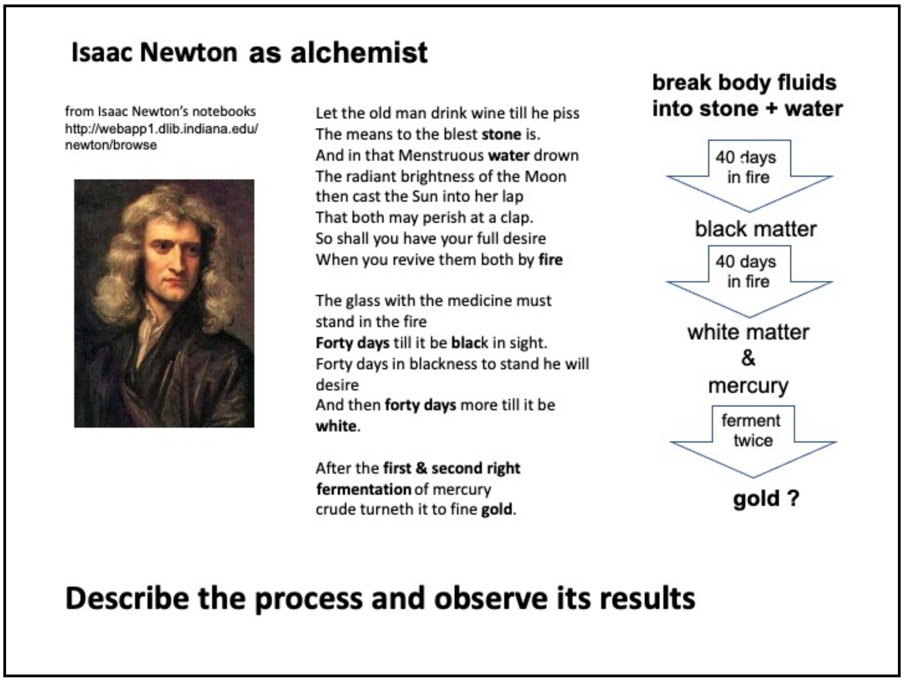
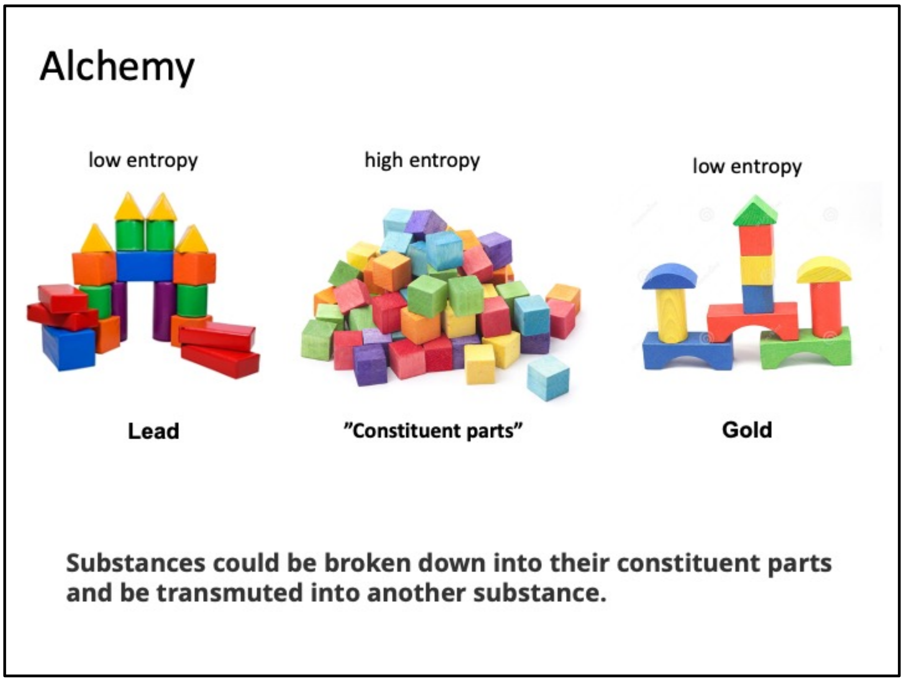
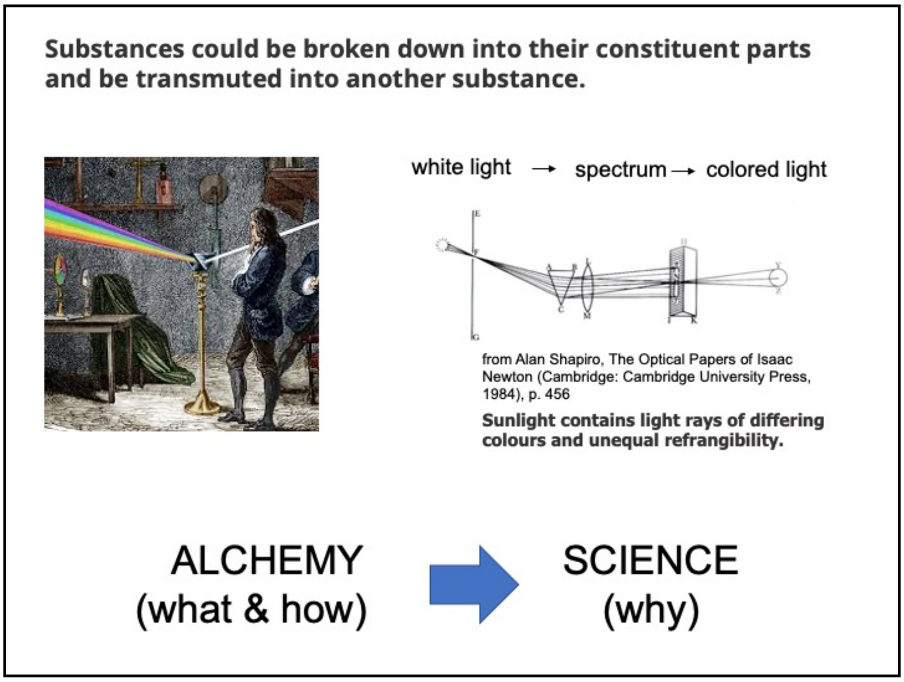


Alchemist such as Newton believed that substances can be broken to individual components and new substance restructured from the primary components.

One recepie for an experiment from Newton's notes suggest "special" stone and "special" water to be boiled for a particula time while observing emergency of "black matter" and :white matter". After "fermentatiuon" of the resulting white matter with mercury, gold is produced.
What we see her is the description what needs to be done and how. We do not know why



One of the key beliefs of alchemists was that materials can be broken to their “constituent parts” and these part can be reansembled to form new material.



Even when Newton probably did not succeded in breaking body fluids into fundamental componets and restructure then into gold, his principle of “breaking up” and ”restructuring”
allowed him to break the white light into different spectral componets and “restruring” some of the spectral components into a particularly colored new light. So here he also succeded to understand **the** this happened, the lihght consists of elements with various properties (he may not known about the wavelenghts) but understood enough
to form his theory of light which remains until today. **Newton’s alchemy turned into Newton’s science.**

Scientific method : set up your hypothesis and decide on running experiments which would reject it

- Why we did it?
  - Hypothesis: We speak in order to be heard ...
- What we did ?
  - Emulated some basic properties of human hearing in extracting features which carry messages in speech. We got better results in recognizing speech from multiple speakers.
- How we did it?
  - Mel cepstrum

    auditory-like nonuniform spectral resolution

    smoothed by cepstral smoothing
  - Perceptual linear prediction (PLP)

    critical-band *Bark" spectral resolution

    emulated equal loudness curce at 40 dB

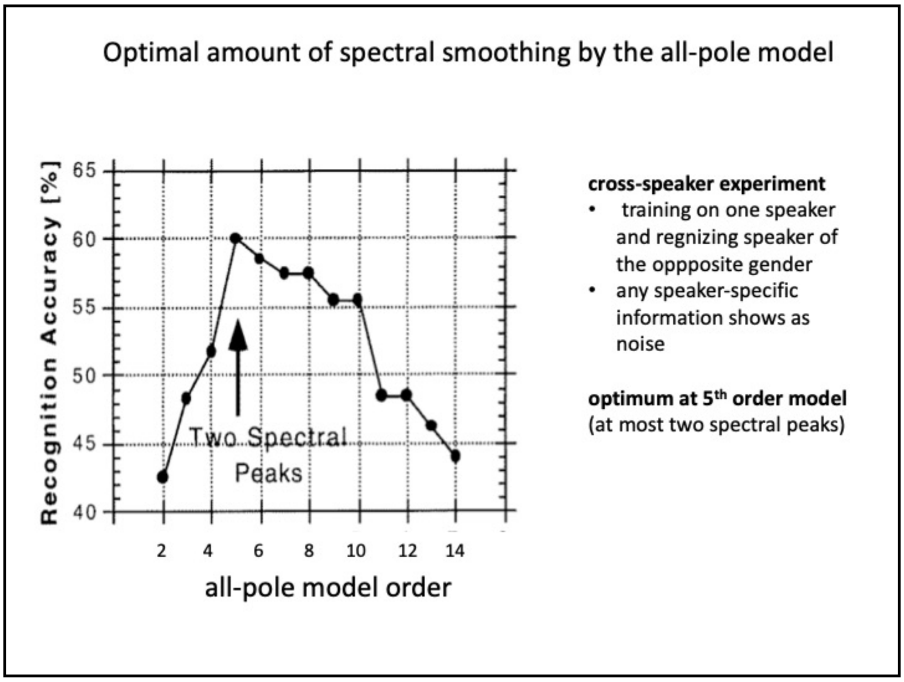
    loudness domain for further processing

    autoregressive spectral fit to approximate spectral pears in the auditory-like spectrum

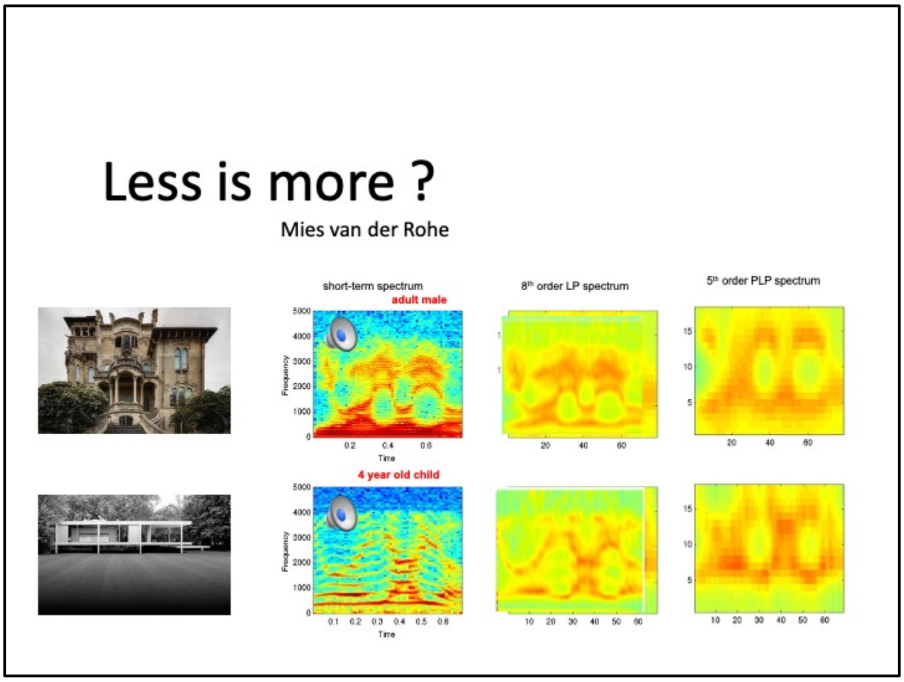
Question: Is the hypothesis supported?:

Why did we get results that we got ?

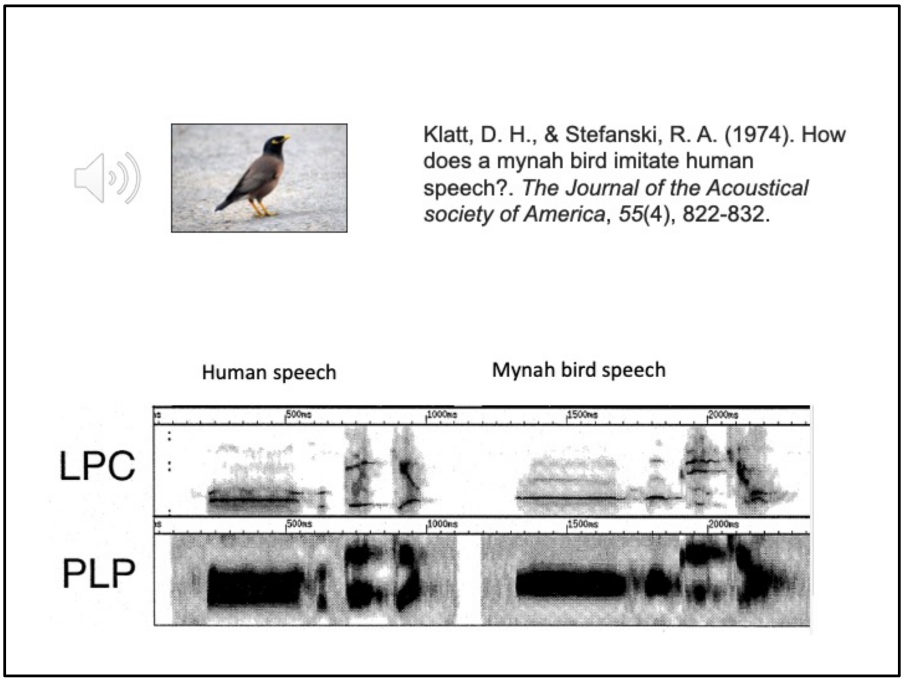
You remember the three questions which we need to set up our hypothesis – why we decided to do something, what we will do, and how we will do it?
At the end, we need to evaluate if our hypotheis was supported by our resuts, i."e. why the results turned in a certain way,. That leads to answering the question what is it that we learned



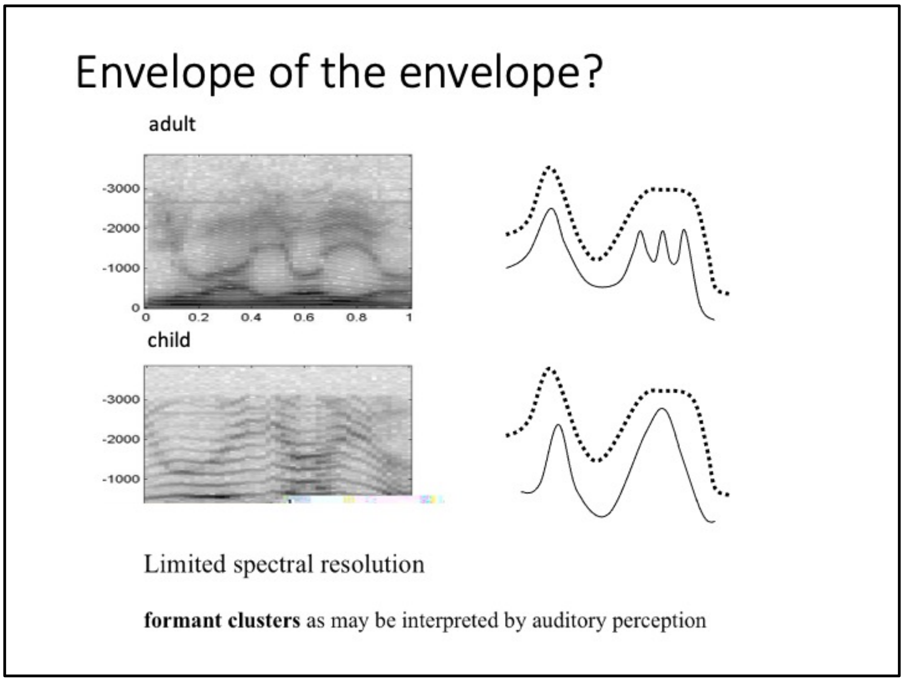
For alleviating the speaker-specific information, some additional spectral smooting of the auditory-like spectrum may be required. The amount of smooting can be derived by speech recogntion experimenst, where the training of the system (spectral templates) are provided by one speaker and the test speech 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.



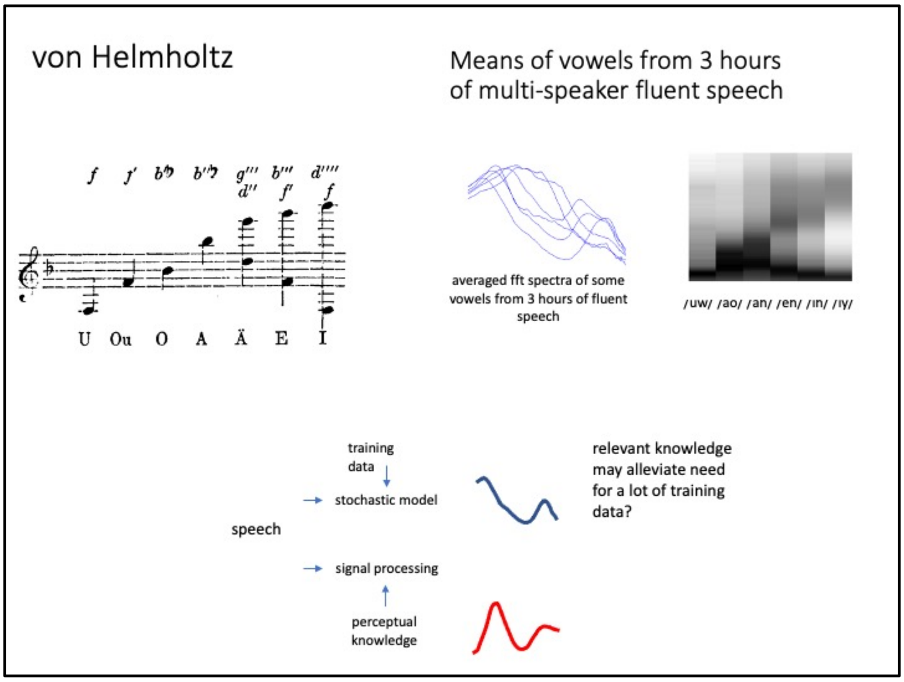
0th century arichitecture came with the concept of simplifying. Comparing to earlier arcitectural designs, the new ‘Functionalist” buildings look cleaner since many unecessary details which served no purpose but to impress the observers, were left out. The slogan was “Less in more”. This should be also said about low orged PLP. We need to remember that speech carries information from many sources and many of these sources maybe evident in the shoty-time speech spectrogram. The only information which is missing in the specrogram is the short-time phase of the signa, otherwise everything is left in. Smoothing out the fine spectral structure bu finding spectral envelopes alleviateg some speaker-specific information but the highly speaker-specific formant structure was still left in. %th order PLP does further spectral smoothing through the combined effects of the hearing-consistent critical-band spectral integration and the low-order spectral smoothing in the auditory-like domain.



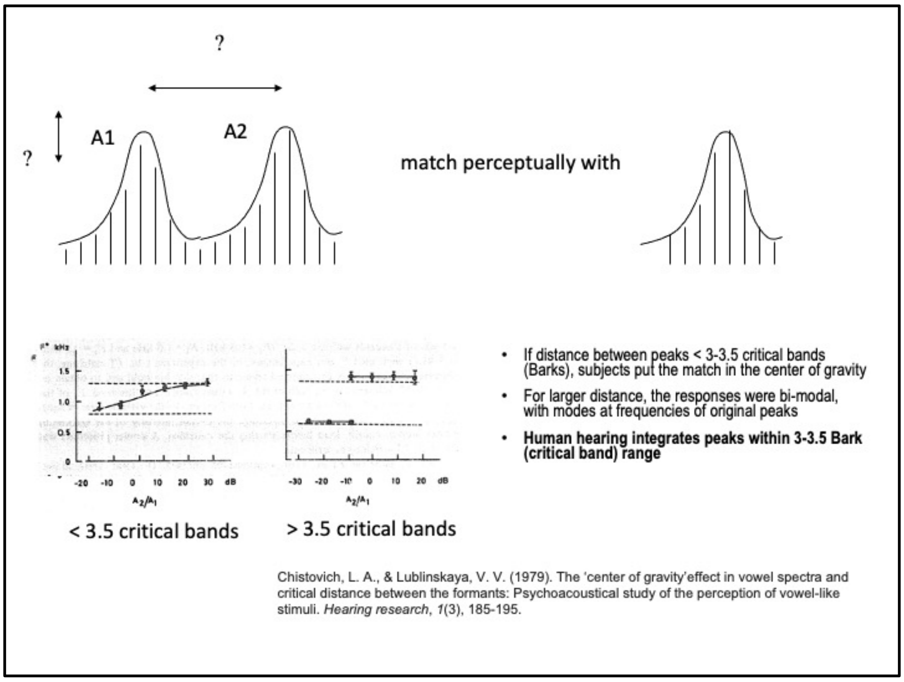
Mynah bird can imitate human speech very well in spite of having very different means for speech production. Spectral peaks extracted by LPC analysis are quite different but when analyzed by low order (5th) PLP, the perceptual similarities become apparent.



In effect, the low order PLP finds the “envelope of the envelope:\”, i.e. it integrates the higher spectral clusters into a singlespectral peak.



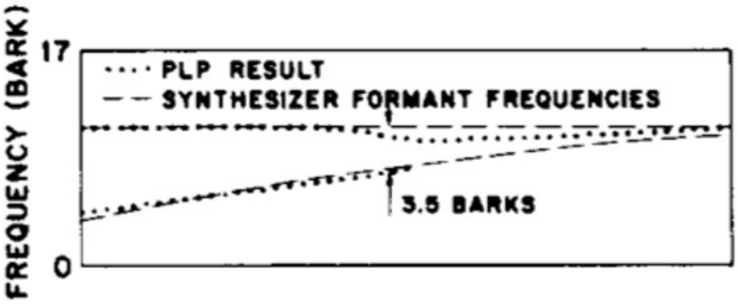
Indeed, the two-peak vowels were proposed by von Helmholts as speaker-independent representation of linguistic messages in vowels. When finding averaged spectra of vowels from continuous speech, the two peak spectra emerge. So providing directly the two-peak speech representation as models of speech in speech recognition could help in making the recognizer less speaker dependent in a similar wa as training it on large amounts of speech data.



Some fifty years ago, experiments in perception of spectral peaks indicated that human hering indeed integrates spectral peaks oner 3-4 crtitical bands. In this experimenst, the subjects were asked to match the two spectral peak stimuli by a single spectral peak stimuli. When the two peaks were further apart than 3-4 critical bands, the responses were bibodal, i.e. the match was into the one or the another peak. That means two peaks were pereceiven. When the peaks in the two-peak stimuli got closer that 3 crtical bands the frequenct=y if the matching single peak was in the center of gravity of the two peaks, i.e. tke two peaks were integrated into one.

Chistovich: 3.5 Bark spectral peak integration in human speech perception

Spectral peaks (formants) that are closer than 3-3.5 critical bands (Barks) are perceived as one peak (center of gravity of the two peaks)



When this experiment was emulated uring the low order PLP analysis, the same phenomenon was observed. The model formed two spectra paks when the signal spectral peaks were further than 4 critical bands and it formed a single peak when the spectral peaks in the stumulus were closer that 3 critical bands. This suports the notion that the 5th order PLP emults well this particular perceptual phenomenon of human hearing.

$$F'_2 = \frac{F_2 + c(F_3F_4)^{1/2}}{1 + c}$$

$$c = \left(\frac{F_1}{500}\right)^2 \left(\frac{F_2-F_1}{F_4-F_3}\right)^4 \left(\frac{F_3-F_2}{F_3-F_1}\right)^2$$

Fant and Risberg 1962
Effective perceptual second formant F2’

- F2’ (effective second formant) does not coincide with any higher formants

The two-peak representations of vowels was studied quite extensively by the prominent Slockholm speech group. The full vowel stumuli were perceptually matched by two spectral peak stimuli. The averaged responses indoicated that the first peak of the matching stimuli was at the frequency of the first formant but the second (so called “effective perceptual second formant” peak was at the frequency which was gicen by weighted average of all formants.

Fant and Risberg 1962
Effective perceptual second formant F2'

Match all Swedish vowels by 2-formant stimuli
The first perceptual formant F1' at F1
The second perceptual formant F2' is a function of all higher formants

$$F'_2 = \frac{F_2 + c(F_3F_4)^{1/2}}{1 + c}$$

$$c = \left(\frac{F_1}{500}\right)^2 \left(\frac{F_2-F_1}{F_4-F_3}\right)^4 \left(\frac{F_3-F_2}{F_3-F_1}\right)^2$$

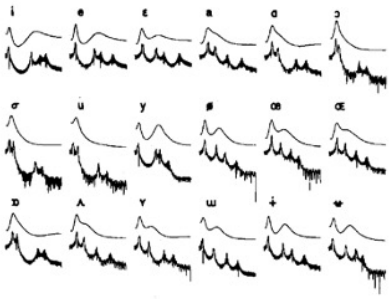
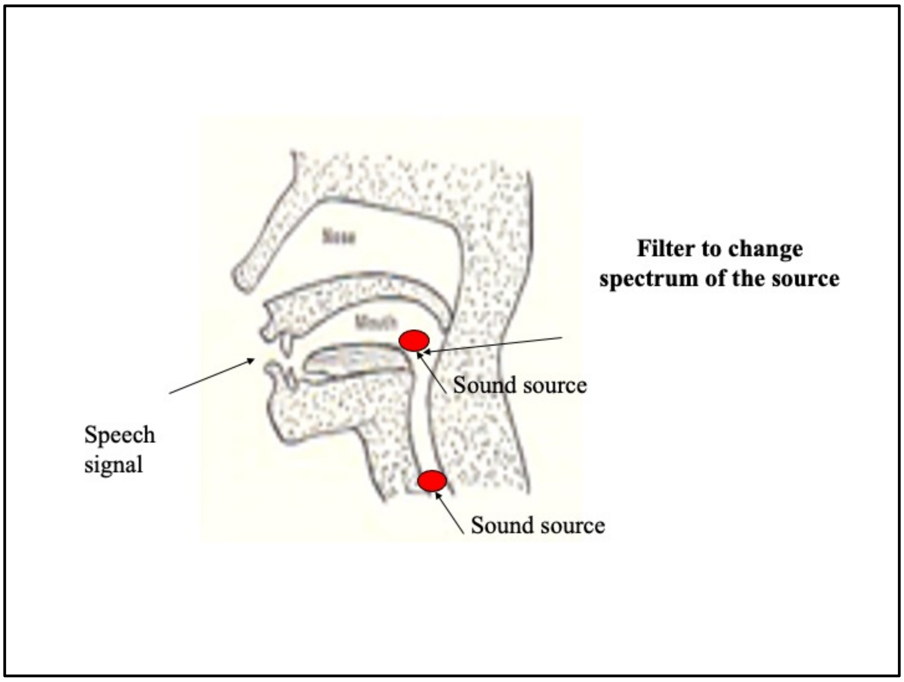


TABLE I. Perceptually estimated (Bladon and Fant, 1978) and PLP estimated frequencies of perceptual formants of 18 cardinal vowels.

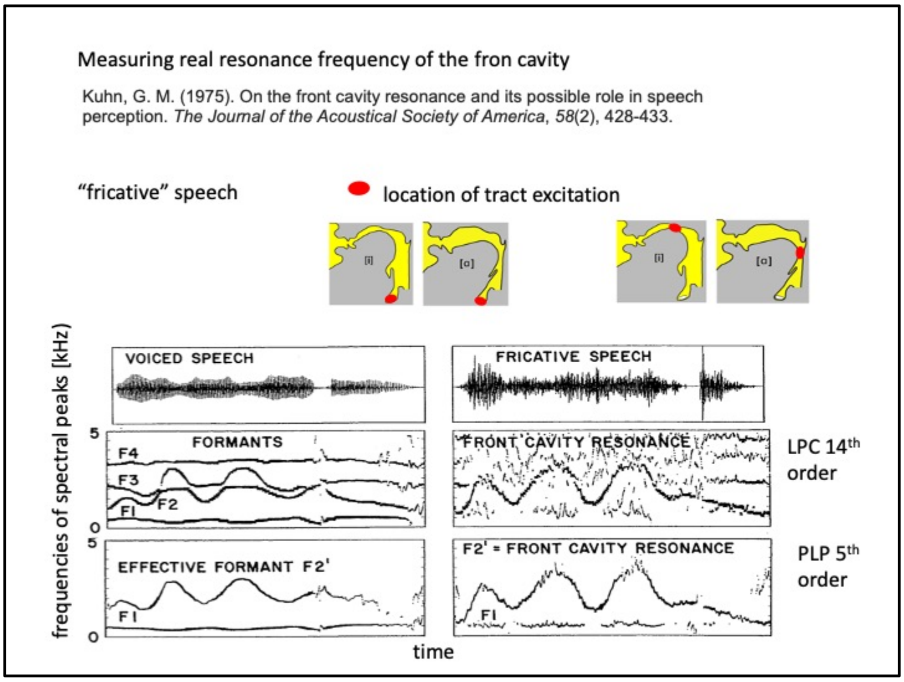
	Perceptual			PLP		Error	
Vowel	$F1$ (Bark)	$f2'$ (Bark)	$f2'-f1$ (Bark)	$F1'$ (Bark)	$F2'$ (Bark)	$F1'-F1$ (Bark)	$F2'-f2'$ (Bark)
i	2.9	14.1	11.2	3.4	13.3	0.5	−0.8
e	4.3	12.5	8.2	4.3	12.8	0.0	0.3
ɛ	5.8	11.7	5.9	5.3	11.7	−0.5	0.0
a	6.4	9.7	3.3	6.2	merged w/$F1$	−0.2	n/a
ɑ	5.7	8.2	2.5	5.6	merged w/$F1$	−0.1	n/a
ɔ	5.1	6.6	1.5	5.3	merged w/$F1$	0.2	n/a
o	3.5	6.0	2.5	4.8	merged w/$F1$	1.3	n/a
u	2.8	5.8	3.0	4.7	merged w/$F1$	1.9	n/a
y	2.9	11.8	8.9	3.4	11.8	0.5	0.0
ø	4.2	10.1	5.9	4.3	10.7	0.1	0.6
œ	5.6	10.4	4.8	5.4	10.7	−0.2	0.3
oE	6.0	9.7	3.7	5.7	9.7	−0.3	0.0
ɑ	5.8	7.4	1.6	5.9	merged w/$F1$	0.1	n/a
ʌ	5.4	9.0	3.6	5.3	merged w/$F1$	−0.1	n/a
ɤ	4.2	9.2	5.0	4.3	9.7	0.1	0.5
ɯ	2.9	9.1	6.2	3.4	10.0	0.5	0.9
ɨ	3.6	10.8	7.2	3.8	11.0	0.2	0.2
ʉ	3.4	9.9	6.5	3.8	11.1	0.4	1.2

The two-peak representations of vowels was studied quite extensively by the prominent Slockholm speech group. The full vowel stumuli were perceptually matched by two spectral peak stimuli. The averaged responses indoicated that the first peak of the matching stimuli was at the frequency of the first formant but the second (so called "effective perceptual second formant" peak was at the frequency which was gicen by weighted average of all formants.

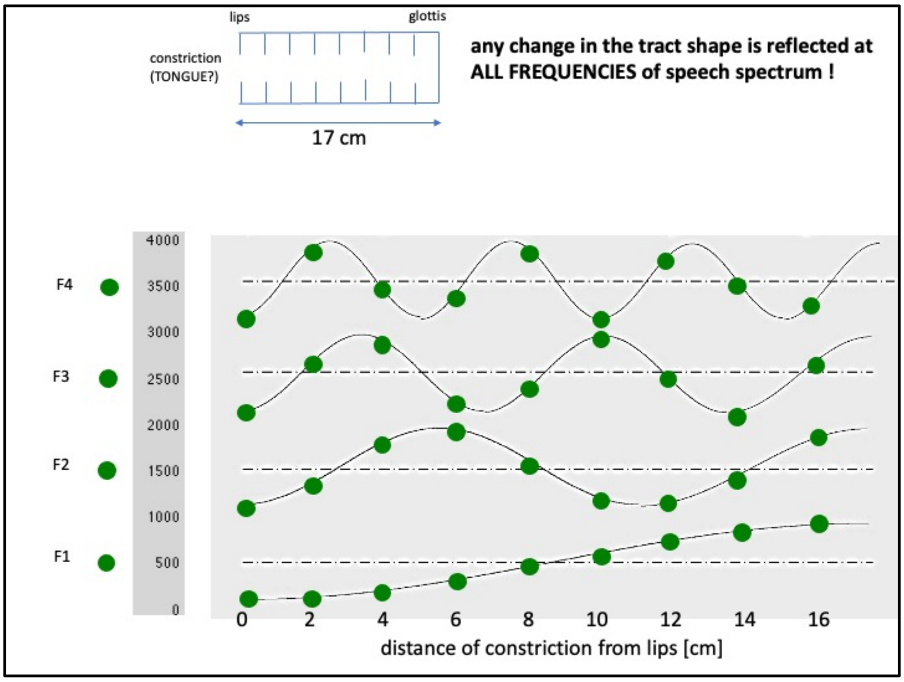
When the synthetic vowels were analyzed by 5th order PLP, the secod peak of the PLP model of front vowels coincided well with the effective second formant. The back vowels, where the first and the second effective formants were closer that 3 critical bands, were modeler by PLP model in which the two peaks merged into one peak. Again, the low order PLP emulated well the phenomena reported in speech perceptual experiments.



Two basic elements of the vocal tract are the **tract cavities** which filter the spectru of the **sound source**.

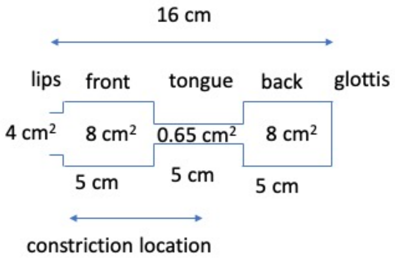


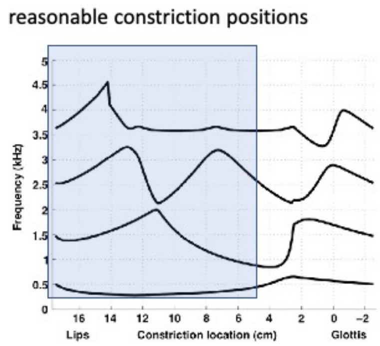
Does the effective perceptual second formant have any correlate in speech production? The answer may be "yes". It was proposed in [Fant, *On the accoustics of speech*,1960] that thei effective perceptual second formant F2' may coincide with the resonace frequency of the uncoupled front cavity in production of speech. Kuhn (1975) proposed and carried out a simple but ingenious experiment, where he produced speech by exciting the front cavity of vocal tract by making the tract constriction narrew enough so that the friction formed at the constriction point, which excited the front cavity. The front cavity resonace frequency coincided sometimes with the seconf formant f2 and sometimes with the third formant f3 in the normal voised speech. When both the normal voiced and the fricative speech was analyzed by 5th order PLP analysis, the reults were rather similar. always putting the second peak of the PLP model at the resonance frequency of the front cavity. This shows that low-order PLP finds the resonance frequency of the uncopled fron cavity in production of speech and support the hypothesis of F2'-fron cavity correspondence proposed by Fant.



A we move the constriction of the tube along its length, all formant frequencies are changing folowing the sensitivity functions derived from the perturbation principle. So any change of the vocal tract shape is reflected at most frequencies of the speech spetrum.

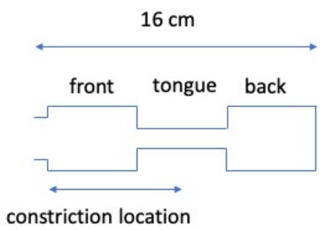
Four-section vocal tract model

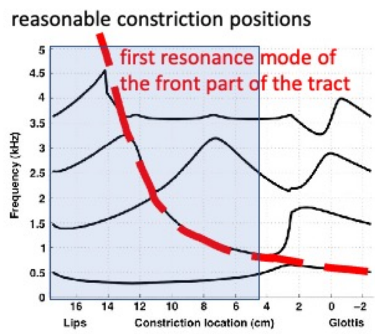




Schwartz, J. L., Boë, L. J., Badin, P., & Sawallis, T. R. (2012). Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial–coronal–velar stop series. *Journal of Phonetics*, *40*(1), 20-36.

Four-section vocal tract model



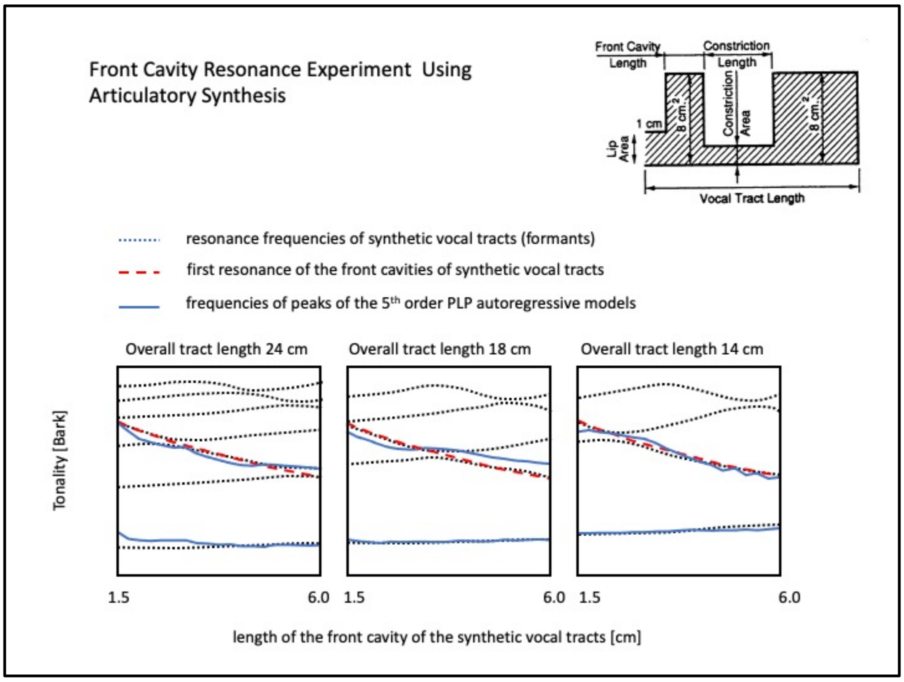


Schwartz, J. L., Boë, L. J., Badin, P., & Sawallis, T. R. (2012). Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial–coronal–velar stop series. *Journal of Phonetics*, *40*(1), 20-36.

PLP-estimated F2’ and Front Cavity Resonance Frequency

- Articulatory Synthesis
  - formants known
  - resonance frequency of decoupled front cavity can be computed
  - synthetic speech is available for analysis by PLP (F2’ can be estimated)

It is possible to synthesize speech with known speech production parameters. Whe we know both the vocal tract resonances (formants) and where the fron part of the vocal tract would reconante when decoupled from the whole tract configuration, plus having the synthetic speech signal for different vocal tract configurations, allows for testing the hypothesis.



Reasonably realistic vocal tract shapes for several overall vocal tract lenghts in productions of front vowels (where the PLP model forms two spectral peaks) were used to generate sythetis speech with known formant frequencies and with the knowkledge where the forn part of such vocal tract configurations would resonate. By analyszing the synthetic vowels the frequencies of the two peaks of the PLP model were extracted. These frequencies were relatively invariant with the changes of the overall tract lengtsh and mainly reflected shapoes of the front parts of the vocal tract emulations.

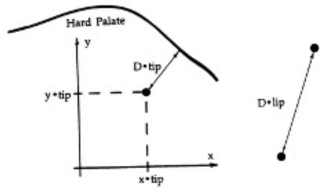
Result of Experiment with Synthetic Vowels

- **correlations on about 11 000 synthetic front vowels**
  - (back vowels for which PLP formed only one peak were excluded)
  - tract length varied between 14 and 24 cm

	tract length	front cavity resonance
Second peak of PLP model	**-0.18**	**0.9**
formants (averaged)	-0.71	0.22

Statistical analysis of the results indicated high correlations of the frequency of the second peak of the PLP models with the resonance frequency of the front cavity of the tract shapes and low correlations with the overall tract lentght. The oposite correlation trends were observed for the formants extracted by the LPC analsysi.

X-ray Microbeam Experiment
(Broad and Hermansky 1989)



$$L = k1 - \alpha \mathbf{x}$$

(a) $$\mathbf{x} = x_{tip} \cos\theta + y_{tip} \sin\theta$$

$$\Phi = k2 + b1 \ln D_{tip} + b2 \ln D_{lip}$$

(b) $$\boxed{\frac{1}{F2'} = \frac{4L}{c} \, \frac{2}{2 + \Phi}}$$

(c) PARAMETERS:

$k1,\ k2,\ \alpha,\ \theta,\ b1,\ b2$

- Shape approximated by cosine with period of 2**L** and amplitude Φ
- Resonance frequency given by **L** and $\boldsymbol{\Phi}$ (Schroeder, Mermelstein)

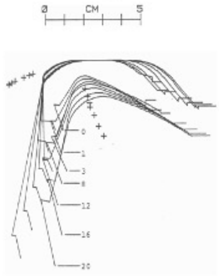
- two male speakers
  - "where were you a year" three times each
- front cavity resonance from articulations
- PLP-estimated F2' from acoustic data

CORRELATION BETWEEN RESONANCE FREQUENCY OF FRONT CAVITY AND PLP-DERIVED F2'

Speaker	Speaker 1	Speaker 2
Correlation	0.95	0.92

X-ray microbeam provides information about several samples of the front part of vocal tract in speech production. The generated speech is also available. From the several samples of the tract shape , a simple approximation of the front part as of tract shape as a half cosine function can be derived. From the estimated half-cosine length and the cosine ampliture the resonance frequency of the front cavity can be computed. At the same time, the second peak frequency of the low-order PLP model is derived from the speech signal and correlated with the front cafity frequency. The correlations for two studied speakers were rather high, further lending support for the fron cavity –F2' hypothesis.

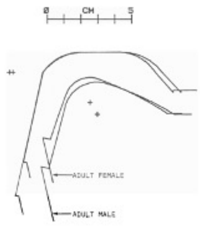
Vocal tract growth with age



	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

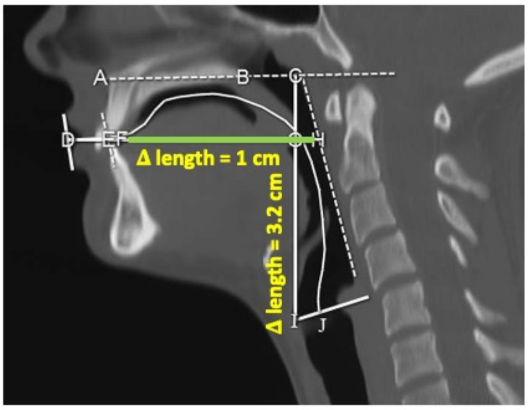


	pharynx	oral
adult male	89 mm	81 mm
adult female	63 mm	78 mm
	41 %	4 %

Ursula Goldstein, MIT PhD. Thesis 1980

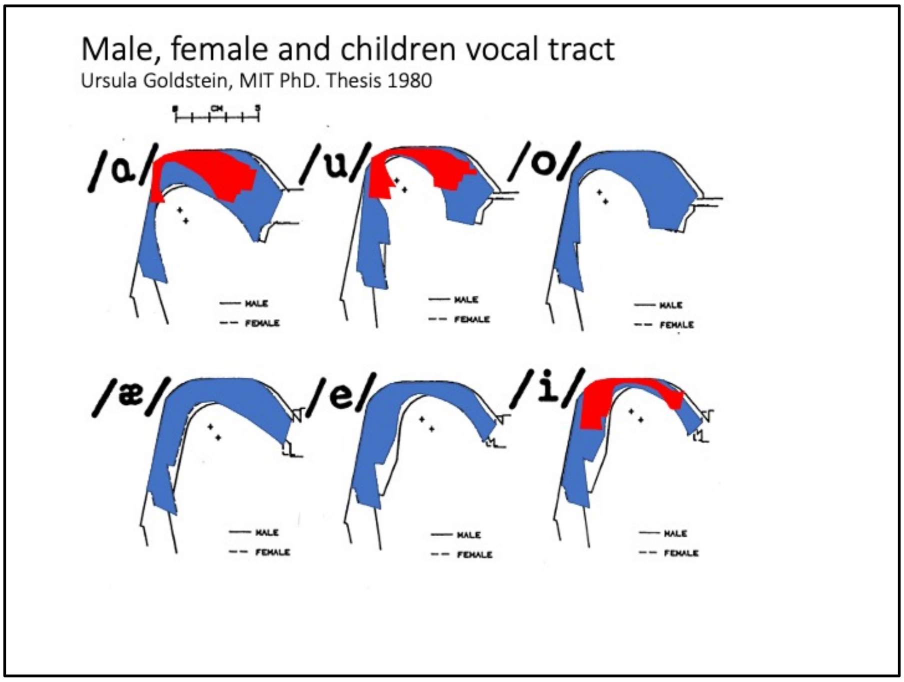
Back part of the vocal tract grows three times more with age that does the fron part. So it has much more sense that as children learn to speak, they learn how to correctly form the front cavity of the vocal tract. In general, the back cavity is much more difficult to control anyways. Only actors may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



Between 4 of age the back part of the male vocal tract grows 3 times more that its front part.

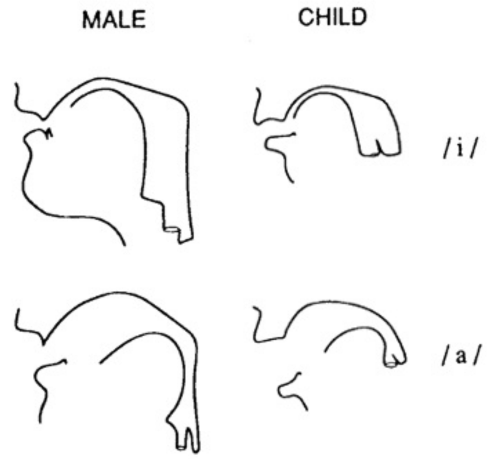
Vorperian, Houri K., et al. "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study." *The Journal of the Acoustical Society of America* 125.3 (2009): 1666-1678.



PhD works of Ursula Goldstein from MIT studied evolutions of vocal tract during the lifetime. Here are her estimates of the vocal tracts of males, females and children in production of vowels. The similarities of the front part of the vocal tract are seen, the back parts are strikingly different.

X-rays of Male and Child Vocal Tract in Production of Vowels

- In production of vowels, the front part of the vocal tract appears to be less speaker dependent than its back part
  - Hermansky and Broad 1990

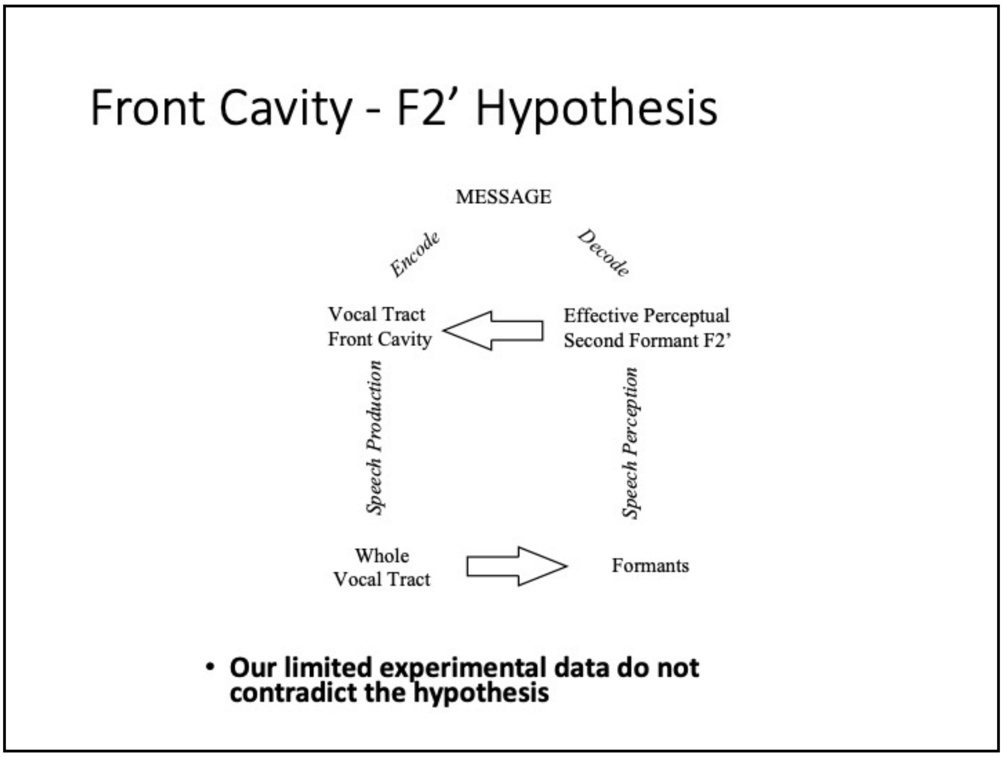


These shapes were traced down from x-rays of real productions of two vowels by an adult and by a child. DIfferences in the pharyngeal (back) part of the tract are again striking.

Front Cavity - F2’ Hypothesis

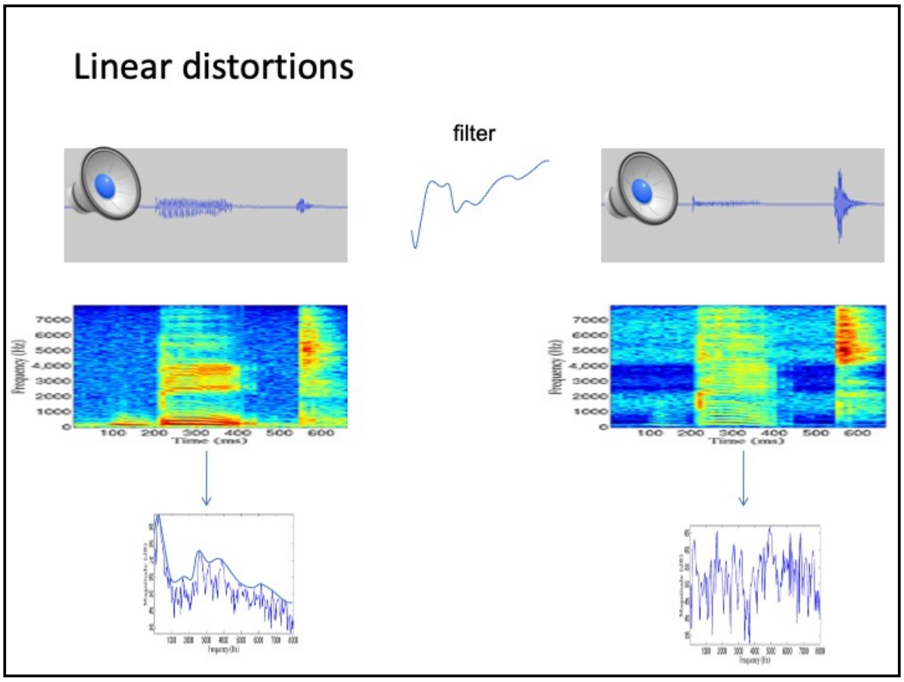
- F2’ correlates with resonance frequency of decoupled front cavity of vocal tract in production of vowels
  - Fant 1960
- Front part of the vocal tract
  - grows less during lifetime
  - is easy to manipulate without special training
  - for many consonants, the front part dominance is well accepted

The front cavity–F2’ hypothesis is temptimg. It would explain how the speech production skills evolve over the lifespan. Since the front part of the tract grows less than the back part, and it is easier to manipulate, it makes sense that this is what children learn g=how to manipulate. As a matter of fact, when one learns new language, the instructions of how to produce different sounds relate to the fron part. Also, in many consonants, it ios only the front part before the consonantal constriction, which contributes to a sound.



Here is the whole hypothesis. The ,essage is coded in the shape of the front part of the vocal tract. However, the speech is produced using the whole vocal tract and therfore the speech spectrum is formed by the tract resonances and is higly speaker-specific, also he information about the speaker besides the infromation about the message. During decoding the message, human hearing modifies the the speech spectrum, enhancing the messge information which is carried in the smoothed auditory spectrum.

ANOTHER PROBLEM WITH SPEECH SPECTRUM



Another significant problem with short-time spectra is their excessive sensitivity to linear distortions, which can be caused by, e.g., different acoustic environments. SUch distortions, even when heard, do not change the perecived message in speech. One extreme example is shown here. Spectral envelope in the vowel /ee/ in the work ”beat” was estimated and the filter with its frequency response being inverse to this spectral envelope ised for filtering the word. As expected, the spectral envelope of the vowel /ee/ in the filtered word in almost flat, not exhibiting resonances expected in the sound /ee/. In spite of that, the /ee/ is clearly heard in the filtered word.

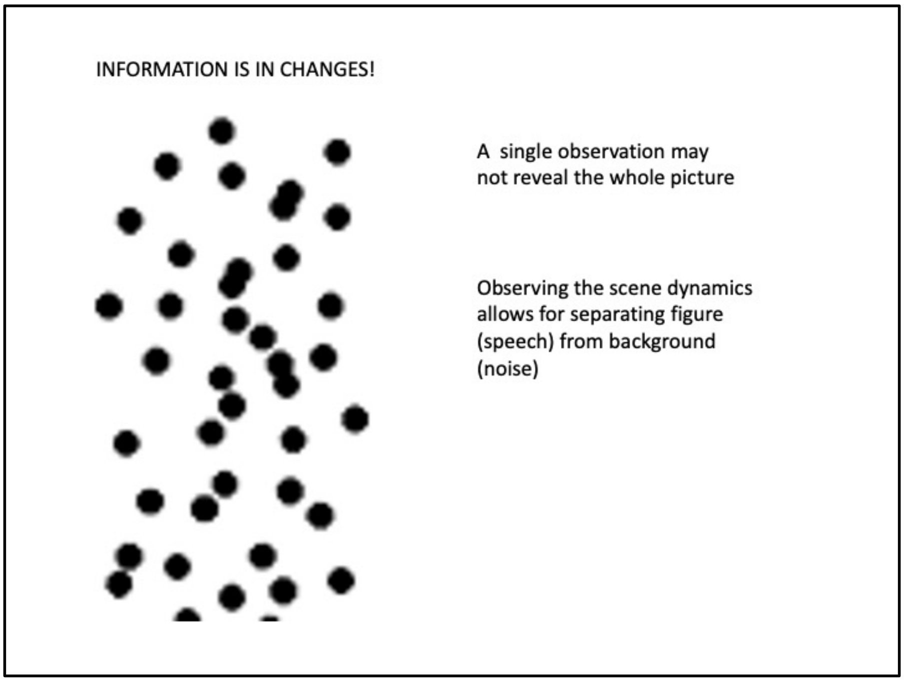
Speech short-time spectrum?

It is fortunate that speech intelligibility does resist the erosion of frequency selectivity, for our normal environment plays havoc with the speech spectrum. The world is full of objects, and the objects all cast shadows. Sound travels around corners, of course, but not all sound waves travel around corners equally well. Low frequencies get around far better than high frequencies. Consequently, the acoustic shadow of an object contains the low-frequency components of the sound while the high-frequency components are considerably attenuated. The speech spectrum behind a talker's head, for example, contains much less high-frequency energy than the spectrum in front of his head If speech were highly dependent upon faithful transmission of the spectra of the different speech sounds, it would necessarily reduce to a line-of-sight method of communication and many of the great advantages of vocal signaling would disappear.

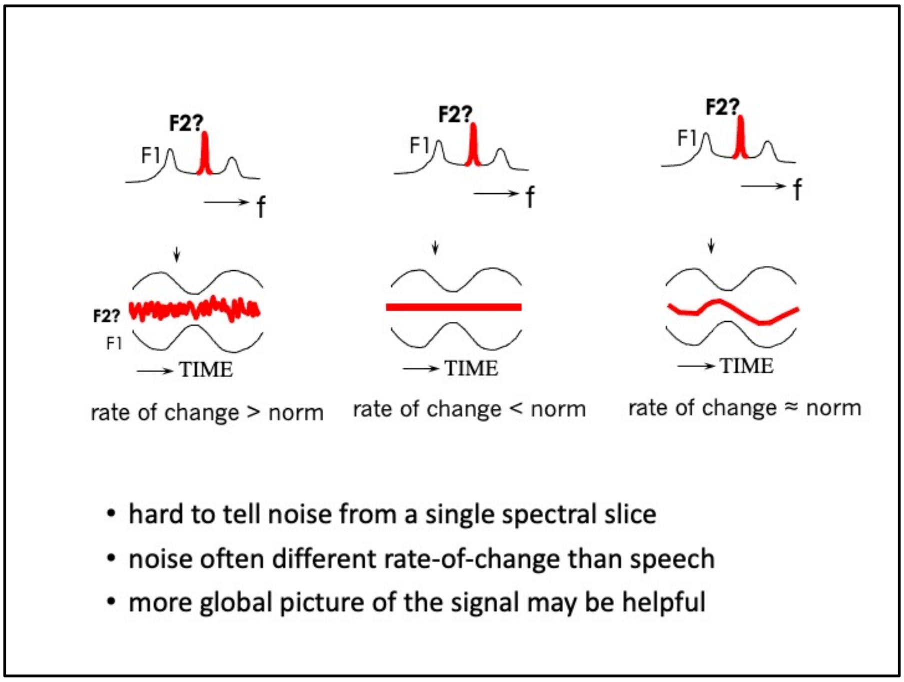
G.A. Miller: *Language and Communication*, p.96

George Miller discusses the problems with short-time spectrum of speech here.

Another nice example shows that steady patterns on the retina diminish. When focusing on the white dot oin the center of the figure, the colors in the background gradually dssapear, only to appear again when the attention moves away from the dot. We are not aware of the importance of canges in vision because our eyes are paremanetrly moving in a short saccades every 200-300 ms so that the pattern on the retina normally never stays steady.



In general, perception is paying attention to changes and not to steady patterns. A noice example is here. When looking at the pattern of dots, one has no idea what the dots represent. When some dots which represent samples from the figure start moving, the figure is clearly perecived on the background of the steady dots,



In speech, the spectra also permanetly change. SOme of the changes are faster and some are slower. We have seen earlier that major spectra; changes due to modulation if the sourse signal by changing sfape of the vocal tract are in the range 1-15 Hz, with the p[ean typically somewhere around 4 Hz. That may allow to separate these speech induced changes form the steady or very fast changing peak here (maked here as F2?).
The changes obviously cannot be seen in a single spectral slice but require more global view of speech over longer time spans.

Additive and convolutive noise

signal x(t)

speech **s(t)**

environment e**(t)**

additive noise **n(t)**

$x(t) = s(t)*e(t) + n(t)$

for uncorrelated additive noise and power spectral domain

$X(\omega,t) = S(\omega,t)E(\omega,t) + N(\omega,t)$

If **noise is known**, it can be first subtracted

$S_{CLEAN}(\omega,t)E(\omega,t) = X(\omega,t) - N(\omega,t)$

If **environment is known**, it can be later subtracted in log domain

$\log S_{REALLY\ CLEAN}(\omega,t) = \log S_{CLEAN}(\omega,t) - \log E(\omega,t)$

With additive noise, deal with the noise first !
When taking the log of the noisy signal, the noise is not additive anymore.

Noise which is additive in the signal and uncorrelated with the signal, remains additive in the spectral domain. In principle if the noise is known, it can be subtracted in the spectral dimain

The problem is that the noise spectrum is not known and when trying to estimate it, the estimates are not accurate. Still, it is advisible to deal with the noise before the logarithm of the signal is taken. Once in the logarithmic spectral domain, new components of the noise and the signal spectrum appear due to the logarithmic nonlnearity and the noise is not additive anymore but becomes signal dependent.

Linear distortions show as convolutions of the signal with the impulse response of the environment.In frequency domain it means that the spectrum of the signal multiplies with the spectral characteristics of the environment, i.e., as additive constants in the logarithmic domain. The aditive constants are different at different frequencies. Here the logarithic domain can be useful, if the spectrum of the environment is known, it can be subtracted in the logarithmic domain. Typically, the spectrum of the environment is not know. However, if it is not changing, it can be estimated by averaging the signal (spectral subtraction). However, the spectrum of the environments may change. Howevert, when the spectrum of the signal is changing at different rate than the spectrum of the environment, it can be filtered out in the modulation spectral domain.

Additive and convolutive noise

if the additive log E(ω,t) or N(ω,t) are unknown but changing slower (or faster) than speech, they can be filtered out

Filtering signal elements which are out-of-range of typical changes of speech make the output signal invariant to these harmfiul elements.

Noise which is additive in the signal and uncorrelated with the signal, remains additive in the spectral domain. In principle if the noise is known, it can be subtracted in the spectral dimain

The problem is that the noise spectrum is not known and when trying to estimate it, the estimates are not accurate. Still, it is advisible to deal with the noise before the logarithm of the signal is taken. Once in the logarithmic spectral domain, new components of the noise and the signal spectrum appear due to the logarithmic nonlnearity and the noise is not additive anymore but becomes signal dependent.

Linear distortions show as convolutions of the signal with the impulse response of the environment.In frequency domain it means that the spectrum of the signal multiplies with the spectral characteristics of the environment, i.e., as additive constants in the logarithmic domain. The aditive constants are different at different frequencies. Here the logarithic domain can be useful, if the spectrum of the environment is known, it can be subtracted in the logarithmic domain. Typically, the spectrum of the environment is not know. However, if it is not changing, it can be estimated by averaging the signal (spectral subtraction). However, the spectrum of the environments may change. Howevert, when the spectrum of the signal is changing at different rate than the spectrum of the environment, it can be filtered out in the modulation spectral domain.

$X_{CLEAN}(\omega,t)=S(\omega,t)E(\omega,t)+N(\omega,t)$

$\log S_{REALLY\ CLEN}(\omega,t)=\log X_{CLEAN}(\omega,t)-\log E(\omega,t)$

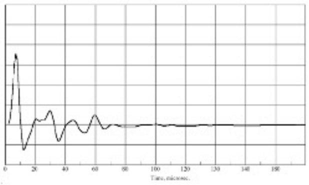
assumption

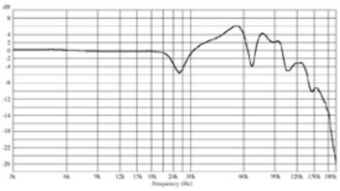
$S(\omega,t)$, $X(\omega,t)$ and $E(\omega,t)$ are known

i.e., their spectral resolutions are appropriate

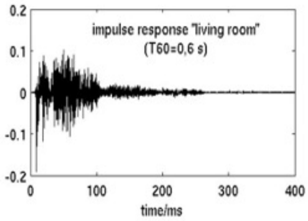
spectral analysis window for computing $S(\omega,t)$ need to be long enough to cover most of the impulse response of the environment easier for the impulse response of a microphone

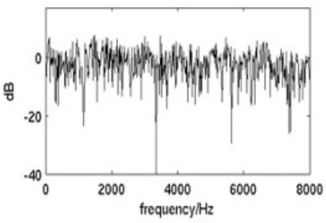
Kherkin (Earthworks)





more difficults for impulse response of rooms (revereberations)

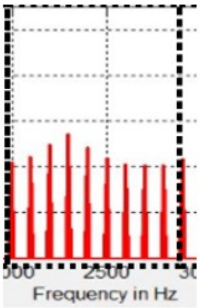




Hirsch nd Finster 2015

Modifying spectral resolution

sum FFT power spectral values within the auditory-like band



Σ

additive noise spectrim remains additive and the

$S'(\omega,t)E'(\omega,t) = X'(\omega,t) - N'(\omega,t)$

holds for the modified power spectra $S(\omega,t')$, $E'(\omega,t',)$, $X'(\omega,t)$ and $N'(\omega,t)$

multiplicative relation with the spectrum of the environment does not hold anymore

$\log S'_{\text{REALLY CLEAN}}(\omega,t) \neq \log S'_{\text{CLEAN}}(\omega,t) - \log E'(\omega,t)$

C. Avendano and H. Hermansky, "On the effects of short-term spectrum smoothing in channel normalization," in *IEEE ASSP* July 1997,