths of spectral peaks in the model are good for quantication in speech coding.



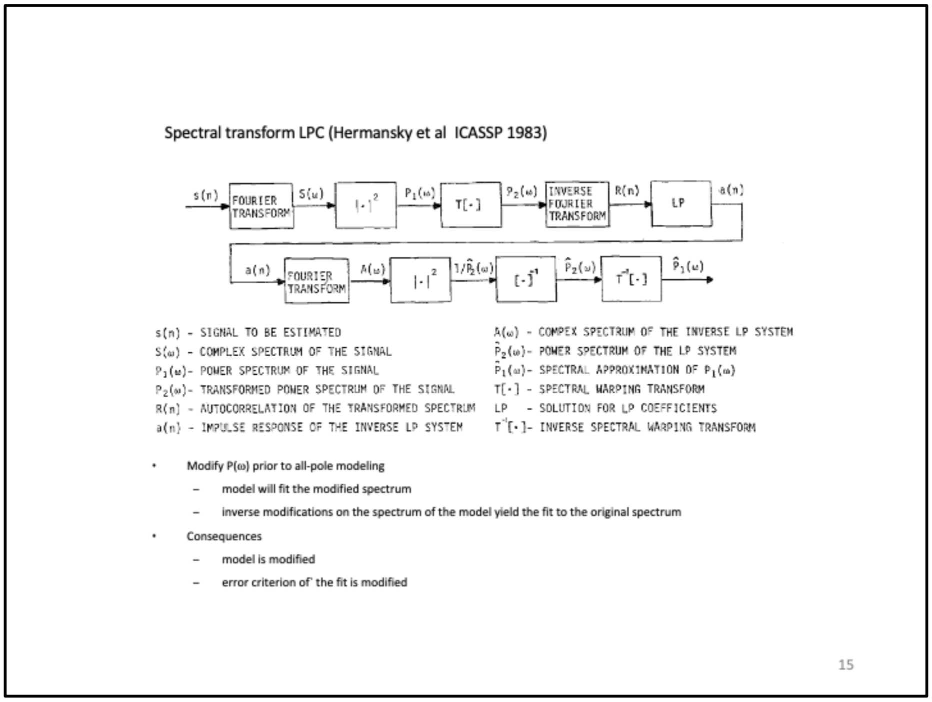
LPC predictor fits the speech power spectrum. The eroror of the fit (and the amout of spectral smoothing) can be controlled by the choice of the model order.



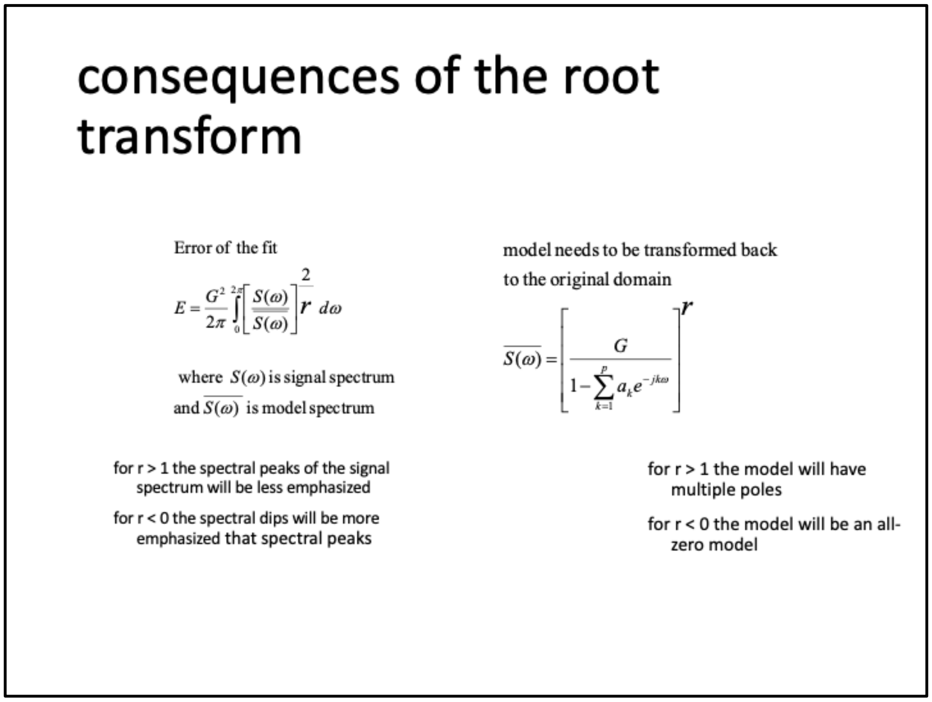
SIgnal spectrum (a) can be approximate by several different methods. A strahtforward way of describing the spectrum with less spectral details is using shorter analysis windw which yields lower spectral resolution (b). The truncation of the signal; cepstrum can be also used (c). The LPC techniques is yetr another way of douns spectral smoothing (d)



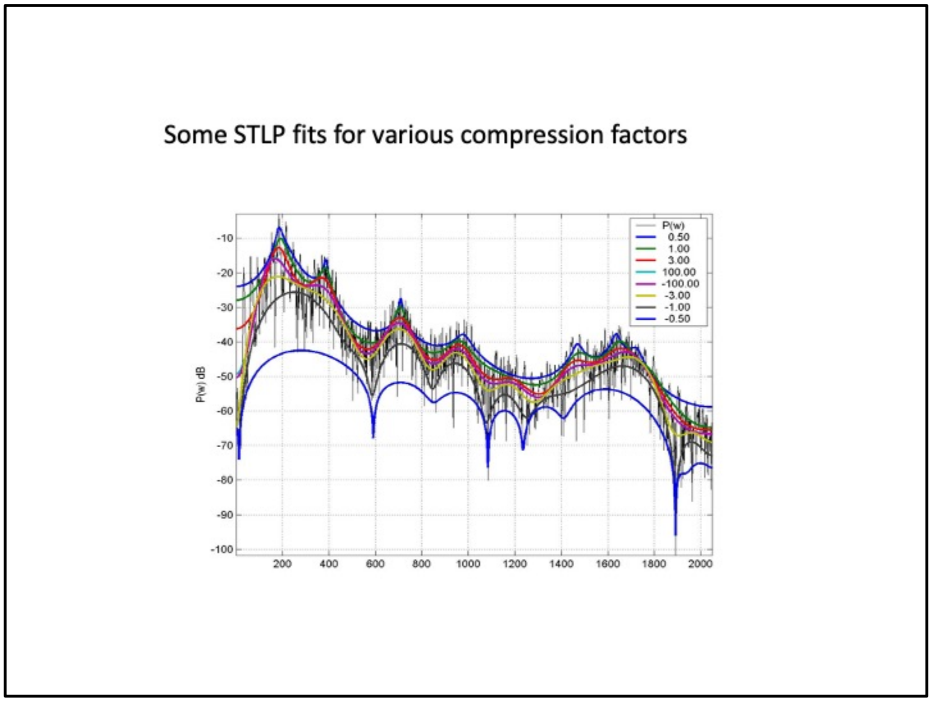
This summarizes the spectral method of LPC. It says that the assumed power spectrum does not need to be a spectrum of any existing signal. The spectral LPC will fit any finction which is non-negative.



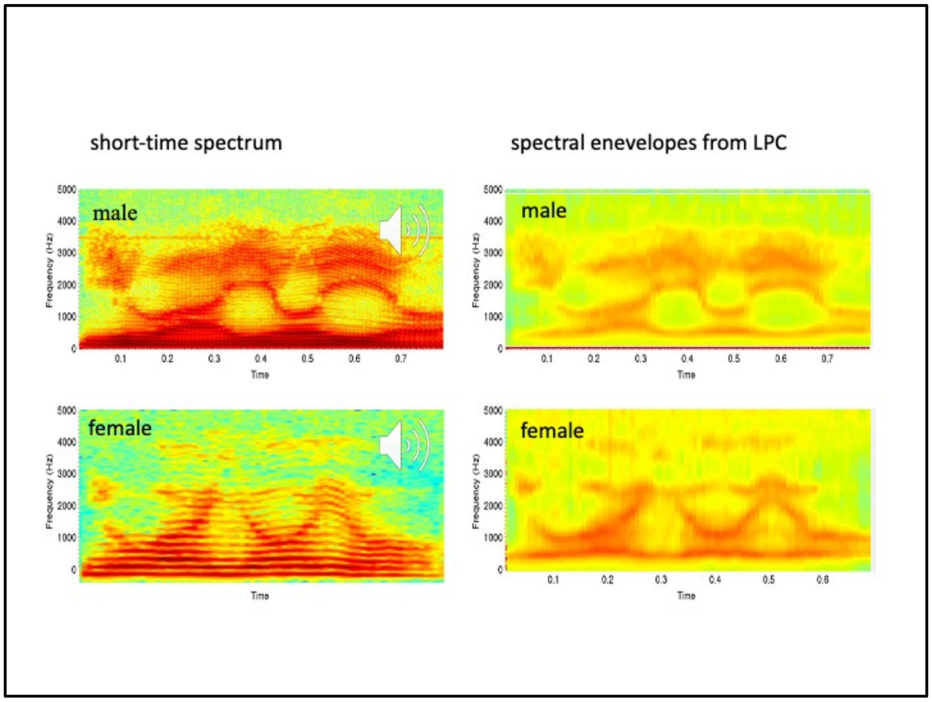
The spectral transform LPC modifies the signal spectrum prior to fitting it with the spectrum of the LPC model. One interesting modification is to take a root of the power spectral values. This compresses the spectrum and allows for different spectral approximation. When the fit to the original spectrum is required, taking the power (inverse of the root function) can be applied.



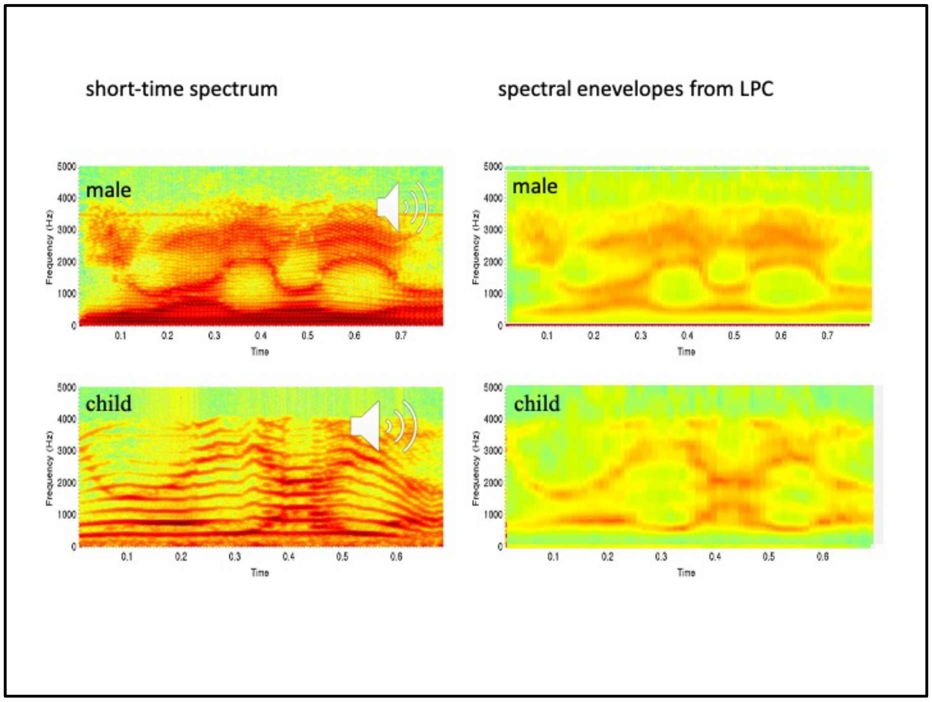
The fit to spectral peaks is emphasizer for roots r>1 and deemphasized ro r <1. Forr r < 0 the model firs the spectral dips better than spectral peaks.



Examples of some root spectral transform LPC. ***For r >1*** the *peaks of the signal spectrum* are approximated more closely than for the conventional LPS (with r=1) When the r < 1, the dominance of spectral peaks is gradually diminished, where for very large ***r,*** the peaks and troughs of the signal spectrum are almost equally important.For ***r < 0,*** the troughs become more important in the model approximation than peaks. Thus, the spectral transform LPC offers considerable flexibility in how the model fit behaves. The conventional LPC is a special case of the spectral transform (root) LPC for ***r=1.***

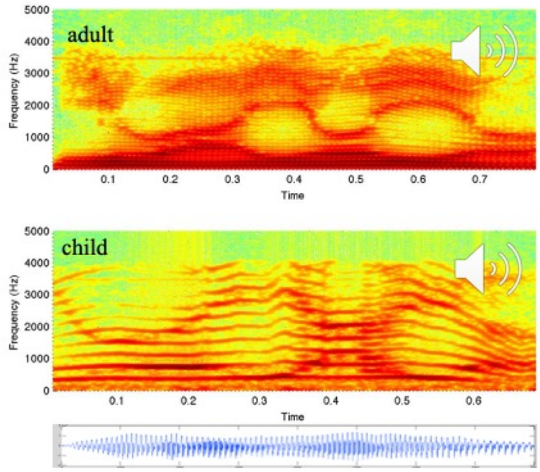


Estimating spectral envelope (here using LPC technique) alleviates differences due to fine spectral structutere (spectrum of the voice source). The spectral envelopes from LPC lokk more similar for both male and female spakers, although some differences in estimated vocal tract reconance frequencies remain.



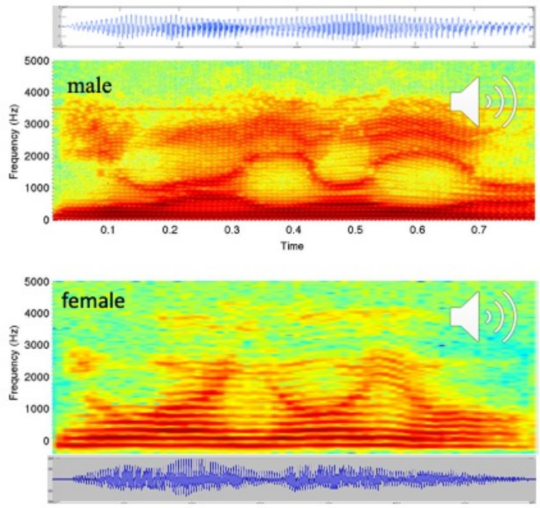
Seeking similarities between male spoeakers and small chilren (here the example is speech of 4 year child) is more difficult. Even when the fine spectral structure is alleviated, differences in spectral envelopes are much more significant. The child hase only two reconance frequencies (formants) in the spectral span of the adult (here 0-4 kHz). The first formant of the child is often at the position of the second formant of the adult and the second formant is as high as the fourth formant of the adult. So far for the formant pattern as carrier of phonetic valus of speech sounds ☺.

Spectrogram from short-time fourier transform



Spectrum from Fourier transform has equal spectral resolution at all frequencies

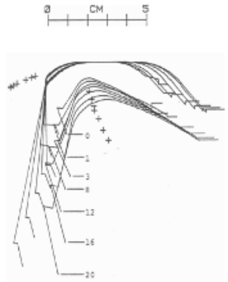
Spectrogram from short-time fourier transform



Spectrum from Fourier transform has equal spectral resolution at all frequencies

Perceptual Analysis

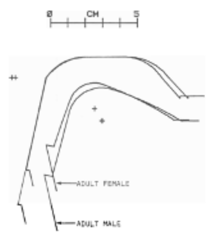
Vocal tract growth with age



	pharynx	oral
3 years	44 mm	61 mm
21years	89 mm	81 mm
	102 %	33 %

between ages 3 and 20
pharyx grows 3 times
more than mouth

Male and female vocal tracts

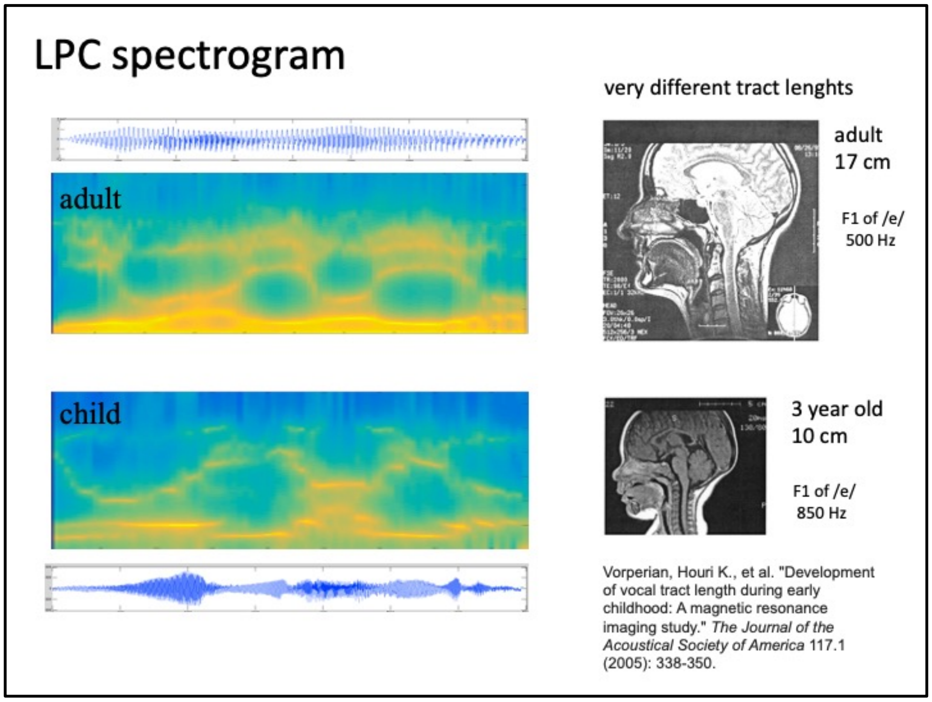


	pharynx	oral
adult male	89 mm	81 mm
adult female	63 mm	78 mm
	41 %	4 %

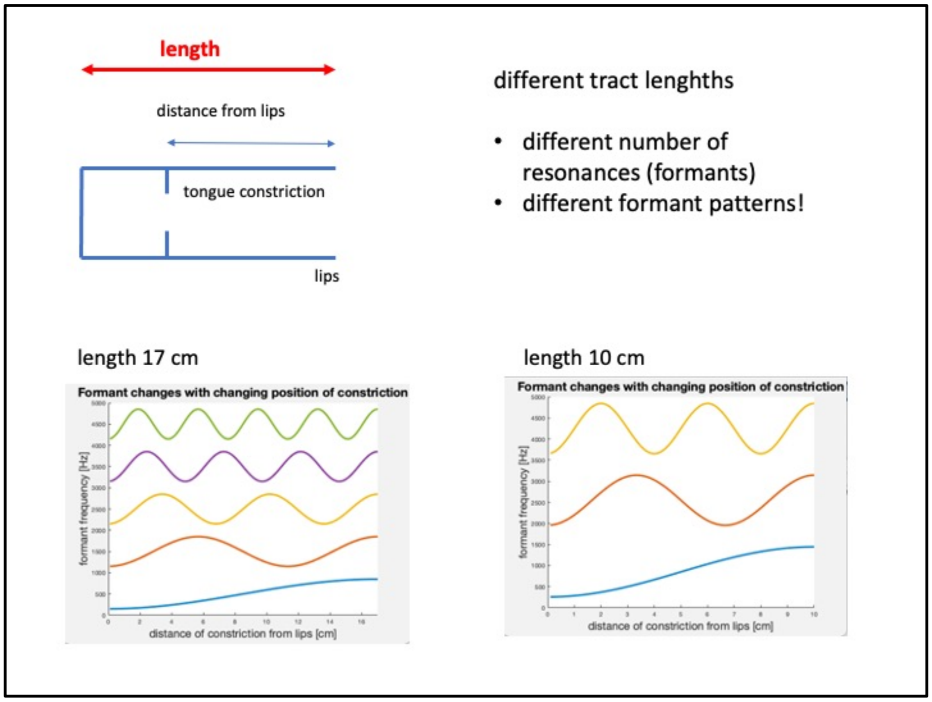
Ursula Goldstein, MIT PhD. Thesis 1980

Back part of the vocal tract grows three times more with age that does the fron part. So it has much more sense that as children learn to speak, they learn how to correctly form the front cavity of the vocal tract. In general, the back cavity is much more difficult to control anyways. Only actors may learn how to do it when they need to emulate different personalities.

The most significant differences between male and female vocal tract lenghts are in the back (pharyngeal) part of the vocal tract.

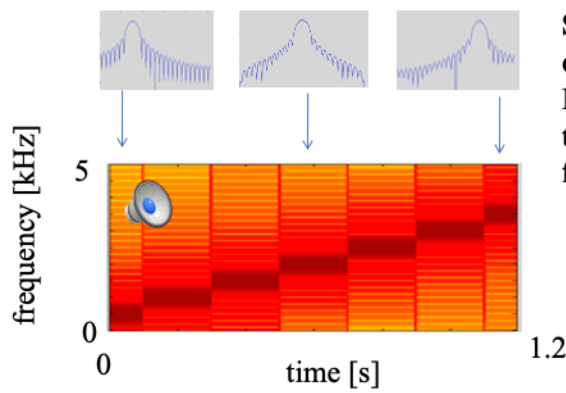


The formant structure of the child is understood from ver known anatomical differences between chldren and adults. WHile the typical tract length of adul males is arounc 17 cm, the length of tract of small chldren is significantly shorter. The first formant of the neutral vowel /e/ in adult make is F1 = c/4 l= 340/68=500 Hz. (c- speech of sound, l- tract length) , the F2 is then et 1500 Hz. The F1 of child's e is F1=340/80= 850 Hz and F2 is 3x850 = 2550 Hz (that is where the F3 of the adult is). No surprize the formant patterna are so different. It is amazing that the concept of formant patterns as carriers of linguistic messages lasted that long.



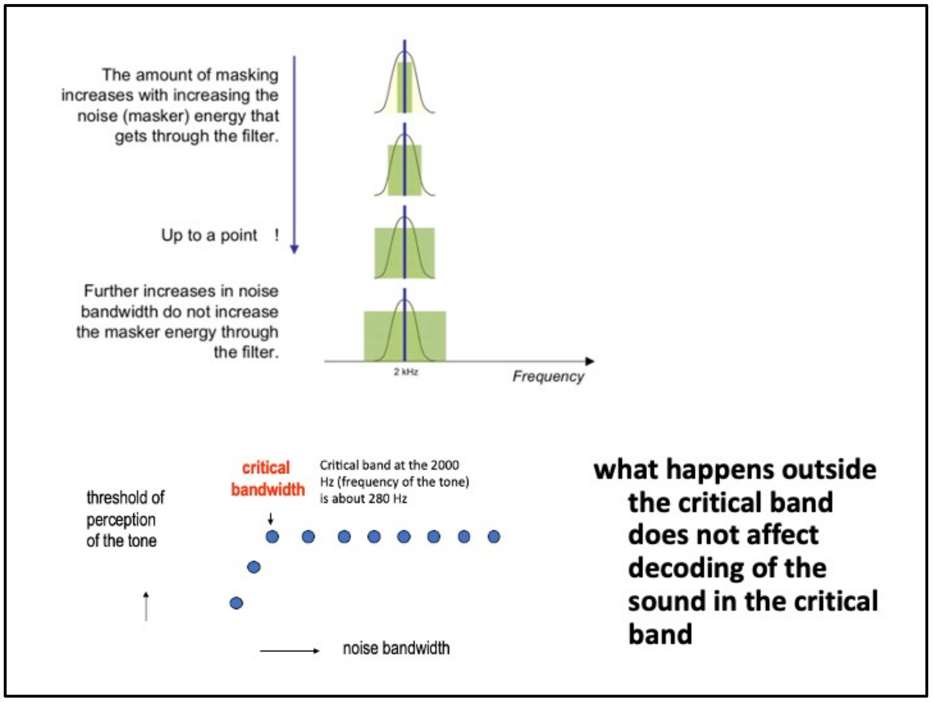
Remembering our brief diacussion about speech production, here we see what can adults and chilren produce by moving tongue in their vocal tracts.

One obvious inconsistency between Fourier analysis and spectral recolution of human hearing



Spectral resolution of the short-term Fourier analysis is the same at all frequencies.

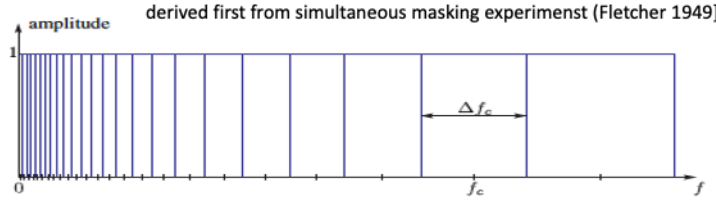
One thing to remember is that the frequency resolution of a short-term Fourier analysis is the same at all frequencies, and it is given by the length of the analysis window. This is one obvous inconsitence between human hearing and the Fourier analysis, and may be one source of problems of Fourier spectral analysis for speech recogntion.



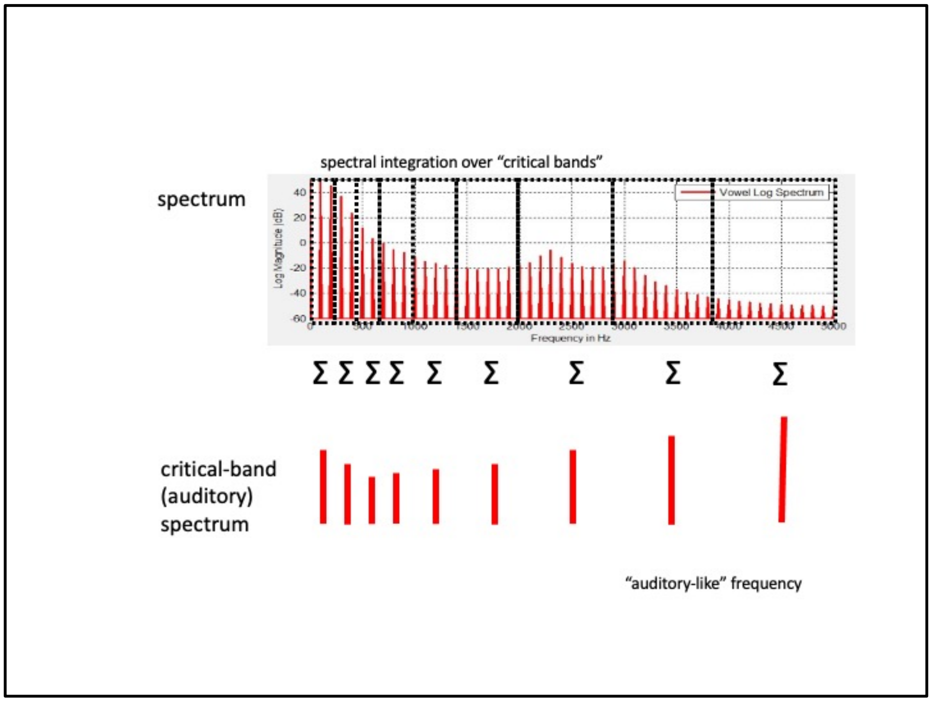
One way to imagine what is happeningas as the bandwidth of the basker signal increases is shown here. Once the masker bandwidth is wider than the bandwidth of the hypothetical cochlear filter, it does not contribute to masking if the signal within the filter.

Spectral resolution (critical bands) of hearing

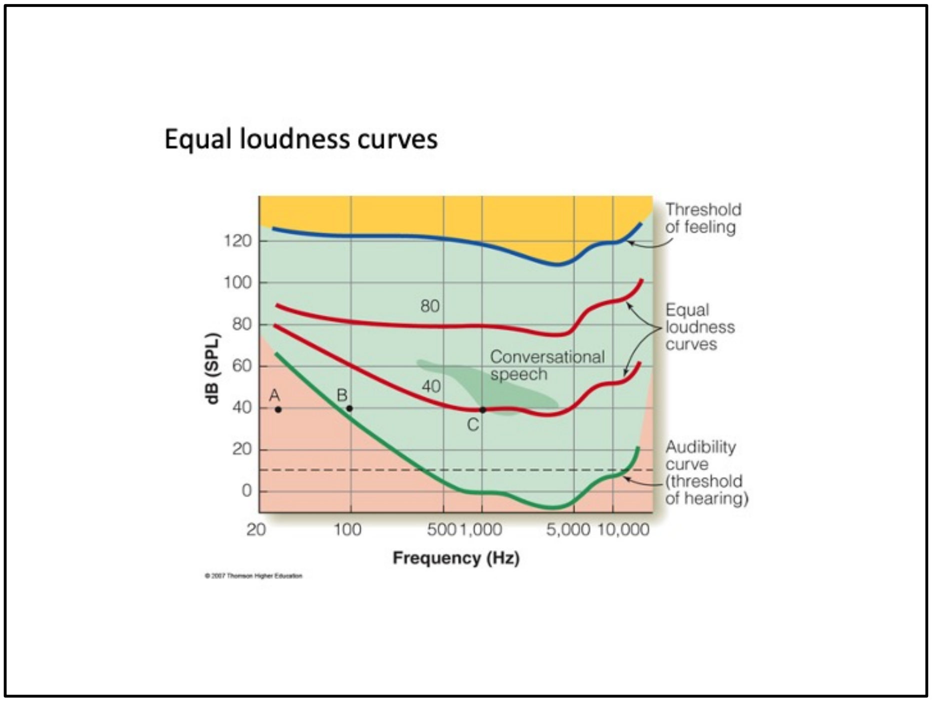
derived first from simultaneous masking experimenst (Fletcher 1949]



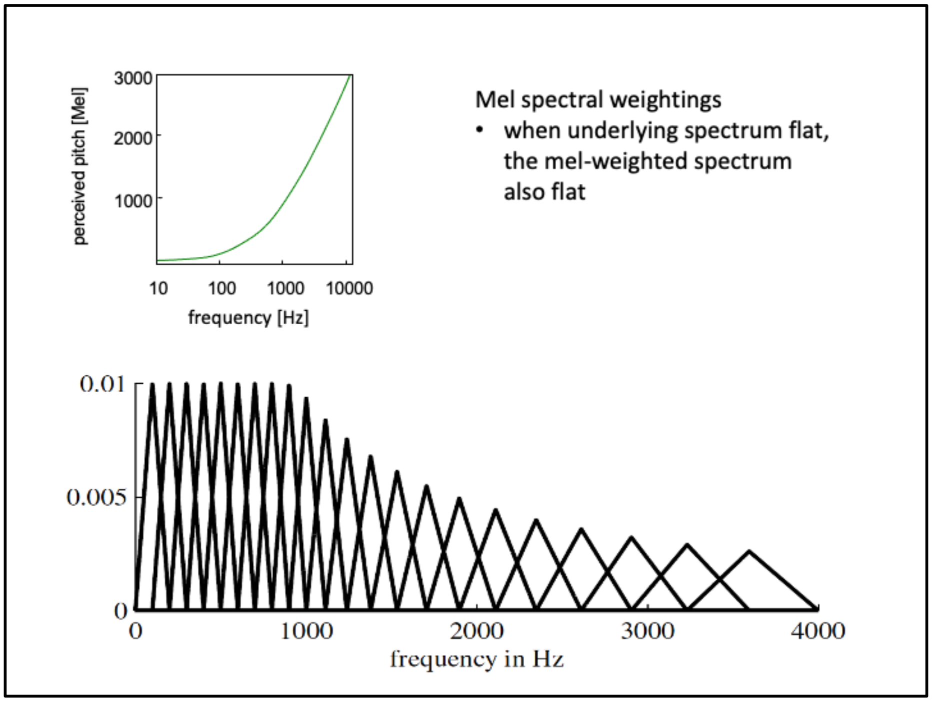
SPectral resolution of human hearing is decreasing towards higher frequencies. We have a lot of evidence for it.



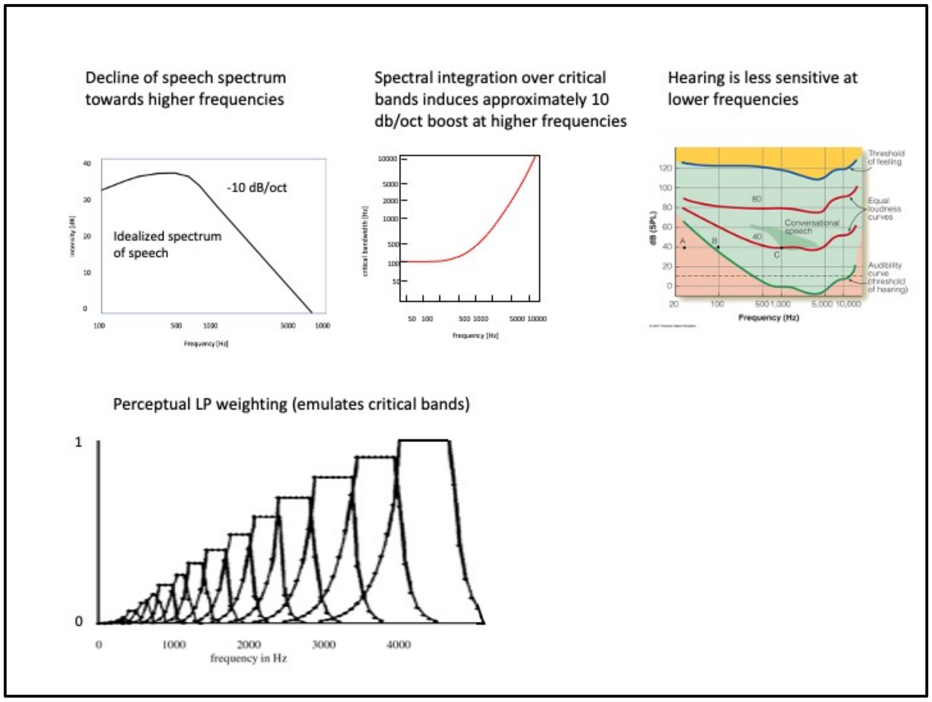
We can change the spectral resolution of Fourier analysis quite simply by integrating Fourier spectrum ofver windows with varying spectral widths.



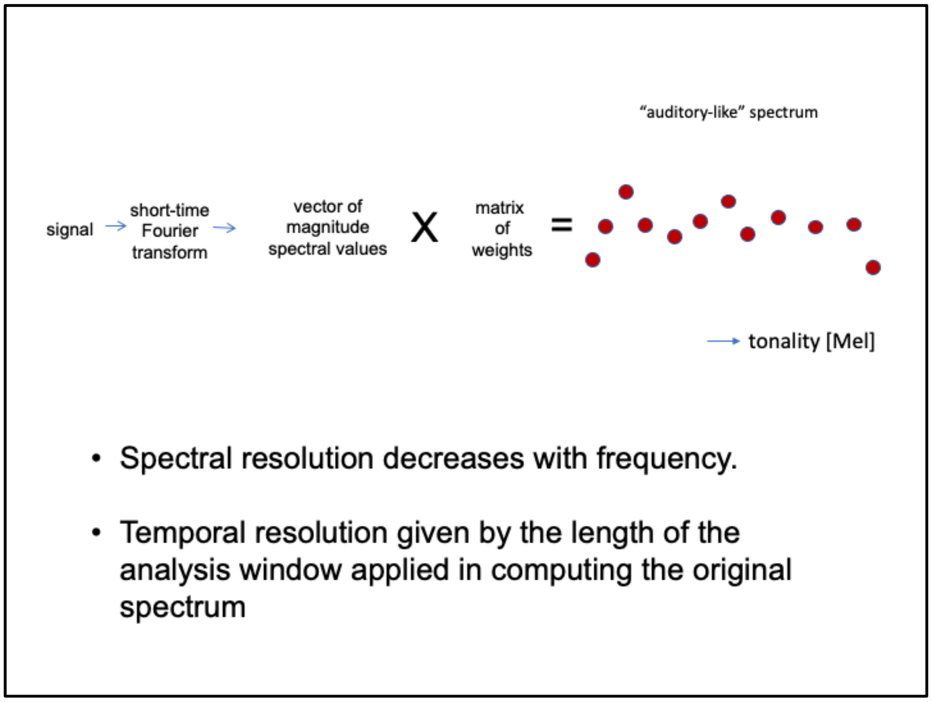
From the perceptual equal lougness curves, we can see that the spectral energies are attenuated at low frequencies below 600 Hz.



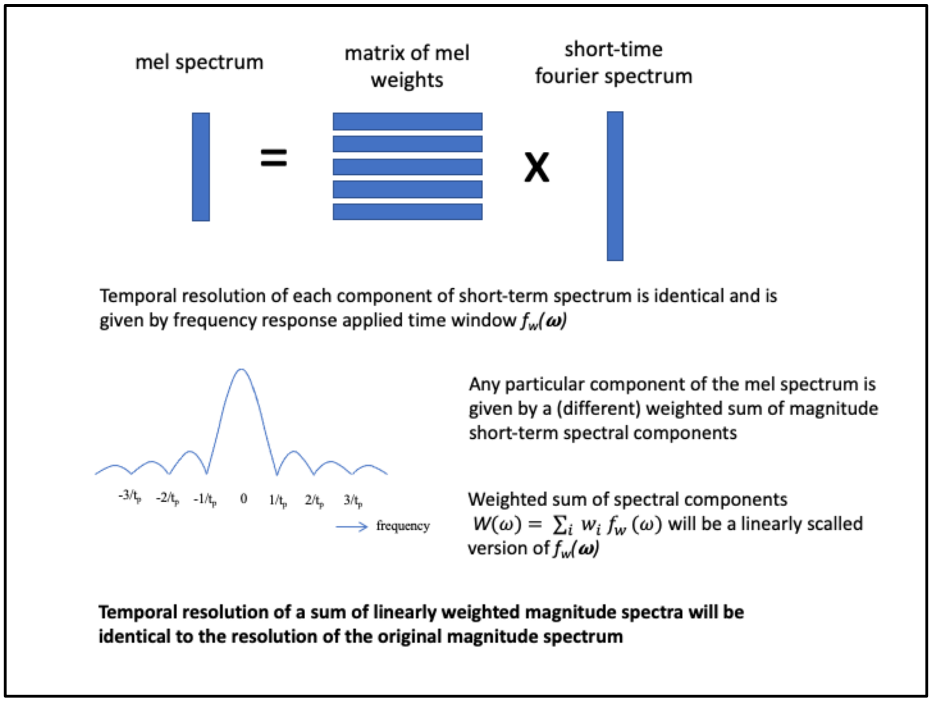
Typicall mel spectral weighting often compensates for the increasing auditory filter with by attenuating the autputs from higher frequency filters. The intention there is to make sure that the while noise spectrum (noise which hase the same energy at all frequencies) remains while after critical band waightinIt preserves the white spectrum equal amplitudes. This compensates for the increasing auditory filter with by attenuating the autputs from higher frequency filters. The intention there is to make sure that the while noise spectrum (noise which hase the same energy at all frequencies) remains while after critical band waighting. This startegy can be questioned.



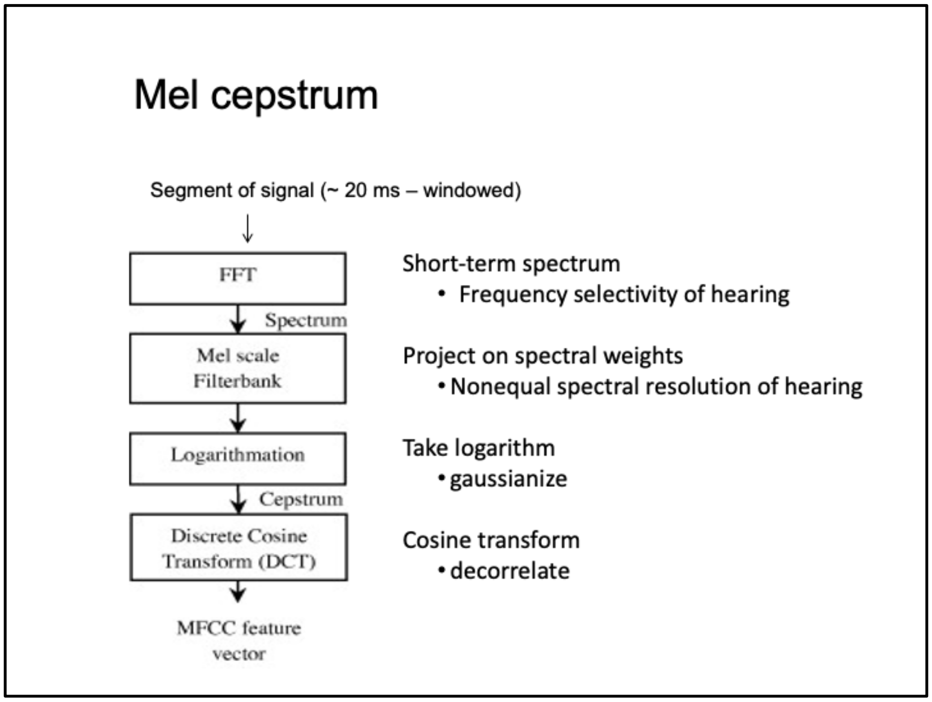
Putting it all together, one may argue that the low frequency spectral energies should be attenuated and the typical mel-frequency weighting may be inconsistent with human hearing. The spectral weighting is PLP models the increasing hearing sensitivity towards higher frequencies by incresing bandwidths of critical band filters any by explicitly including the equal-loundness curve (modeled as 60 dB SPL fixed curve)



Here we see the whole process of mofication of the spectral resolution of Fourier analysis. Substittute your favorite spectral weightingsm here we show the typical way of computing the met spectrum, most often used in ASR. The "auditory-like" sopectrums is typically processed further by some additional transformations.



Linear filter-bank with decreasing spectrl resolution towards higher frequencies woul haveincreasing temporal resultion towards higher frequencies (remember the uncertainty principle in spectral analysis). However, what we are dealing with here is not the linear filtering but the integration of maginitde spectra (while not using spectral phase at all). The temporal resolution of such analysis is fixed at all frequencies and is given by the window length in the Fourier analysis used in deriving the magnitude spectrum. Check it yourself if you want to prove me wrong,

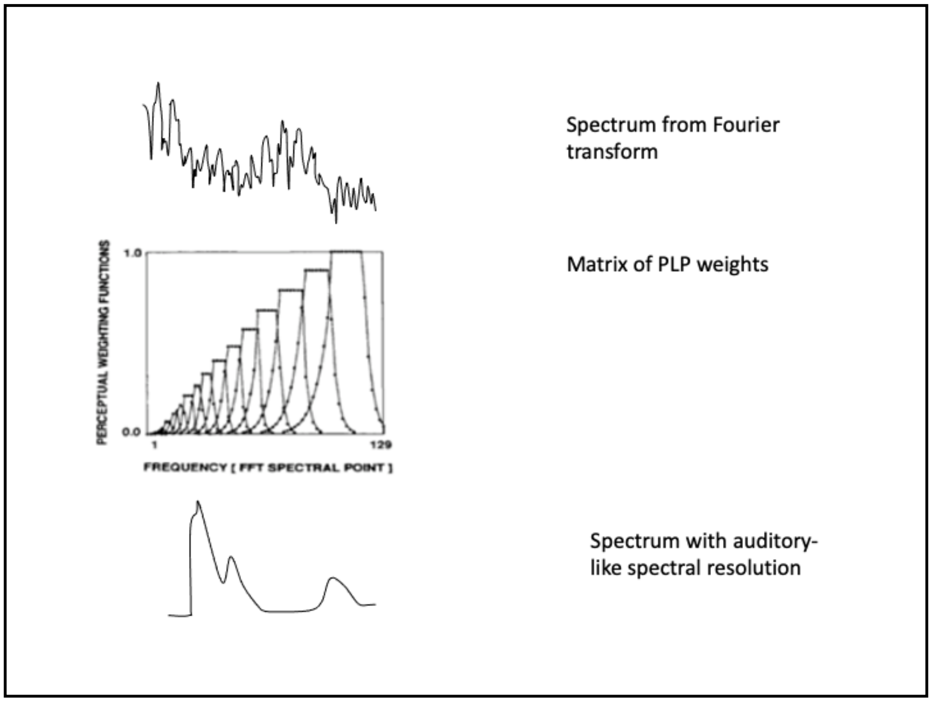


Here we see the block diagram of the whole process of computing the mel cepstrum.

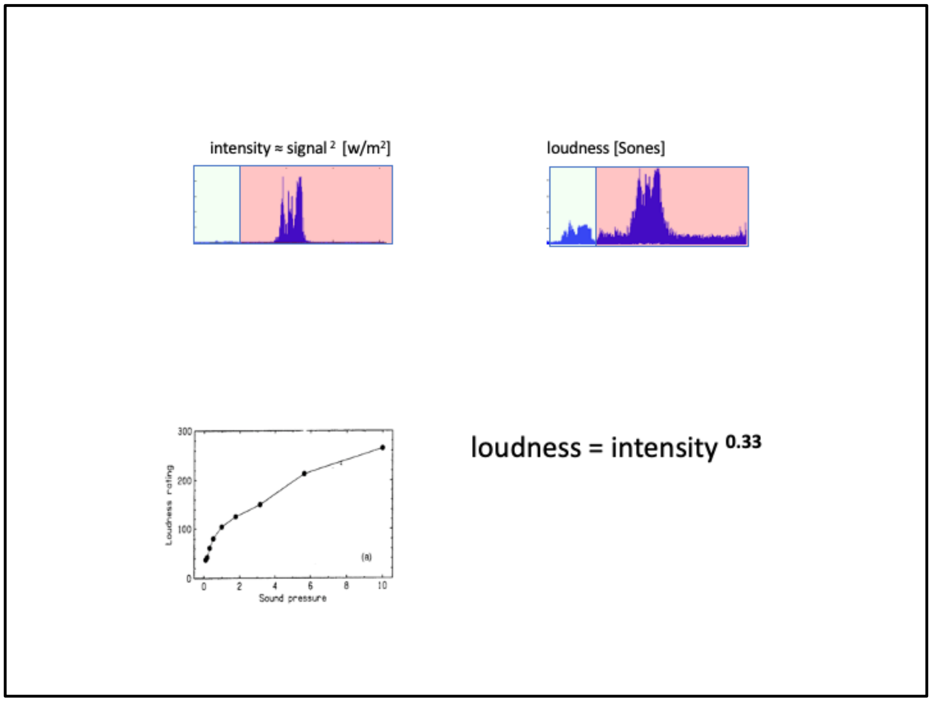
Perceptual Linear Prediction (PLP)

A simple auditory model models some basic properties of human hearing

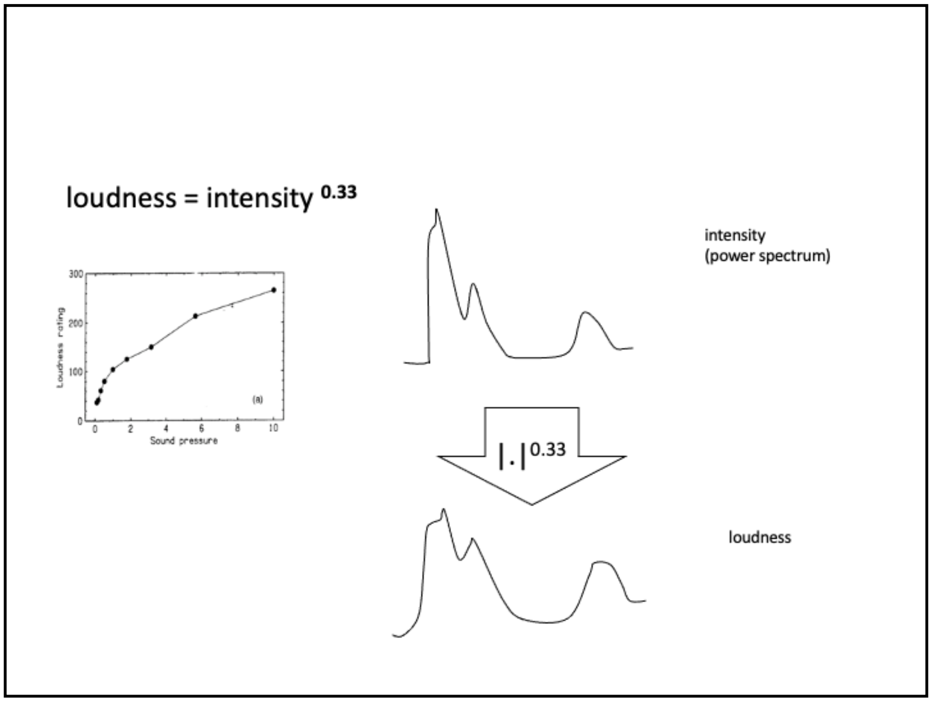
- Critical-band spectral resolution
- From sound intesity to loudnes domain
- Fixed equal loudness-like hearing sensitivily supresses low frequency spectral energies
- Selective spectral smoothing by autoregressive model



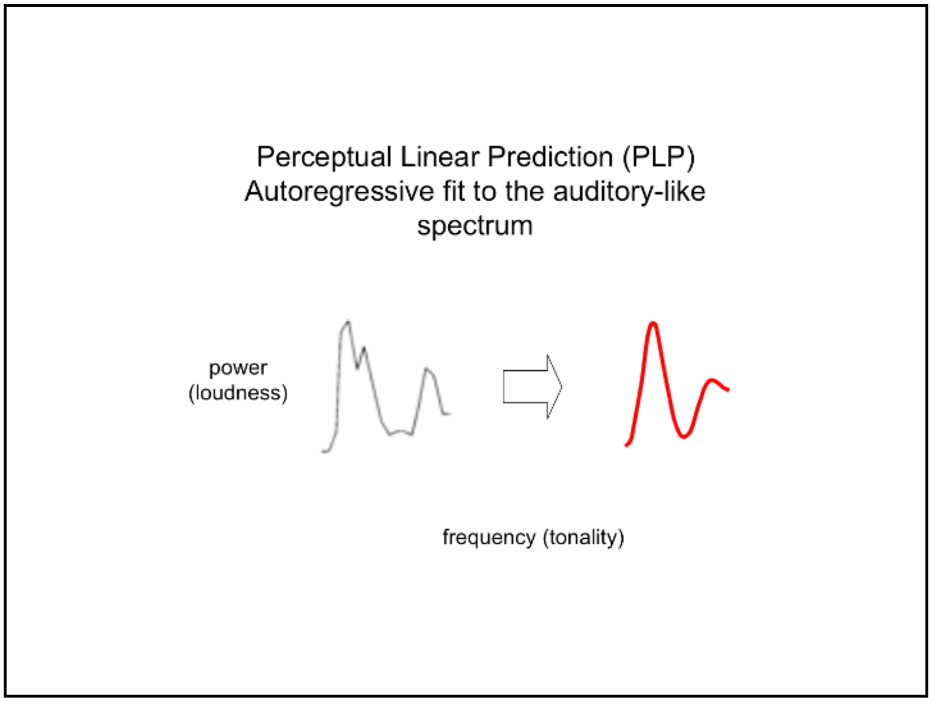
The PLP auditory-like spectrum is then formed by taking the short-time spectrum from shot=rt-time Fourier transform and multilpying is with the PLP spectral weighting which combines the effects of critical=-band integration and the equal loudness curve. This operation results in significantly smoothed “audotory-lile” spectrum and shown here.



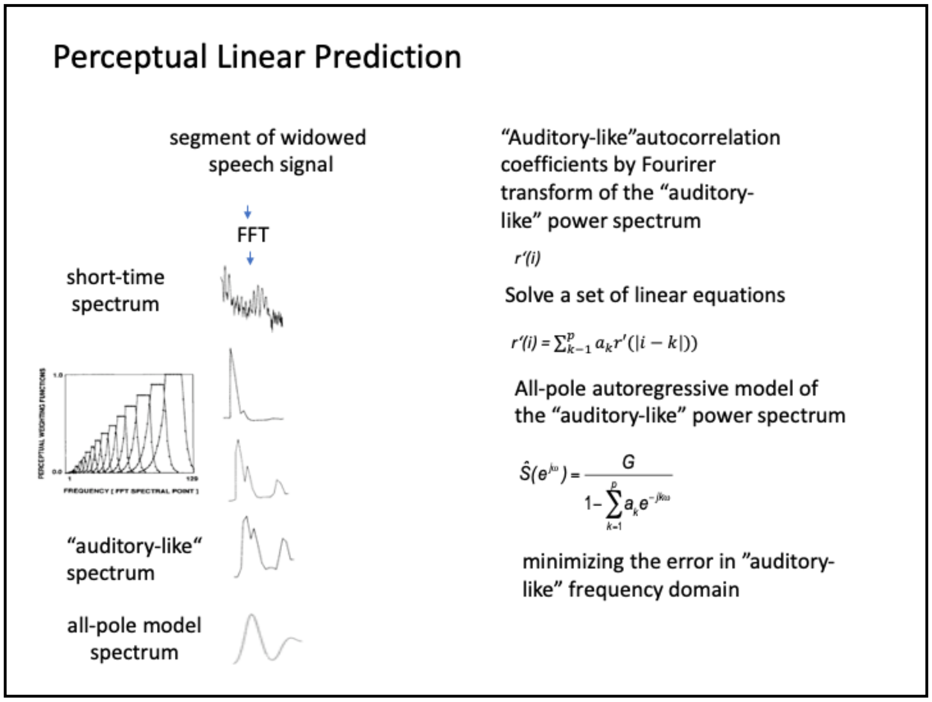
The operation in PLP is converting spectral intensities to loudnesses in the individual frequency bands. THis iperation is justified by experiments in lougness sumation.



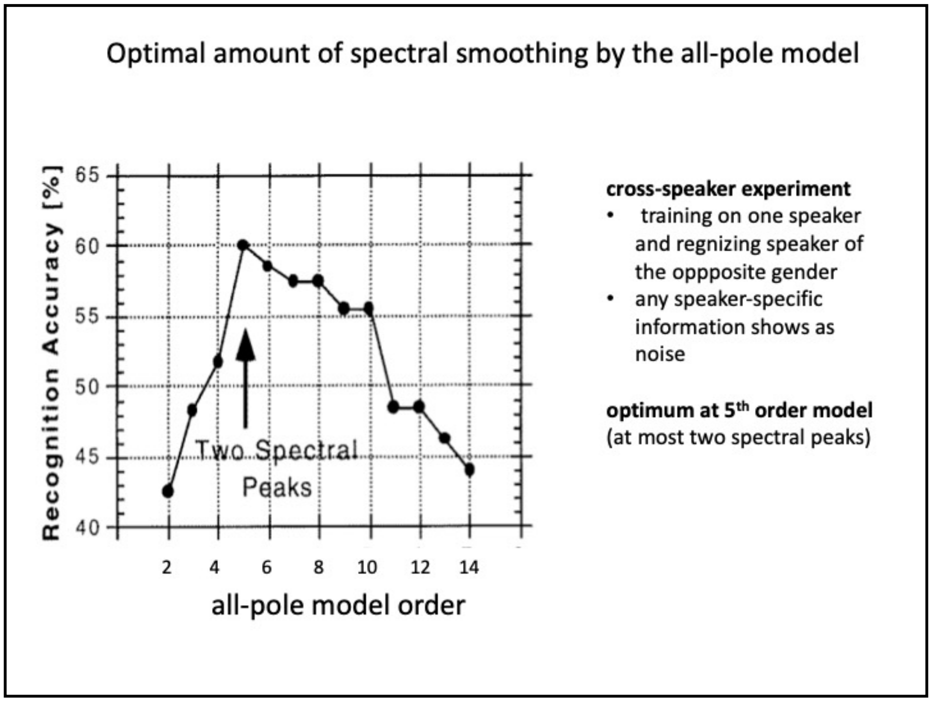
The operation in PLP is converting spectral intensities to loudnesses in the individual frequency bands. THis iperation is justified by experiments in lougness sumation. (S.S. Stevens- power law of human hearing)



The all-pole approximation of the auditory-like spectrum allow for alleviating minor spectral components from the spectrum.



THis schematic diagram summarizes the whole PLP analysis. The short-time magnitude spectrum is computed from a short segment of windowed speech signal. The spectrum is transformed into auditory-like domain thropugh multiplication owith the PLP spectral weighting matrix, which emulates the combined effects of the critical-band weighting and the fixed equeal loundess at 60 dB SPL and power spectrum is computed. Such weghted spectrum is transformed to loudness domai through the cubic root nonlinearity. The loudness spectrum is approximated by the spectrum of the all-pole model using the spectral LPC.



For alleviating the speaker-specific information, some additional spectral smooting of the auditory-like spectrum may be required. The amount of smooting can be derived by speech recogntion experimenst, where the training of the system (spectral templates) are provided by one speaker and the test speech comes from another speakr of the oposite gender. In this experiment, any speaker-specific information representes the unwanted "noise" and only the message–specific information contributes to the recogntion. This experiment is repeated for all possible opposite-gender speaker-test pairs and results are averaged. Even though the recognition rates are not very high, the experiment still indicates the optimal amount of the model smoothing, which was in this case the smoothing by the 5th order all-pole model, which forms at most two spectral peaks.

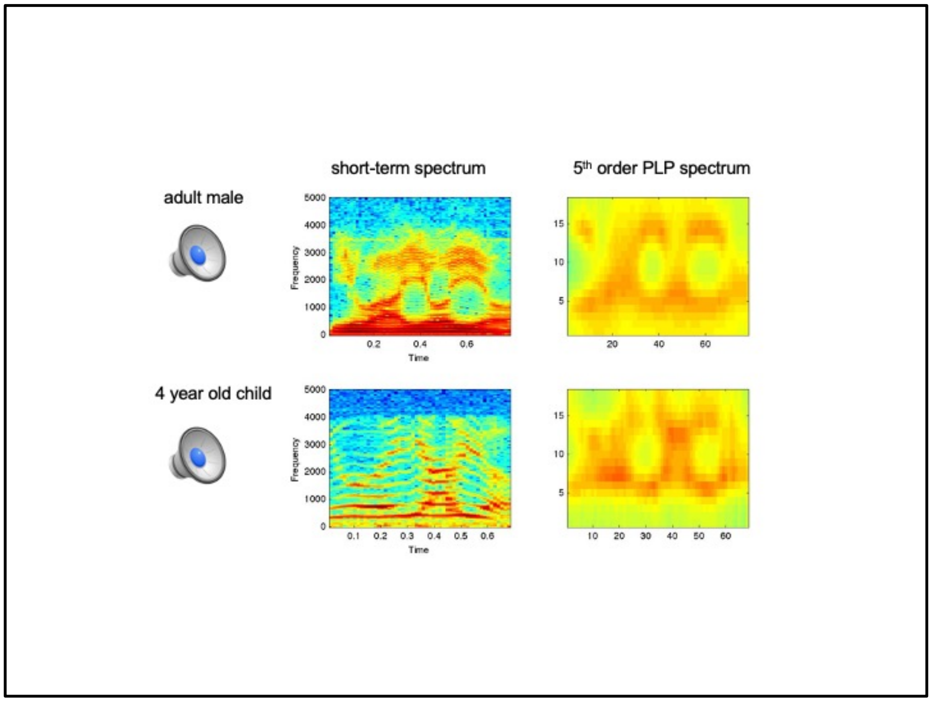
H6. Effect of the spectral model order in automatic speech recognition. Kazuhiro Tsuga and Hynek Hermansky (Speech Technology Laboratory, 3888 State Street, Santa Barbara, CA 93105)

It has been observed that in speaker-independent multi-template digit recognition, the 5th-order perceptually based LP (PLP) analysis method yields about 40% lower error rates than does the standard 14th-order LP.

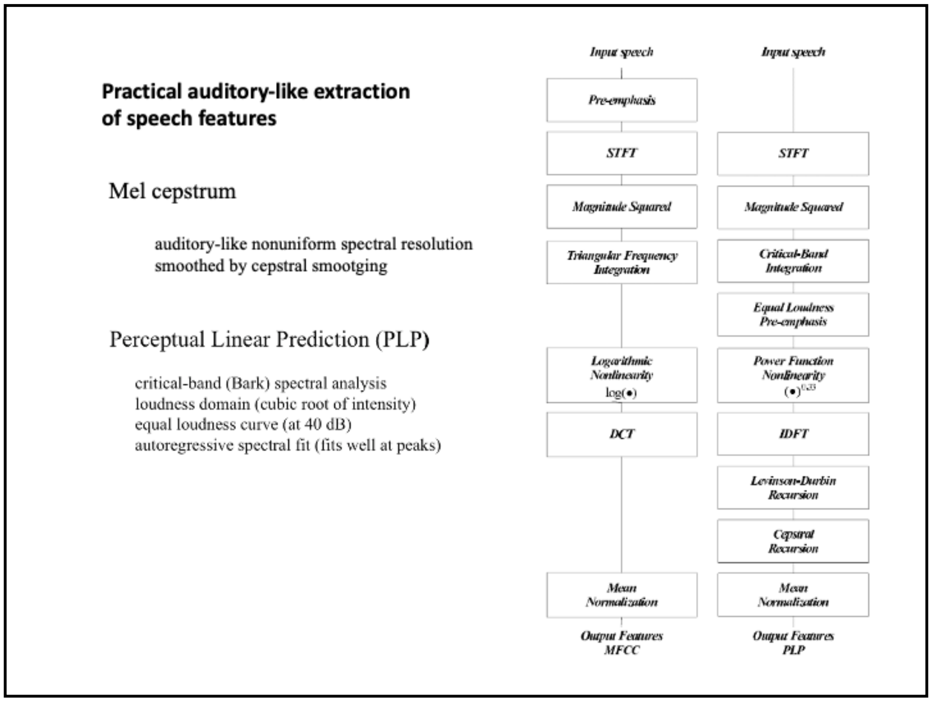
J. Acoust. Soc. Am. Volume 80, Issue S1, pp. S18-S18 (1986)

RE: TWO INVENTIONS BY DR. H. HERMANSKY ET AL
WE HAVE CAREFULLY STUDIED WHETHER TO FILE U.S. PATENT APPLICATIONS FOR THESE INVENTIONS. AS MR. NYUJI TELEPHONED YOU, OUR CONCLUSION IS:
(1) ''PERCEPTUALLY BASED LINEAR PREDICTIVE ANALYSIS OF SPEECH''.
WE CANNOT SEE ANY PRESENT OR FUTURE PRODUCTS TO WHICH THIS INVENTION IS PRESUMED TO BE APPLIED. SO, THIS INVENTION DOES NOT HAVE ENOUGH PRACTICAL VALUE TO BE APPLIED FOR A U.S. PATENT.

The low-order PLP model was effective in larger speaker-independet digit recogntion. The techniques was developed at Speech Technology Laboratort of Panasonic Company. Interesting thing was that at that time Panasopnic declined its patenting, eventhough eventually PLP became one of the techniques of choice of some major companies in the early 21st century.



The low-order PLP enhances what is similar in spectra of adult and children speakers.



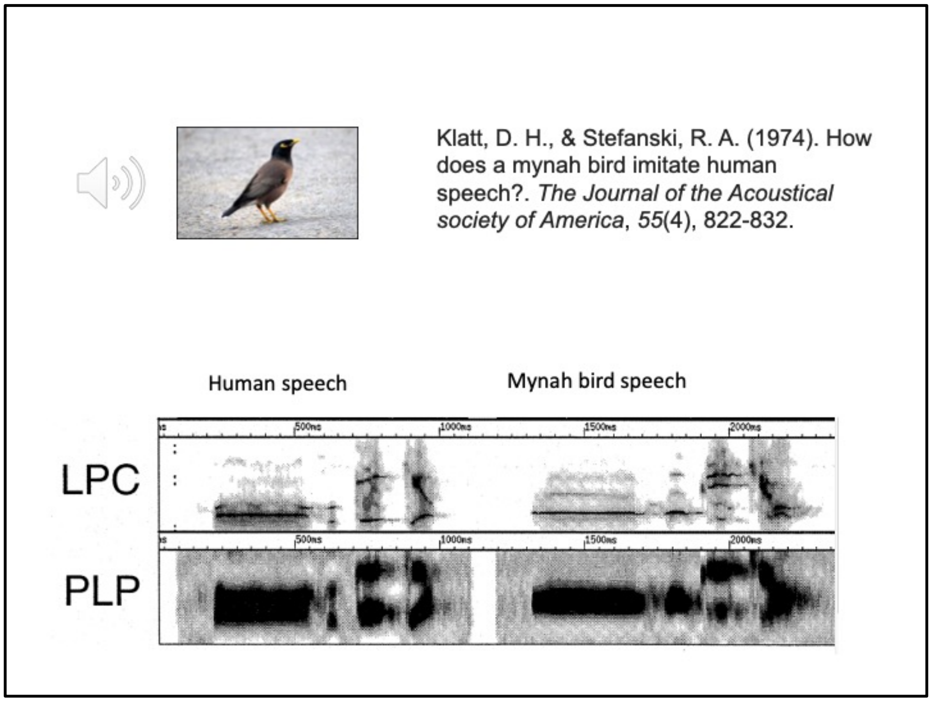
Here we see the similartires and differences between computation of the mel-cepstrum and PLP-cepstrum. One difference is in the form of spectral weighting to form the auditory-like spectrum. Another difference is in the form of the spectral smootnig, the mel-cepstrum using the cepstral smoothing, PLP main spectral smoothing is done by allpole modeling.

Limited spectral resolution

formant clusters as may be interpreted by auditory perception

WHY?

It is one thing to show that the technique works in applications. It is quite another thing to understand what is behind this working. So this quiestion “why?” is with us all the time. It is only when we understand why we see the particula resuls, it is difficult to make more progress. Our success might have been to accidental combination of reasons and it is difficult to generalize to new situations.



Mynah bird can imitate human speech very well in spite of having very different means for speech production. Spectral peaks extracted by LPC analysis are quite different but when analyzed by low order (5th) PLP, the perceptual similarities become apparent.