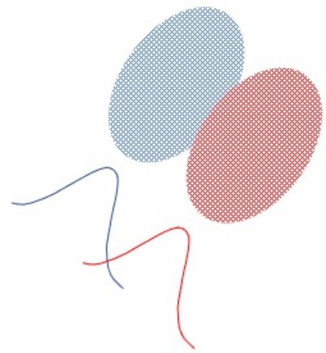
Linear discriminant analysis
(Ronald A. Fisher 1936)

- find projections of data, which preserves most of the discriminability
- data vectors need to be labeled by classes
- yields matrix of discriminant vectors, ordered by their discrimination power
- discriminants are linear and therefore can be easily interpreted



The Acoustic Modeling Problem in Automatic Speech Recognition
Peter Brown, CMU CS Department, AFWAL-TR-87-1161

One well established technique to emphasize discriminability in a feature vector space is the linear discriminant analysis (LDA). LDA needs labeled vector space, where is it know which class each vector represents. LDA finds a projection of the space in directions of the class discrinability.. To peresrve all original discriminability, all dimensions of the new space are needed. However, most discriminability is in the first few dimesions of the new vector space. The amount of discriminability in each direction is indicated by the valys of te eigenvector of the discriminat matrix, used in the projection.

One of the first (if not The First) use of LDA is speech recognition was by Peter Brown at CMU in 1987.

Within-class scatter matrix

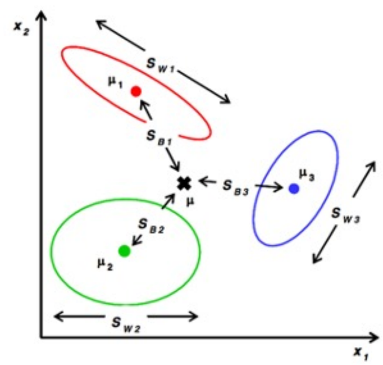
$$S_W = \sum_{i=1}^{C} S_i$$

- where $S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$
  and $\mu_i = \frac{1}{N_i} \sum_{x \in \omega_i} x$

Between-class scatter matrix

$$S_B = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^T$$

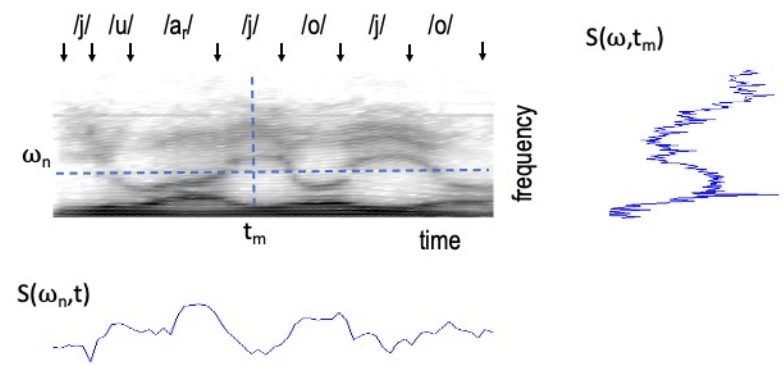
- where $\mu = \frac{1}{N} \sum_{\forall x} x = \frac{1}{N} \sum_{i=1}^{C} N_i \mu_i$



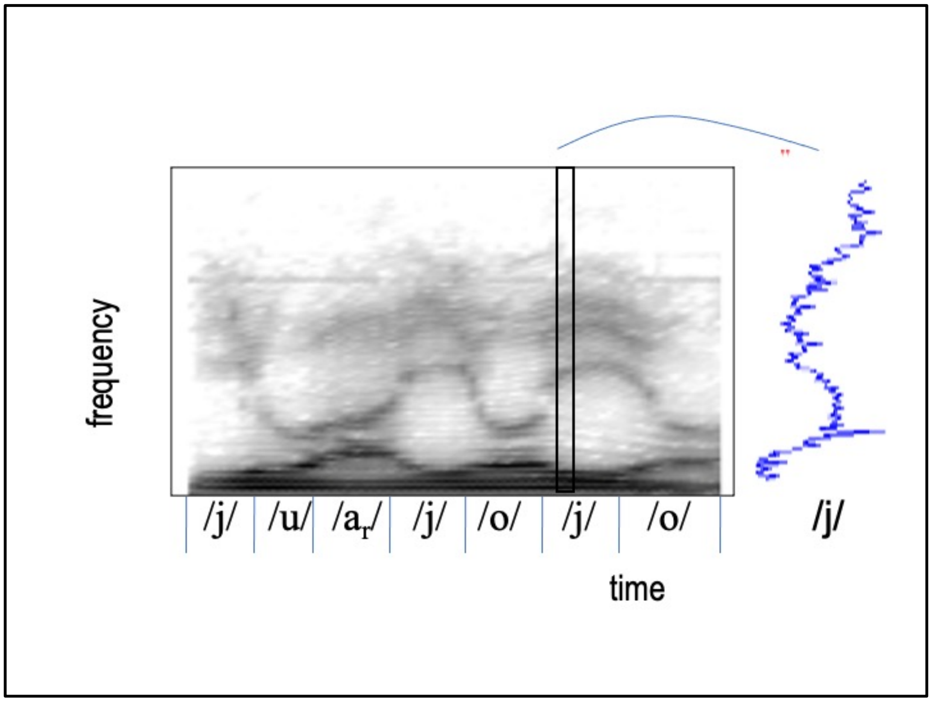
Find the eigenvectors with largest eigenvalues of

$$S_W^{-1} S_B$$

Here we have the math of LDA. Details can be found in any textbook on machine learning (e.g. **Pattern Classification and Scene Analysis, by Duda and Hart).**The bottom line is that we need to compute twio correlation matrices, one rescribing correlarions within classes and one correlationa across classes. MATLAB loves to do LDA but one deeds to carefu; not to ask for impossible, like using too small data sets. MATL:AB always uses some tricks to give you the answers but the answers may be meangless. Of course, this is the problem with most software packages and as a matter of fact with any ”knowledge” obtained from the WEB.

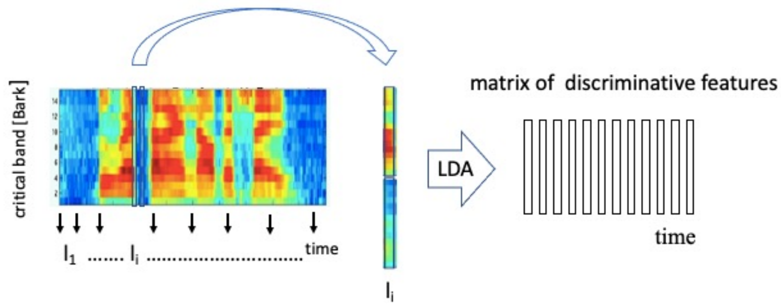


Spectrogram is 2D representation (matrix) of speech signal
Y-axix carries short-time spectra of speech at a given time in each column of the matrix.
X-axis carries evolutions of spectral energies at a given frequency over time in each row of the matric



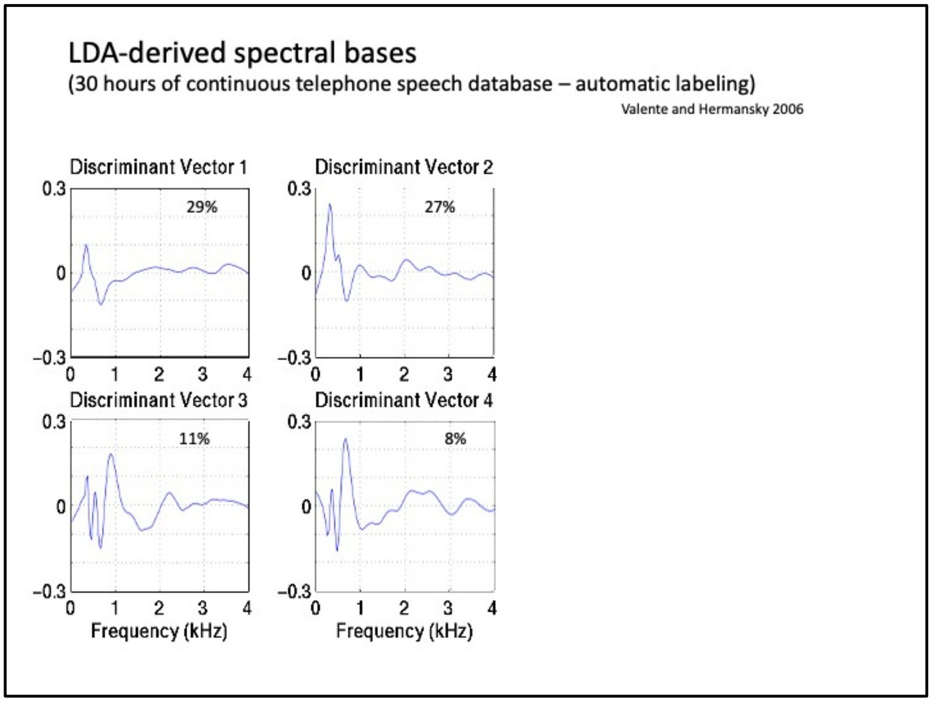
The spectral veclor in blue is describing sound /j/ so it will carry the label /j/. Once we have the labeled vector space, we can do LDA. The result of the LDA will be the matrix of linear discriminants which can be used for projecting the old short-time spectrum vector space in the directions of the most discriminability on a new vector space.

Using LDA to derive optimal projections of short-time spectral vectors

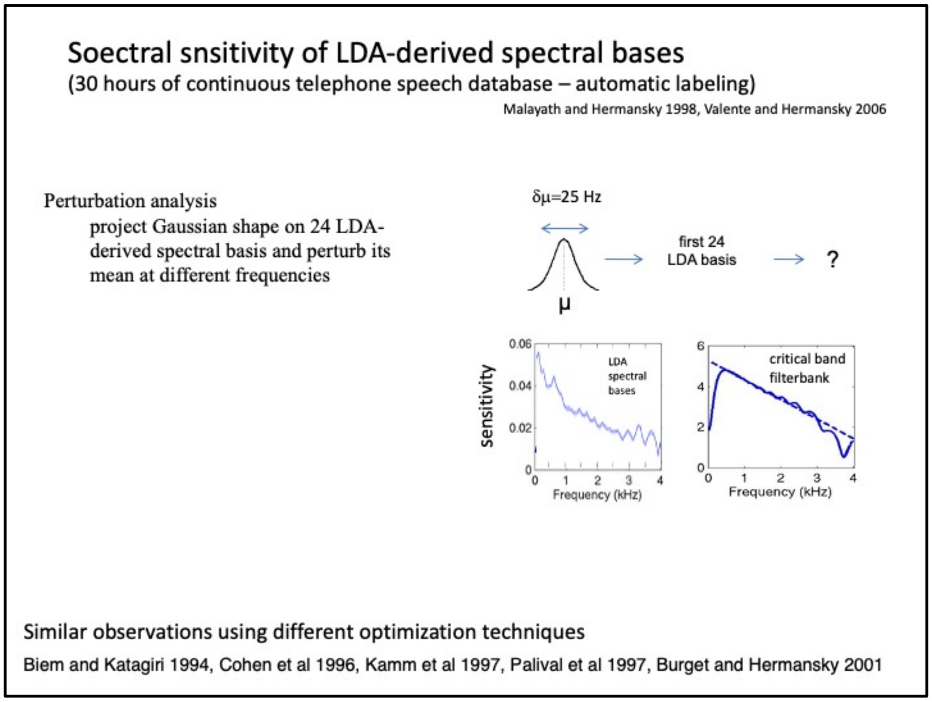


How do the first (most significant) discriminant bases look like?

One ve understand that if we have labeled speech data, LDA can be used to fiond optimal projections of a vector space to discriminate among given classes, we can ask what would be the optimal projection of short-time spectral vectors to optimize their use in classification of speech sounds. When we have labeled data, we know from which sound every spectral vector came, so we have the vectors space and the labels with each vector, so we can try LDA analysis.



These 4 discriminants represent about 75% of discriminating power. The spectral bases oscilates faster al low frequencies, this implies higher spectral resolution at low frequencies. Not much discriminability is carried beyond the 4-5 discriminants



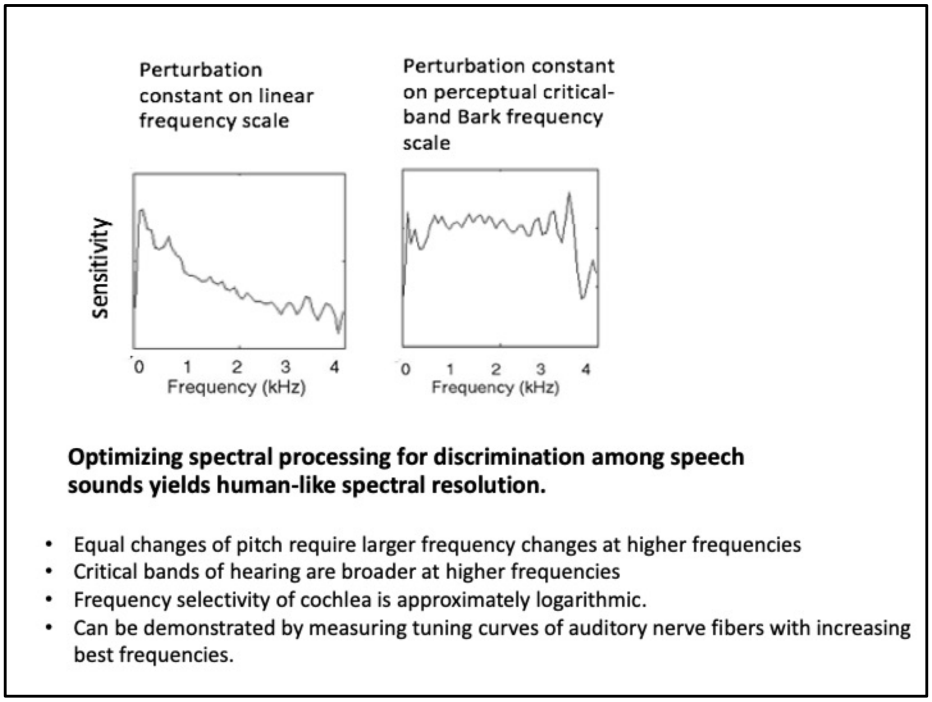
Spectral sensitivity of the derived LDA projections can be formally evaluated by the so called perturbation analysis. In this technique, a reasonable spectral object (here we used a Gaussan form resembling a resonant peak in spectral space with its mean at a given frequency) is projected on the new vector space. Its position is varied (perturbed) by a certain amount (here **Δ**f = 30 Hz constant at all frequencies) and variations in the projected output are evaluated. This yields the sensitivity of the projection at a given frequency. Evaluation of the variationa at different frequencies yields the sensitivity profile of the derived LDA projections.
Perturbation on linear frequency scale show higher sensitivity at lower frequencies.
One can do the same perturbation analysis on the spectrum modified by the auditory-like warping at in the mel spectrum or PLP. Looking at the PLP weigjhted spectrm, we observe similar trends as on the LDA-based projections.

Just a few known facts – most of you know by now after out classe on human perception better than me ☺:
Equal changes of pitch require larger frequency changes at higher frequencies
Critical bands of hearing are broader at higher frequencies
Frequency selectivity of cochlea is approximately logarithmic.
Can be demonstrated by measuring tuning curves of auditory nerve fibers with

increasing best frequencies.

No knowledge of human hearing used, just asking for optimal classification of speech sounds!

Human-like spectral sensitivity is optimal for classification of speech sounds. Known since early 20th century!



To double check. When the perturbationa are equal not on the linear scale (as in the previous experiment wher they were always were **Δ**f = 25 Hz) but constant on the perceptial Bark frequency scale (**Δ**f = 0.8 Bark), the output from the LDA projection is constant, this veryfying the lower spectral sensitivity at higher frequency, consistetly with the properties of human hearing.

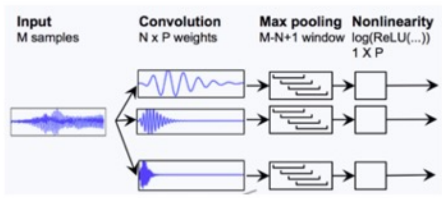
Just a few known fact abiut human hearing that we mentioned in this class:

- Equal changes of pitch require larger frequency changes at higher frequencies
- Critical bands of hearing are broader at higher frequencies
- Frequency selectivity of cochlea is approximately logarithmic.
- Can be demonstrated by measuring tuning curves of auditory nerve fibers with increasing best frequencies.

Such human-like spectral resultion emerging from just requiring optimal classification of speech sounds.

Using deep neural net classifiers, filters directly from speech signal

Sainath et al ASRU 2013



DNN-based ASR

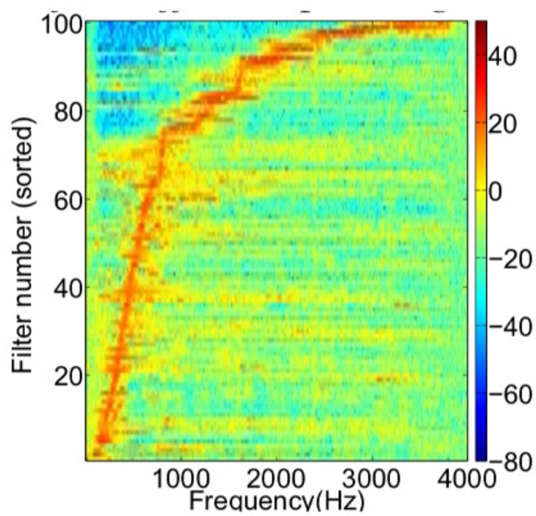
Q: what are the learned weights in the convolution input layer?
A: impulse responses of filters consistent with critical bands of hearing

also Palatz et al 2013, Tueske et al 2014, Golik et al 2015, Gharemani et al 2016, Luo and Mesgarani 2918, …

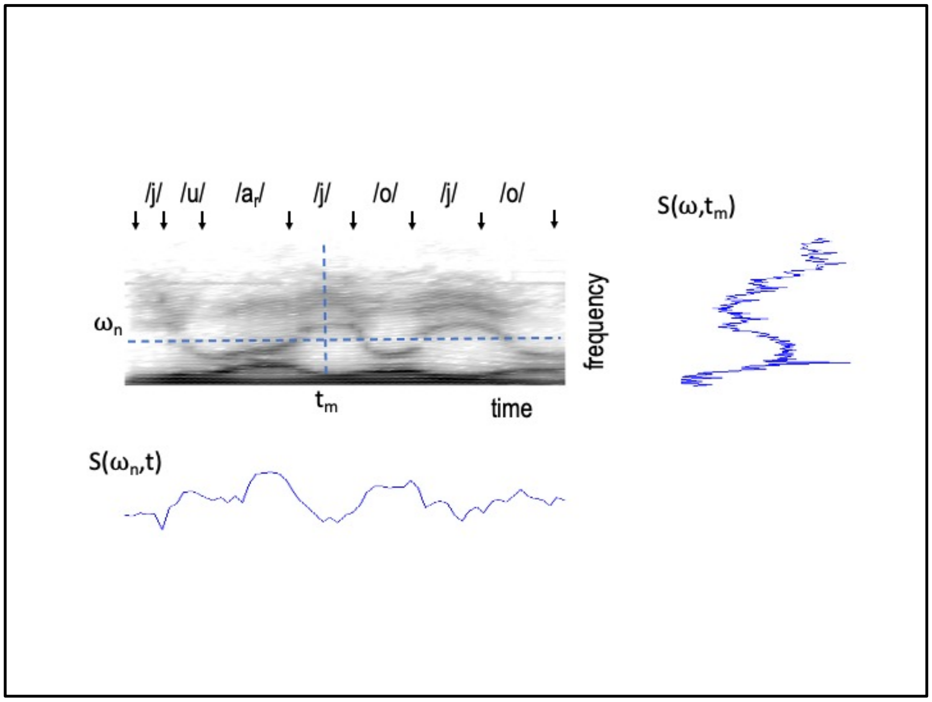
DNN architecture can be set up in such a way that on thei DNN input nodes are weightings of the data which can be interpreted as the finite inpulse response (FIR) filters (signal convolutions for the FIR filters). After the DNN system training, the FIR filters can be examined.

Magnitude response of learned filters ordered by center frequency

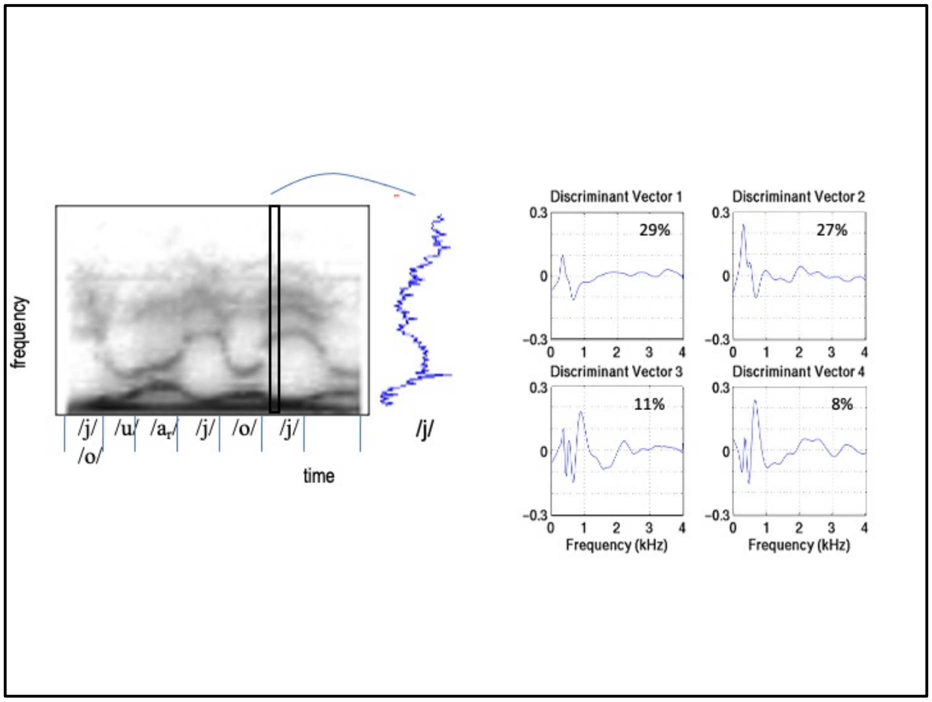
Gharemani et al 2016



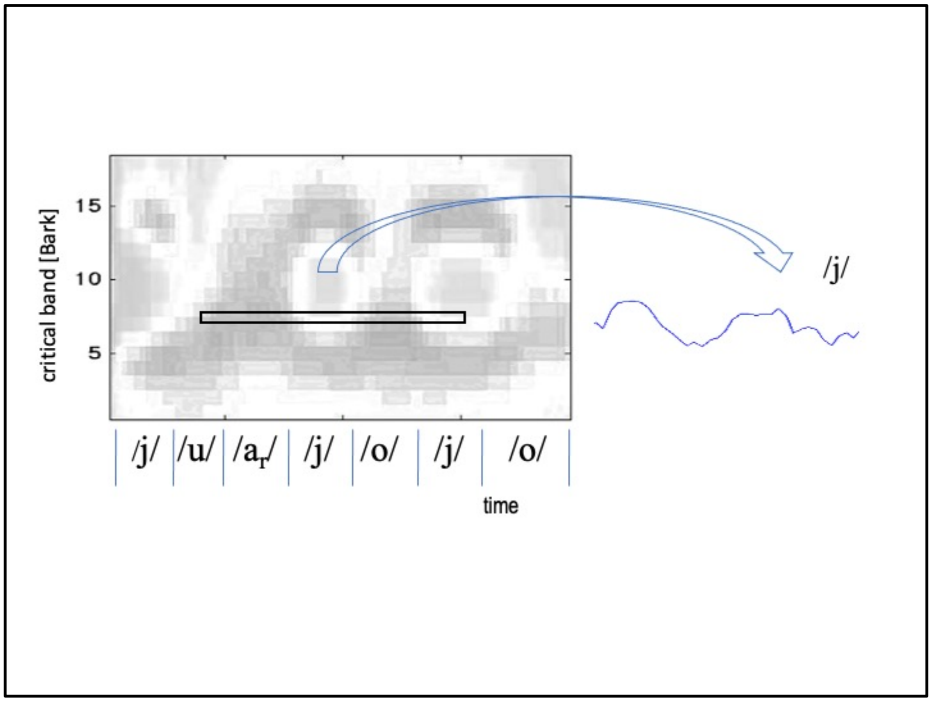
The derived bank of FIR convolutive filters typically show human-like spectral resolution.



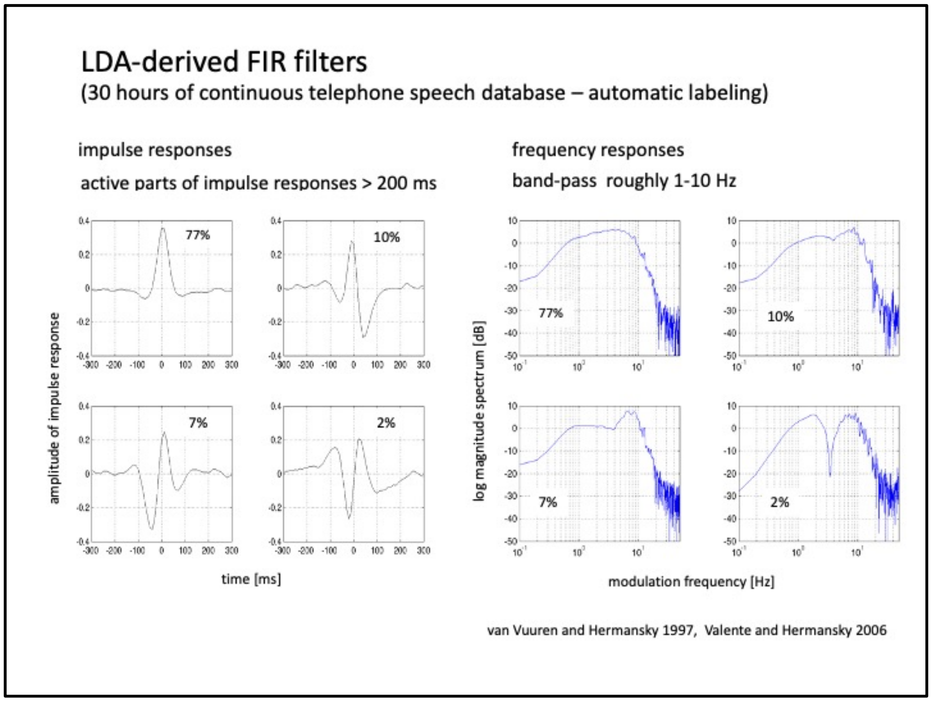
Spectrogram is 2D representation of speech signal
Y-axix carries short-time spectra of speech at a given time
X-axis carries evoltions of spectral energies at a given frequency over time



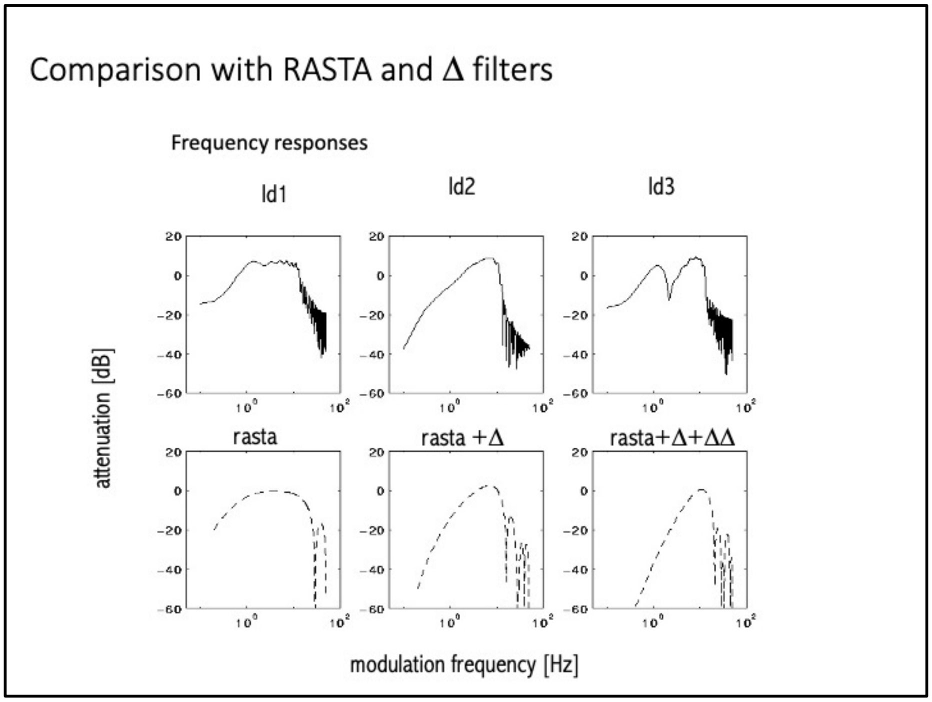
LDA on short-time spectogram suggested the use of human-like spectral resolution (critical-band like)



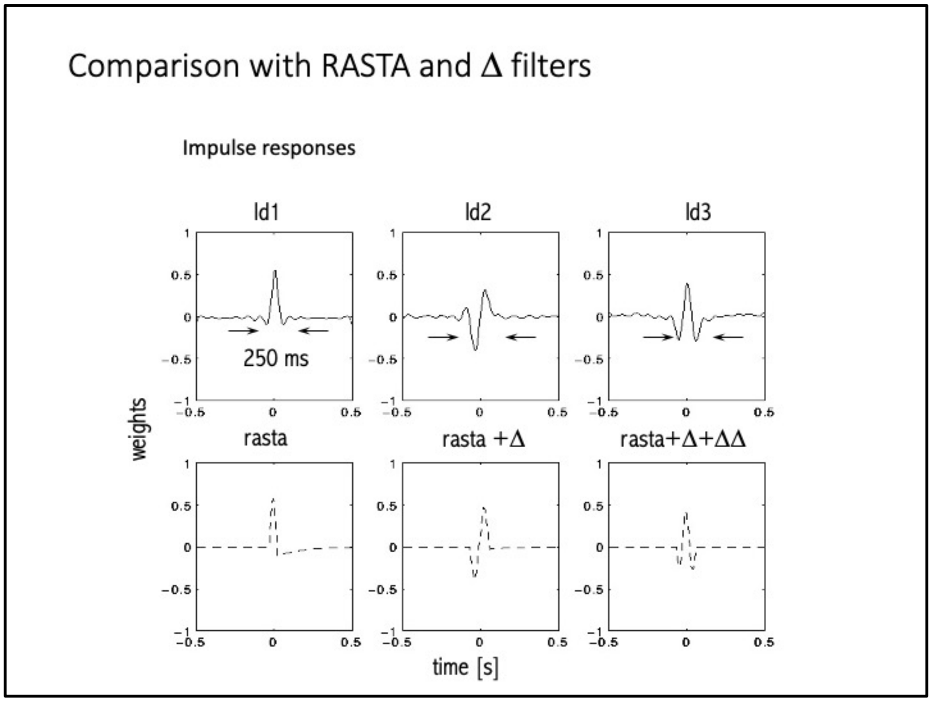
Now we know that we should use critical-band-like spectrum. Using such spectrum, we can re-use the same technique for deriving optimal modulation frequency filters. This time the derived discriminant matrix will surrect optimal weighting of neigbouring spectral values to optimize the speech spound classification. For those skilled in DSP – this represents impulse responses of optimal finite impulse response (FIR) filters.



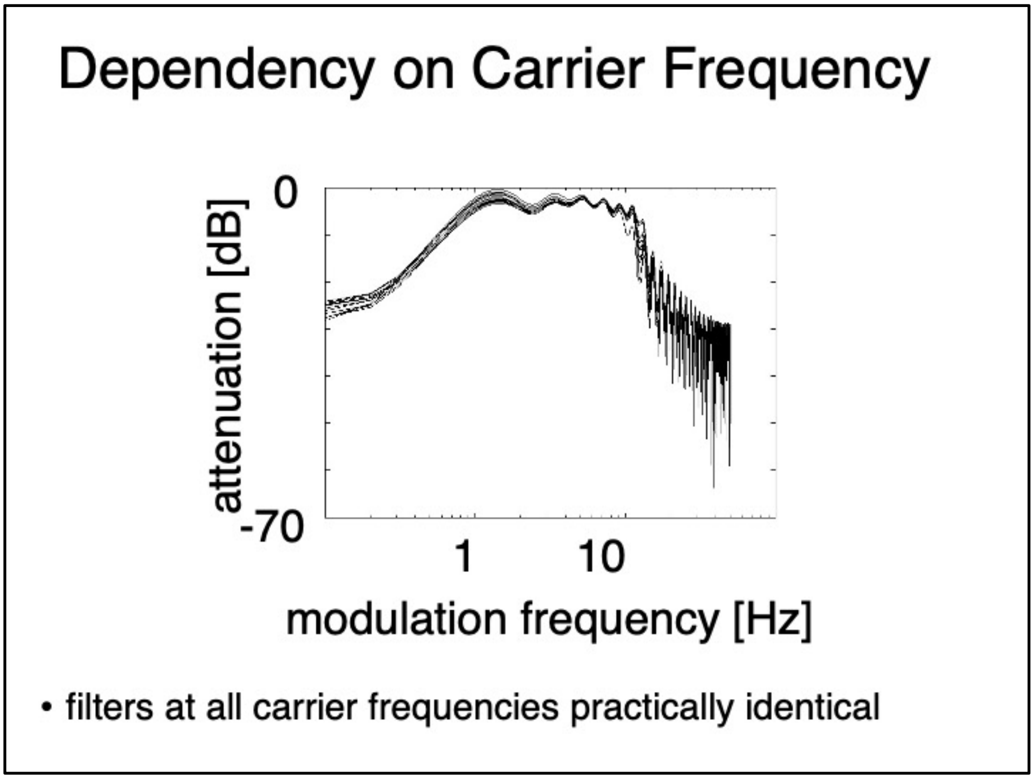
Impulse responses are rather long – more than 200 ms. Frequency responses show that filters suppress very slow modulations as well as modulations faster than around 10 Hz. This is similar to RASTA filters. HOwever, the impulse responses are summetroc, implying the zero-phase linear FIR filters.



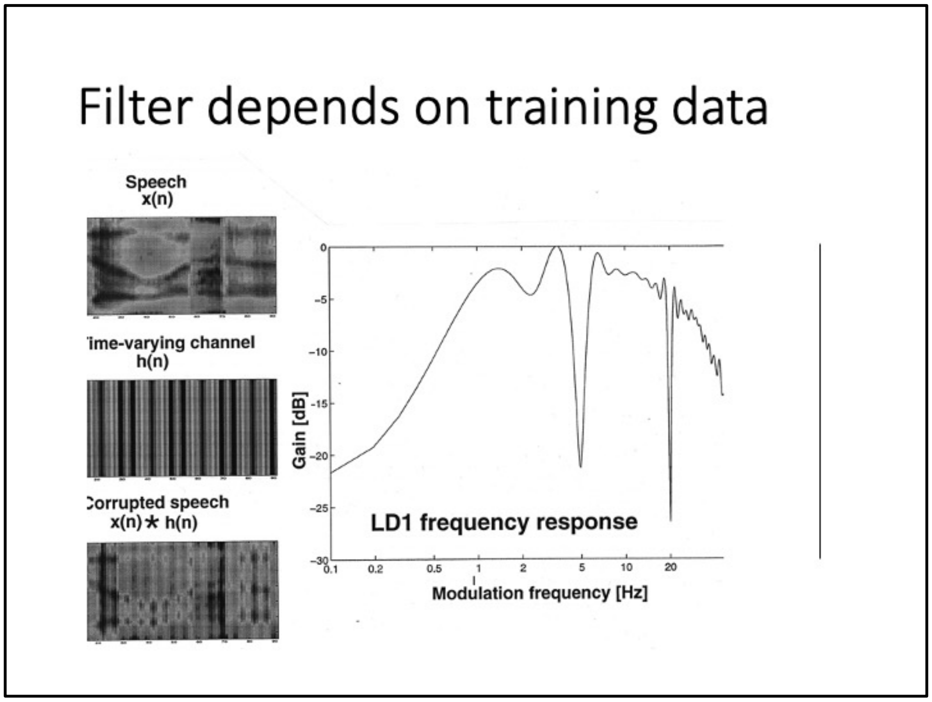
RASTA features were typically more effective when usrd with the dynamic (differential and double-differential) features. I is iinteresting that the second and the third discriminant from thr temporal LDA filter design resemble the combination of the RASTA and dynamic RASTA features.



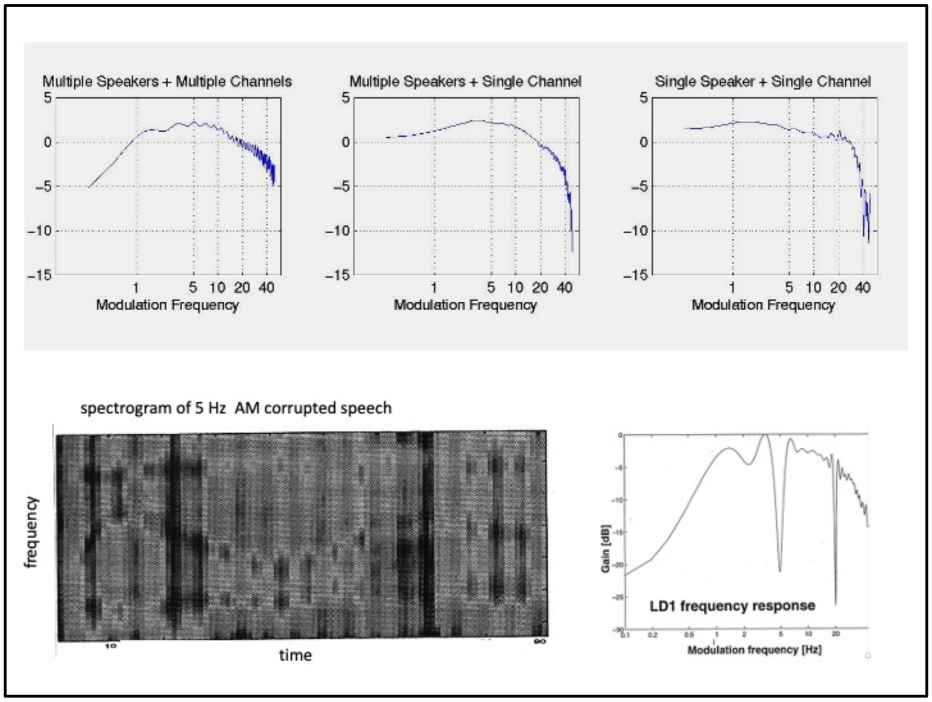
RASTA features were typically more effective when usrd with the dynamic (differential and double-differential) features. I is iinteresting that the second and the third discriminant from thr temporal LDA filter design resemble the combination of the RASTA and dynamic RASTA features.



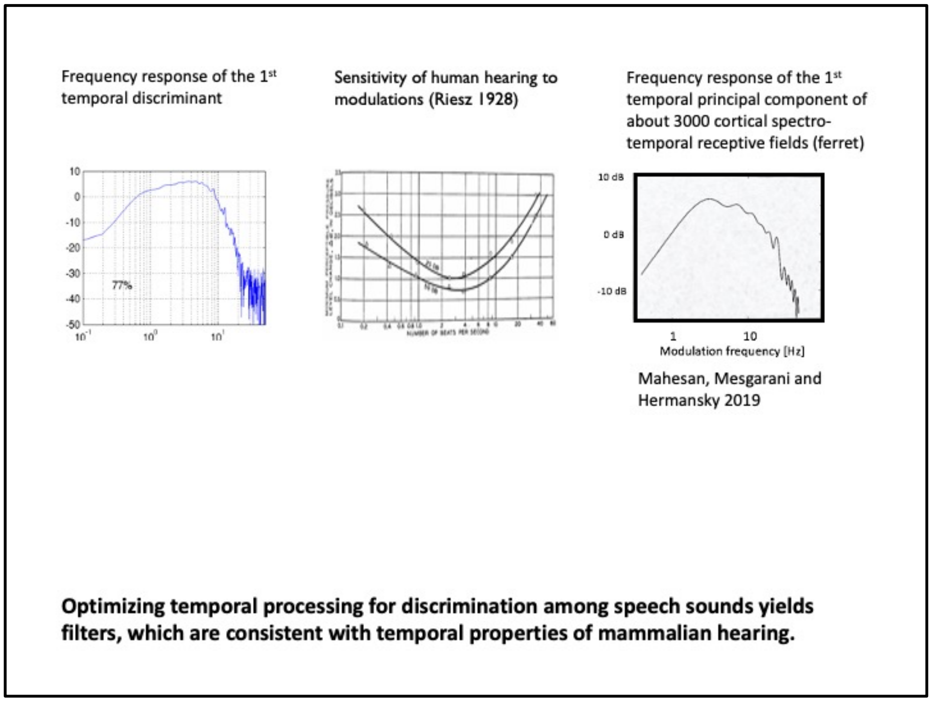
An interesting question is if the filter properties are dependent on frequency trajectories (carrier frequencies) to which they are applied. The ansver seems to be ”no”.



As expected, the filters reflect the data on which they were derived. When the data were artificially corrupted by amplitude modulating the speech signal at 5 Hz, the filter emerged with the noth at % Hz.

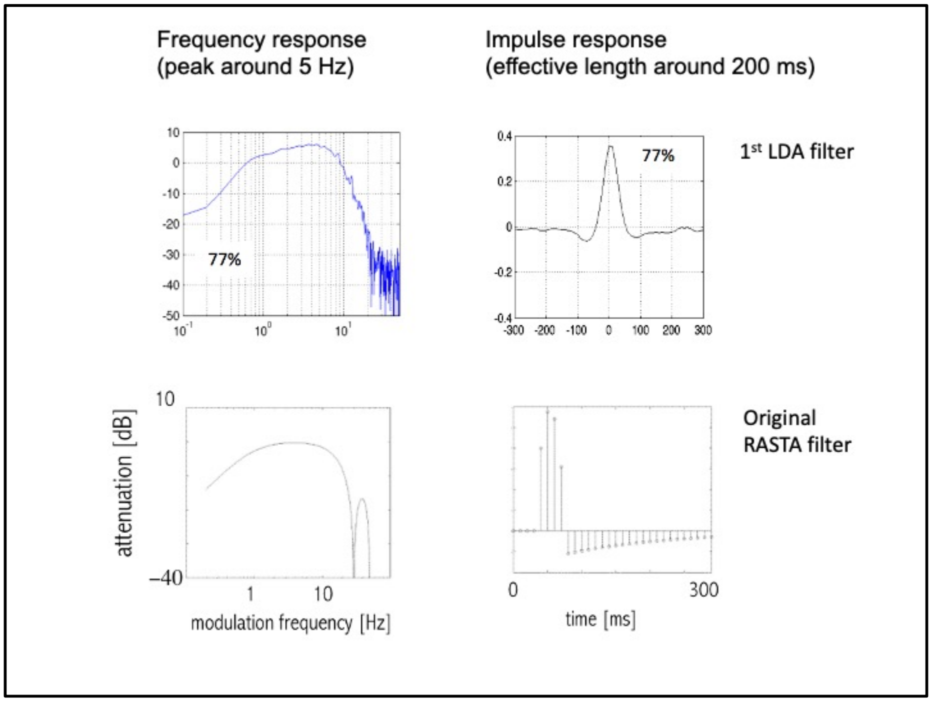


However, they are dependent on the type on speech material on which they were derived. For the general case where different speech utterances come from different speakers and different environments The filters are noticeably suppressing modilation frequencies below 1 Hz. When the hannel variability is not present (all speech spamles comning from the same acoustic envilronment) the low modulation frequeny supression is much weaker. Whe the speaker variability is not present (all files from the same speaker), low modulation frequencies are preserved.

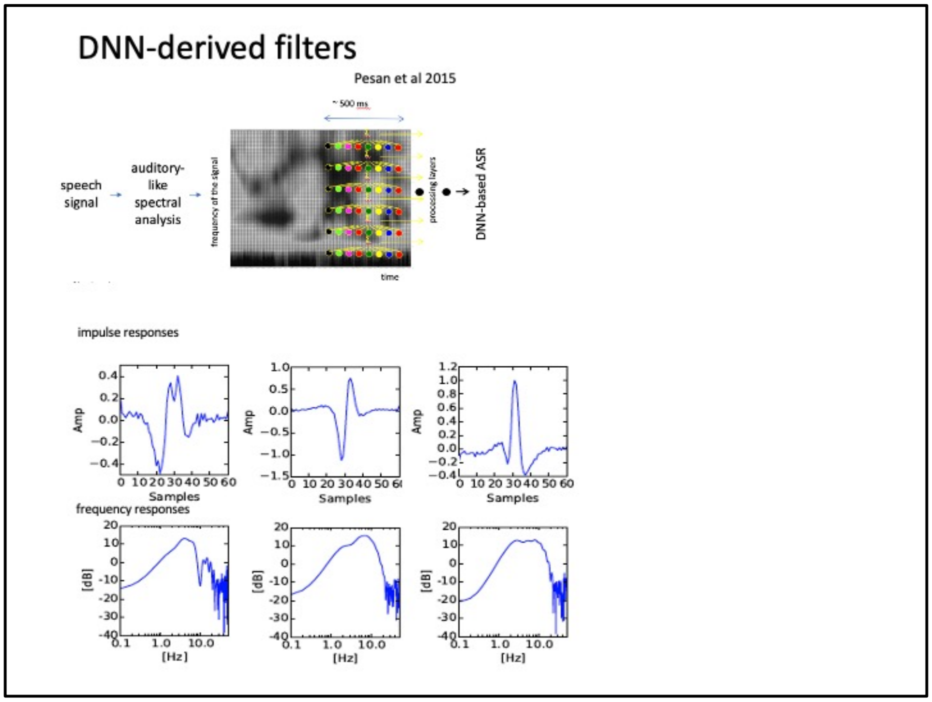


Just a few facts from human hearing:
Sensitivity of human hearing to modulations is highest around 5 Hz
-known to speech engineers since 1928
Auditory cortex seems to be the most sensitive to slow modulations.

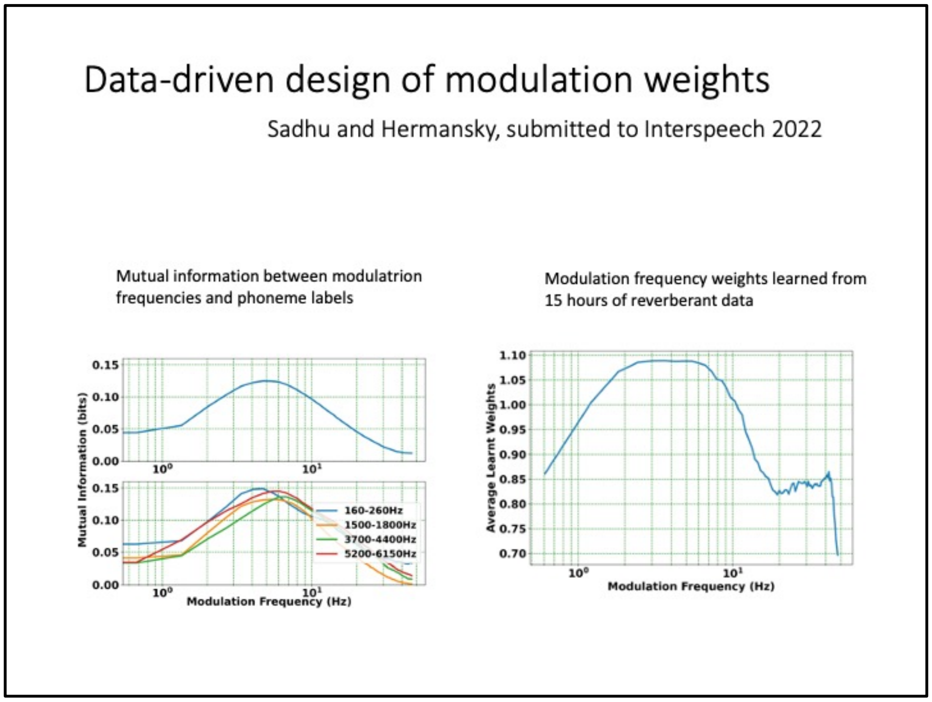
No knowledge of human hearing was used to derive temporal features directly from speech data, just asking for optimal classification of speech sounds!



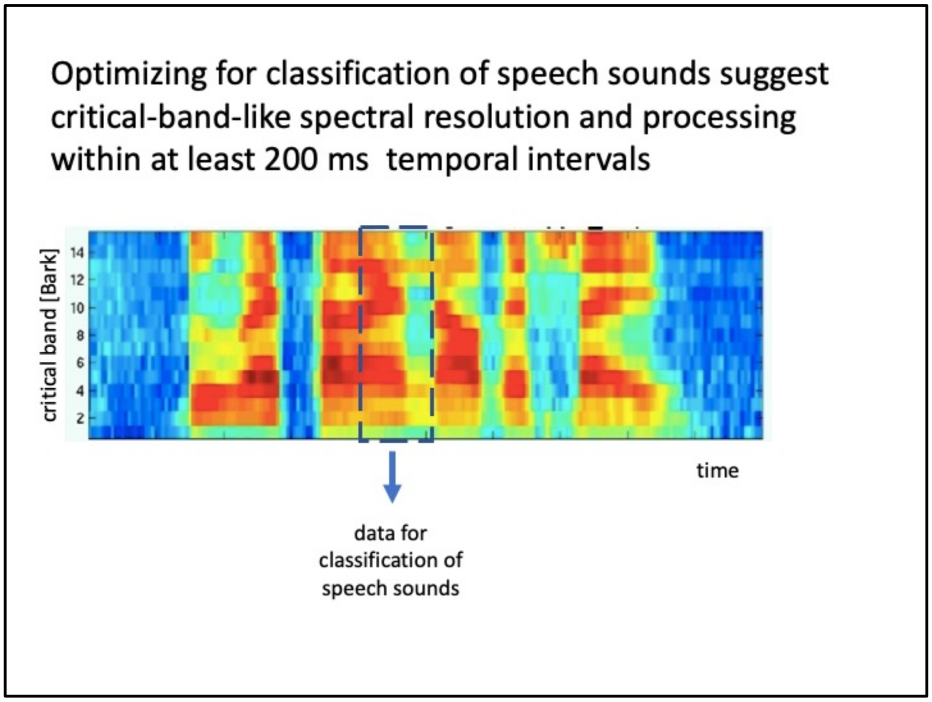
Frequency response
(peak around 5 Hz)
Impulse response
(effective length around 200 ms)
77%
77%
1st LDA filter
Original
RASTA filter
attenuation [dB]
modulation frequency [Hz]
time [ms]



Impulse responses are rather long – more than 200 ms.
Frequency responses show that filters suppress very slow modulations, as well as modulations faster than around 10 Hz.

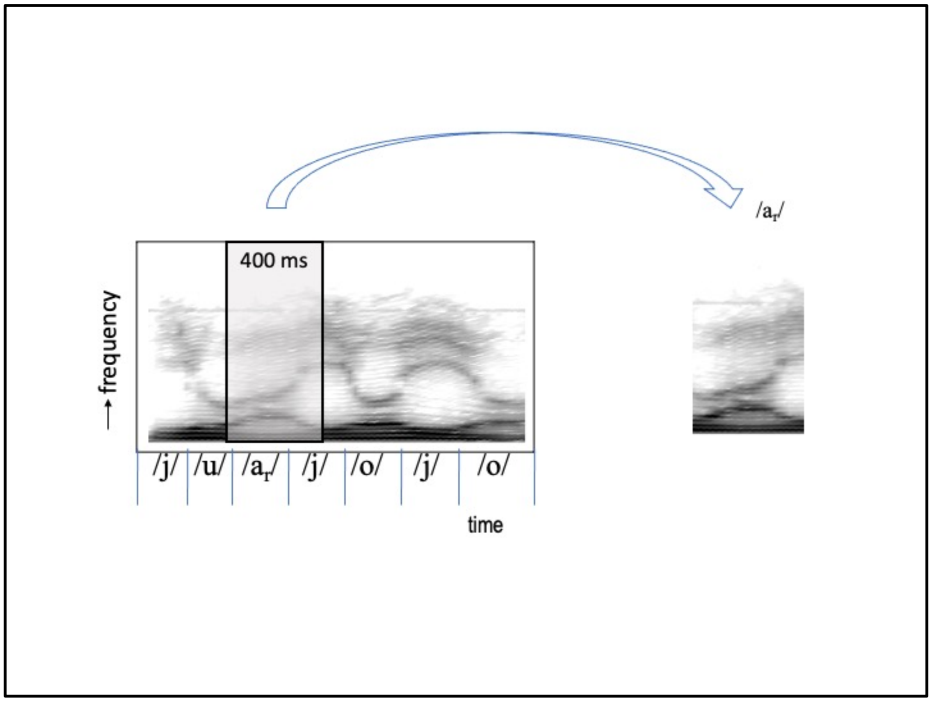


Mutual information studies nd data derived modulation frequency processing can be also studied, conforming relations between linguistic imformation carried in modulation frequency and optimal modulation frequency weighting derived by optimizig speech recogntion system during the machine training.

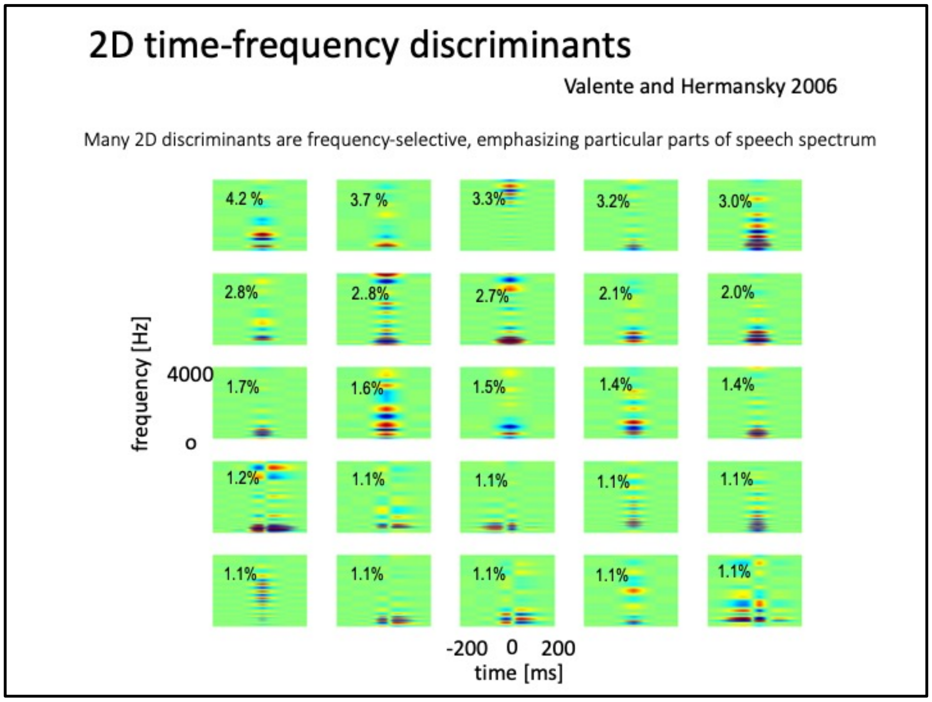


Does all that mean that we should use critical-band spectral resolution
(of course we all know that)
and temporal spans larger than 200 ms
(that is what most advanced ASR systems started to use now).

Should we probe even further for appropriate signal pre-processing?

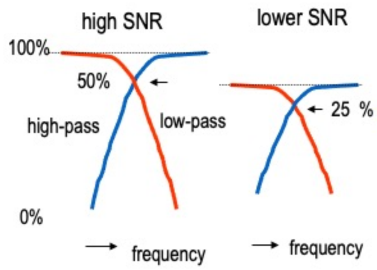


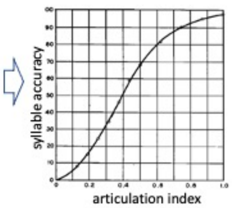
Using the same LDA technique on larger segments of LPC-smoothed spectra should give 2-D spectro-temporal (cortical-like??) projections for next stages of phoneme classification.



Spectral resolution is typically coarser at higher frequencies. Temporal spans of filters typically cover more than 200 ms. We knew that from our earlier LDA spectral and temporal optimizations.

However, most interestingly, different filters often focus on different parts of speech spectrum !! This suggests that the optimal set of time-frequency derived features would be the features which look at different temporal spans of speech signal (up to 200 msec) and often do not cover all fequencies, each feature focusing only on some frequencies of the speech spectrum.





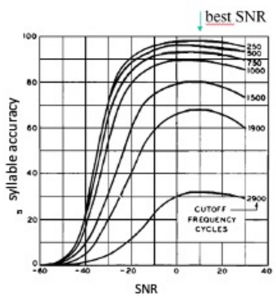
approximated by

$$\mathcal{A}(s) = \frac{log10(1-s)}{log10(1-S_{MAX})}$$

s = phoneme accuracy

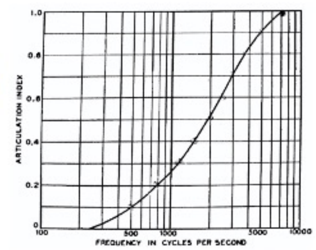
articulation index as a sum of logs (product) of phoneme errors in frequency subbands

High-pass filtering and SNR changes



read accuracies for different cutoff frequencies and convert to articulatory index

$\mathcal{A}$(frequency)

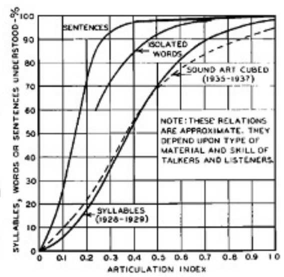


Articulatory Bands

French and Steinberg 1949

250-375-505-654-795-995-1130-1315-1515-1720-1930-2140-2355-2600-2900-3255-3680-4200-4860-5720-7000 Hz

- 20 frequency bands in speech spectral region
- each band with SNR > 30 dBcontributes equally to human speech recognition
- bands with SNR < 0 dB do not contriute at all
- any 10 bands sufficient for 70% correct recognition of nonsense syllables, better than 95% correct recognition of meaningful sentences [Fletcher and Steinberg 1929]



7000

FREQUENCY [Hz]

1800

0

TIME [s]

This turned out to be true not only for two bands, but also for more bands (up to 20 bands). Thus, the multichannel model predicts that the total error will be given by

$$e = \prod_{i=1}^{K} e_i$$

- SERIAL PROCESSING
  - phonemes in a nonsense syllable are decoded independently of each other S=c.v.c (**probabilities of correct recognition multiply**)
- PARALLEL PROCESSING
  - errors in phonetic judgment in nonsense syllables in individual sub-bands are independent (**probabilities of errors multiply**)