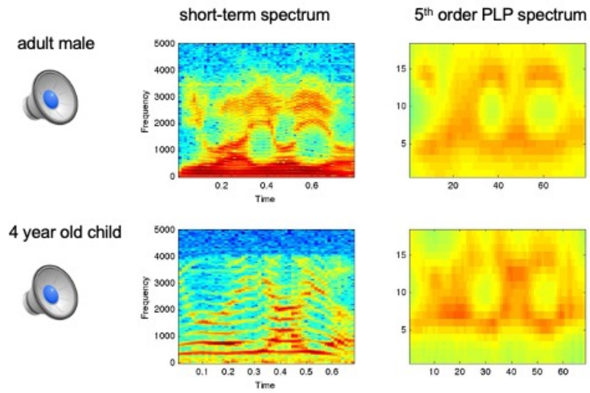
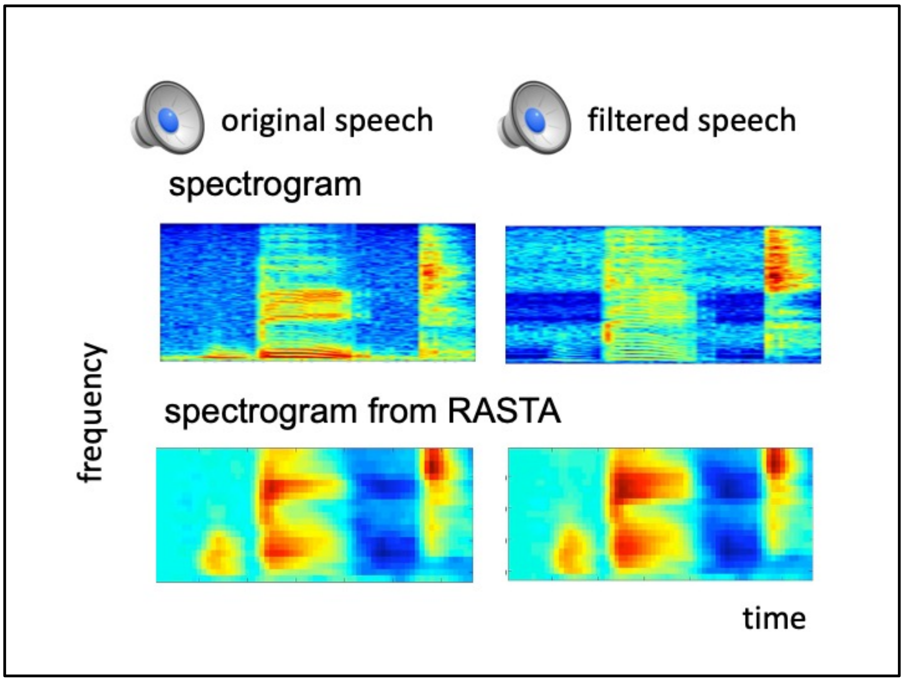
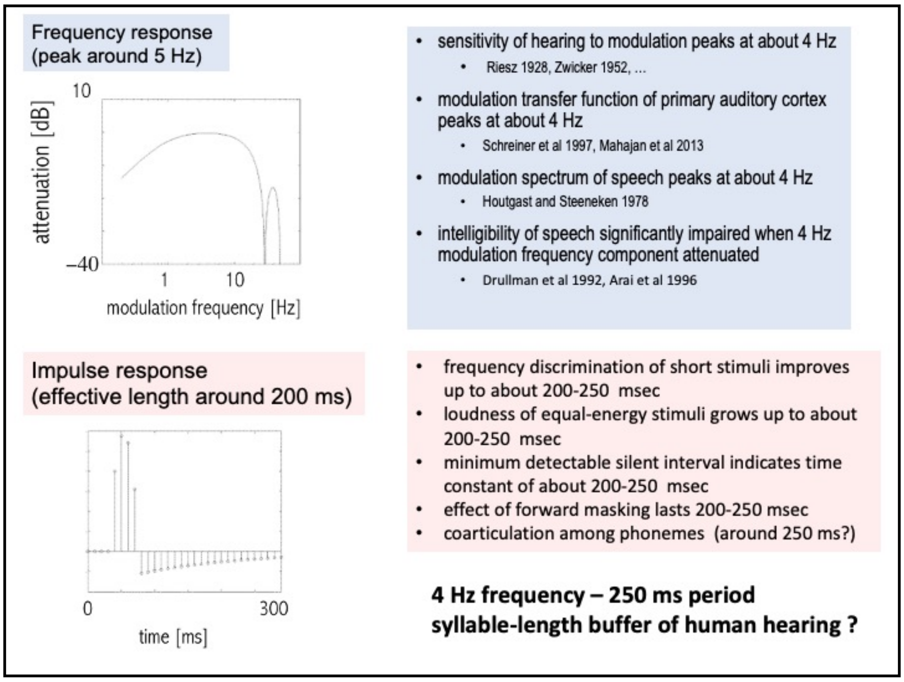
Auditory-like spectral resolution and considerable spectral smootining




original speech
filtered speech
spectrogram
frequency
spectrogram from RASTA
time



To summarize: The RASTA processing, which involves band-pass filtering of temporal trajectories of spectral energies between two nonlinearities, one being compressive (logarihm) and another expansive (exponential), intially developed to address linear distortions in speech, turmned out to be consistent witth several phenomena observed in human hearing.

RASTA processing enhances modulations beween 1 and 12 Hz, which is consistent with human sensitivy to modulations (known since 1923). It is also consitent with the estimated temporal properties of hearing, seen in estimated impulse responses of auditory cortical receptive fields. It is also known that in this region, the energy in modulations in speech is the highest. Finally, it can be shown that when this region of speech modulation is removed from the signa, the intelligibility of speech as well as the accuracies of its machine recognition, decrease.

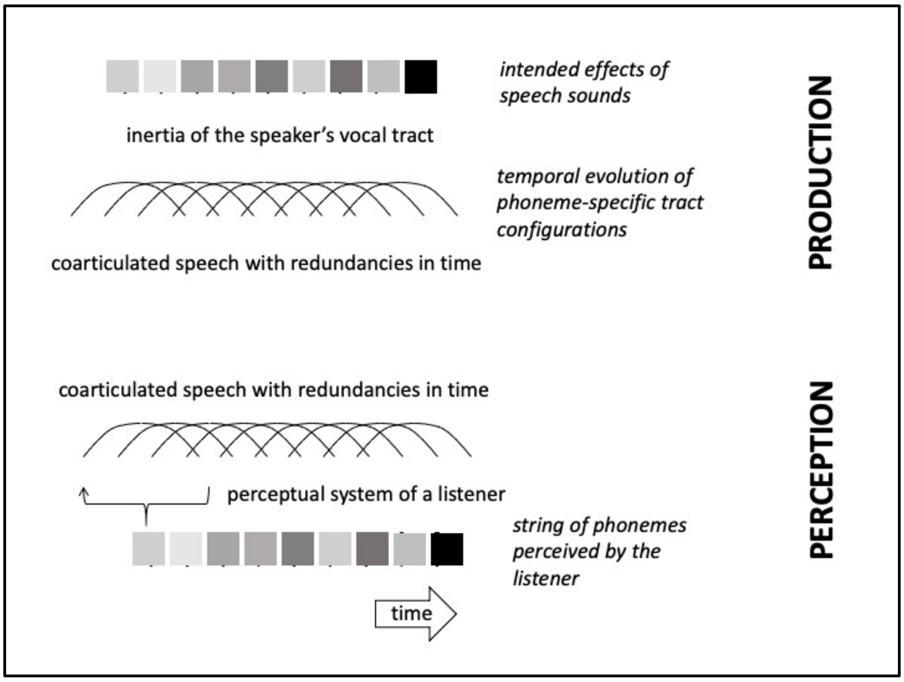
The effective length of the impulse response of the RASTA filter agrees with the concept of temporal buffer in hearing, observed, e.g., in frequency discrimination of short stimuli and in their loudness increase, and with the implied hearing inertia, seen in detection of the gaps in noise and in the temprla forward maskin. RASTA also provides a reasonable model of the forward masking.

Speech Production

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

4

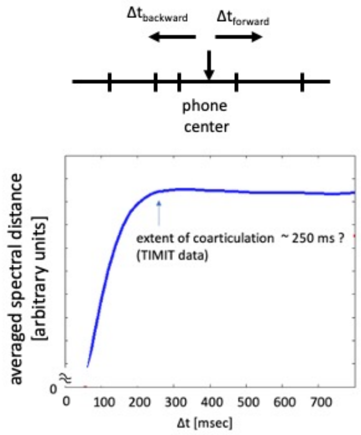
As stated earlier, is is likelu that human speech evolved to emply some properties of human hearing. It would be interesting to see what is the dominant frequency of vocal tract movements in production of speech.



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.

Extent of coarticulation ?

Hermansky 1996 (proceedings DoD Summer worskhop)



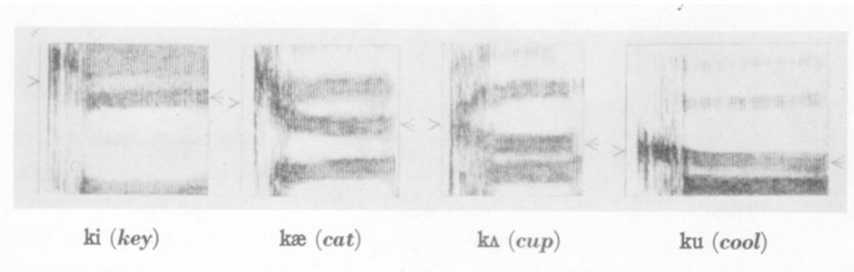
$\Delta t_{forward} \cong \Delta t_{backward}$

$\Delta t = \Delta t_{forward} + \Delta t_{backward}$

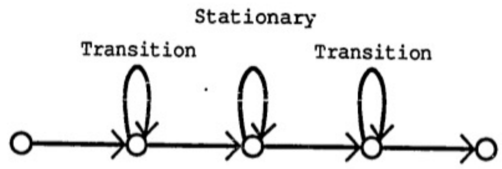
250 ms
one period of 4 Hz modulations

thanks to Katia Yegorova, BUT, Czechia

Real Speech



Coarticulation is our enemy.



HMM model of **context dependent** speech sound (Lee 1988)

Coarticulation is our friend.

Recognizing vowels in German syllables **/d(vowel)t/** presented in a carrier phrase “Ich habe ***syllable*** gesagt.”

- vowels are better recognized from the coraticulated neigboring consonants then from the isolated vowel segments

~100 % accuracy

•••• /d/ (vowel) /t/ ••••

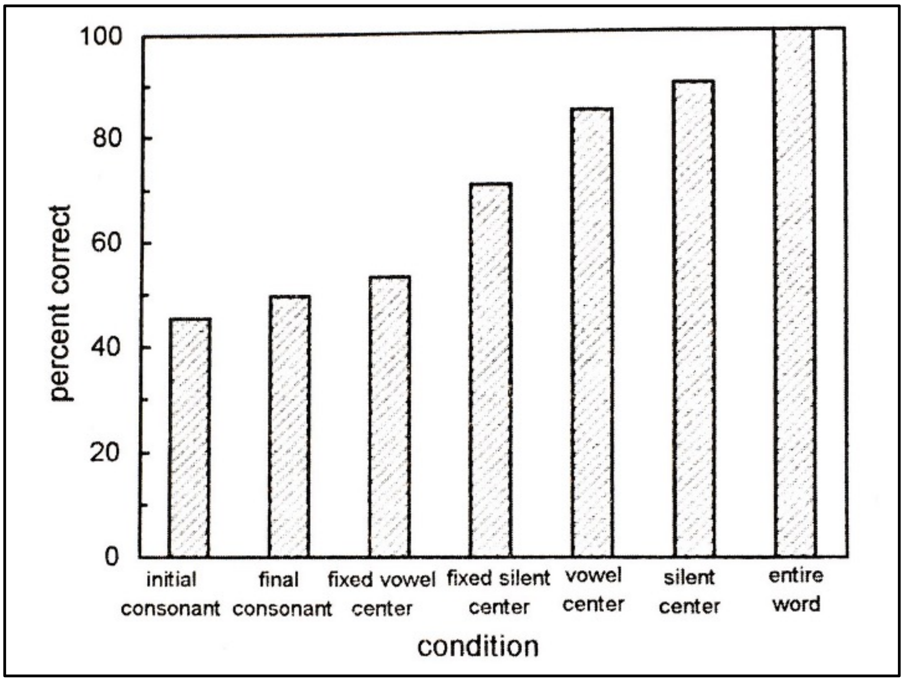
~90 % accuracy

•••• /d/ silence /t/ ••••

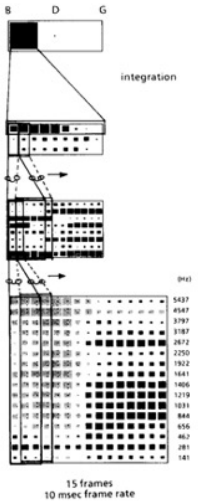
~85 % accuracy

•••• silence (vowel) silence ••••

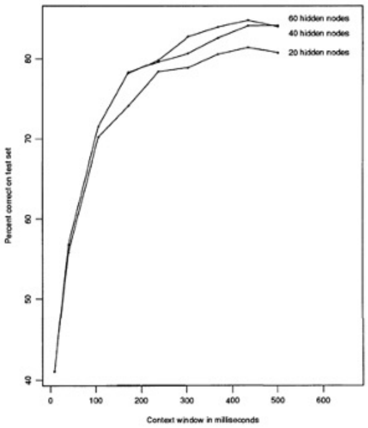
Strange and Bohn 1998


100
80
60
40
20
0
percent correct
initial consonant
final consonant
fixed vowel center
fixed silent center
vowel center
silent center
entire word
condition

TDNN
Waibel et al 1989



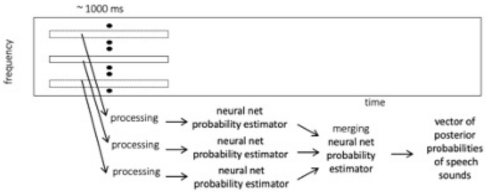
Fanty, Cole, Roginski NIPS 1992



TRAPS

Hermansky and Sharma, ICSLP 1998

Classifying **T**empo**RA**l **P**atterns of **S**pectral Energies



13 telephone quality isolated digits, clean and arificially corrupted by 4 different additive noises

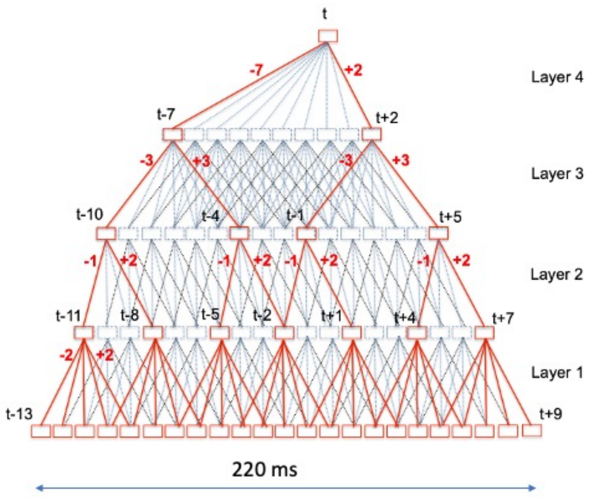
conventional ASR 22.5 % error
combined with TRAP 19.7.% error

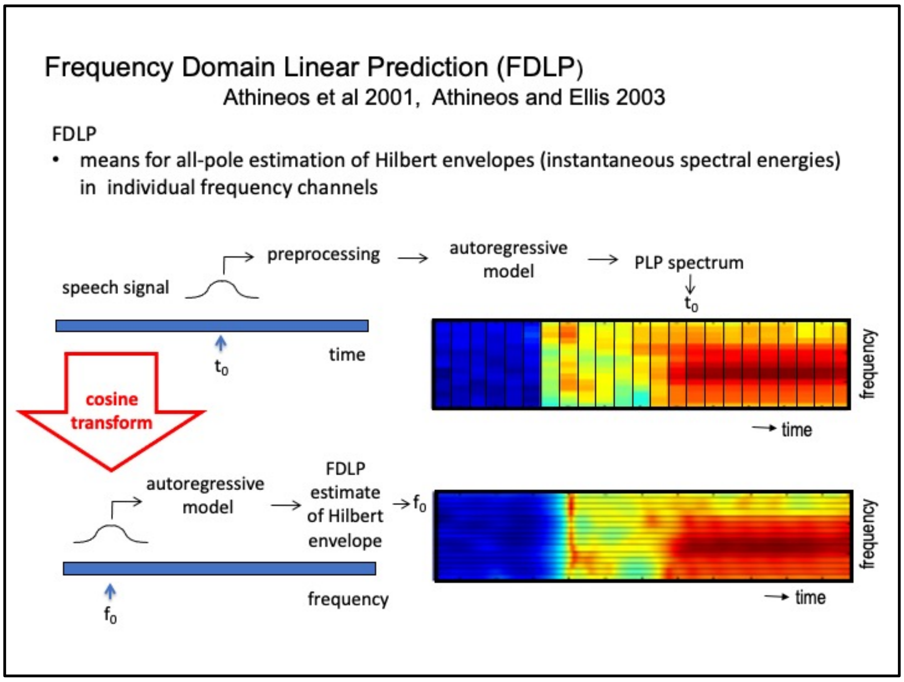
Hermansky and Sharma, ICASSP 99

Some "novel" (in 1998) elements of TRAPS

- Rather long temporal context of the signal as input
- Hierarchical structured neural net ("deep neural net")
- Independent processing in frequency-localized parallel neural net estimators
  - most of these elements typically found in current state-of-the-art speech recognition systems

Modern TDNN (Peddinti et al 2015)

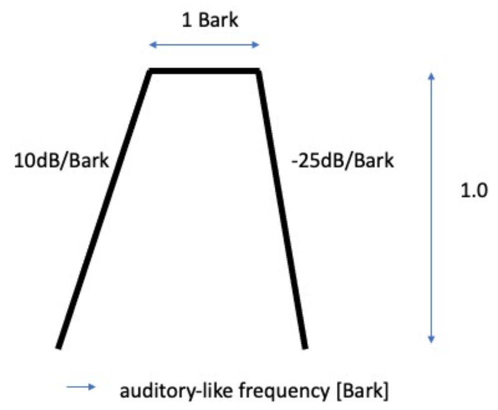


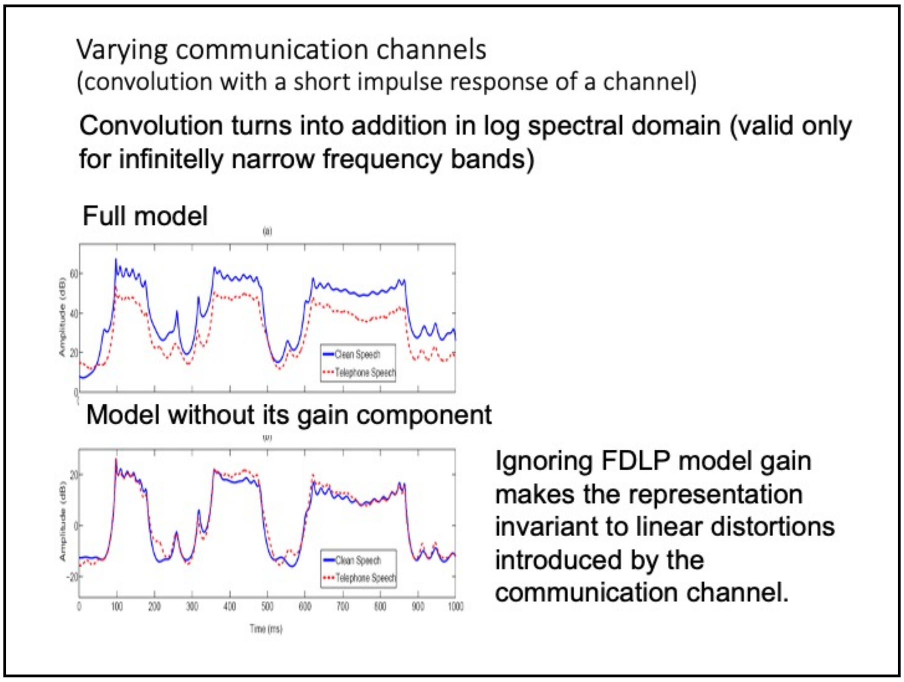


Schematically, the time domain linear prediction (TDLP) of the PLP auditroy spectrum is shown on the upper part of the figure.
For conventional PLP analysis, the speech signal is segmented to short segments by windowing the signal. For every processing frame short-time auditory spectrum is approximated by autoregressive model. PLP spectra represent columns of the spectrogram.

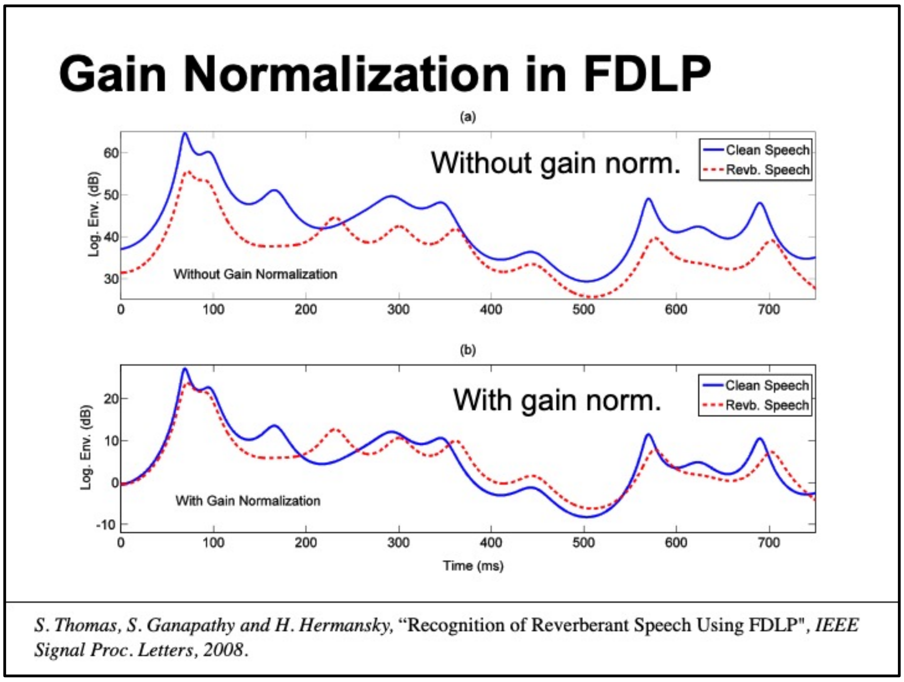
For the frequency domain linear prediction (FDLP) , the speech signal is transformed by cosine transform to frequency domain. WIndows on the cosine transform select desired frequency bands to be processed. Each selected band is processed by the FDLP to yield all-pole model of the squared Hilbert envelope in the particular frequency band. These estimates represent rows of the spectrogram. The better moleing of temporal events in the signal by FDLP is seen here.

Windows on cosine transformed signal

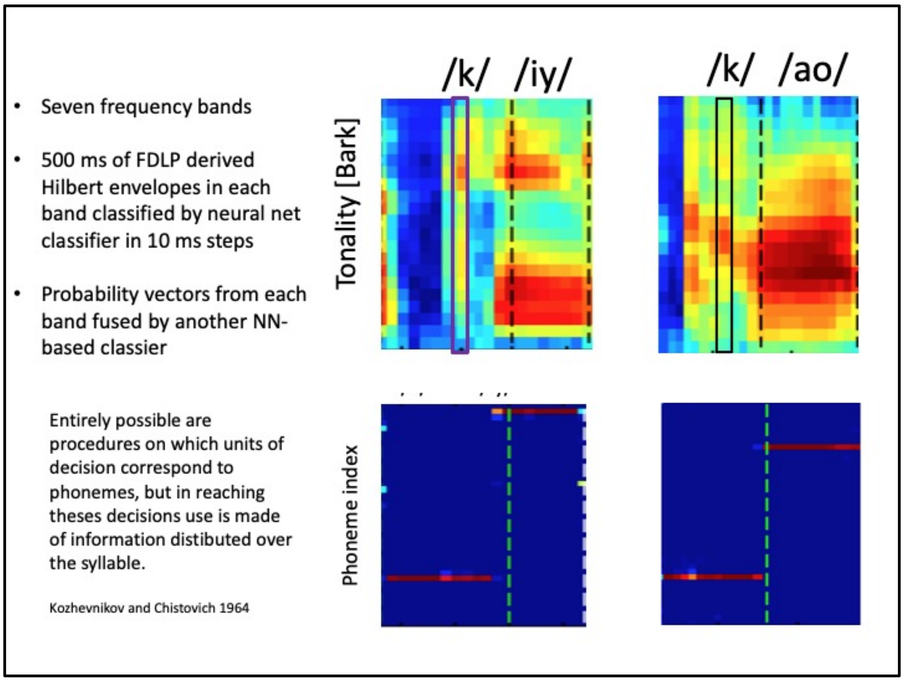




Linear distrortions induced by differences in frequenct responses of recording environments show mainly as different DC biases at different frequencies and are mainly reflected in gains of the FDLP models at different frequencies.
By ignoring the DC (gain) in the model, the FDLP approximations may be made less sensitive on the linear distrortions.



THis is partially true even for speech distortionas created by reveberatioons which are mainly caused by convolutions of the signal with rather long impulse responses of reverberant rooms. SInce the time spans over which the FDLP model can be quite long, event the effect of the long reverberations can be partially handeled.

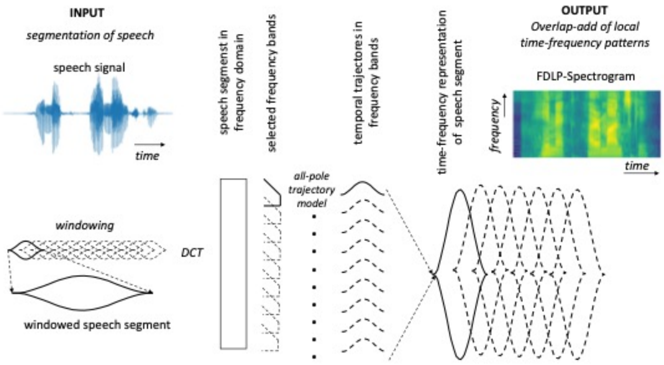


At the first stage, the fullband speech signal is decomposed into seven band-limited streams. A independent threelayer MLP is trained to discriminate the phonemes based o the temporal modulation feature of input narrow-band signal. The MLP phoneme classifier generates a 40 dimensional posterior probability vector. Each item of the posterior vector represents the posterior probability of a particular phoneme given the acoustic evidence. Since each stream only provides marginal information, the seven band-limited streams are fused by a three-layer MLP at the second stage to give a more reliable estimation of the target sound.

The narrow-band speech in each band-limited stream is represented by the frequency domain linear prediction (FDLP) feature [9], which provides a parametric representation of the Hilbert envelope of subband signal. Long-segments (3–5 seconds) of speech are decomposed into critical bands, by multiplying the discrete cosine transform (DCT) coefficients with a set of windows (Eq. 1). The subband temporal envelopes are approximated by an all-pole model using FDLP. The subband envelopes are divided into short frames by multiplying with a 500 ms hamming window every 10 ms. Next the segments of envelope are converted into modulation spectral components by DCT transform, and then concatenated to form a feature for the band-limited strea

FDLP spectrogram

Sadhu and Hermansky 2021



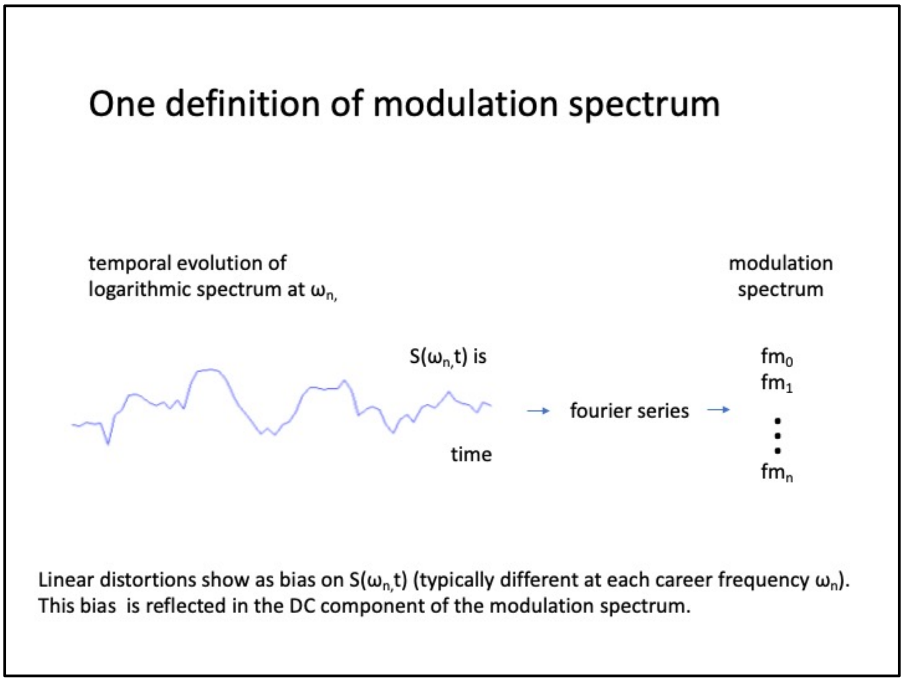
Results

	Wall Street Journal (WER %)		
	clean	street noise (20dB)	babble noise (20 dB)
Guo et al. *	4.9	-	-
Our mel baseline	5.1	24.7	75.2
FDLP	**4.8**	**20.4**	**56.1**

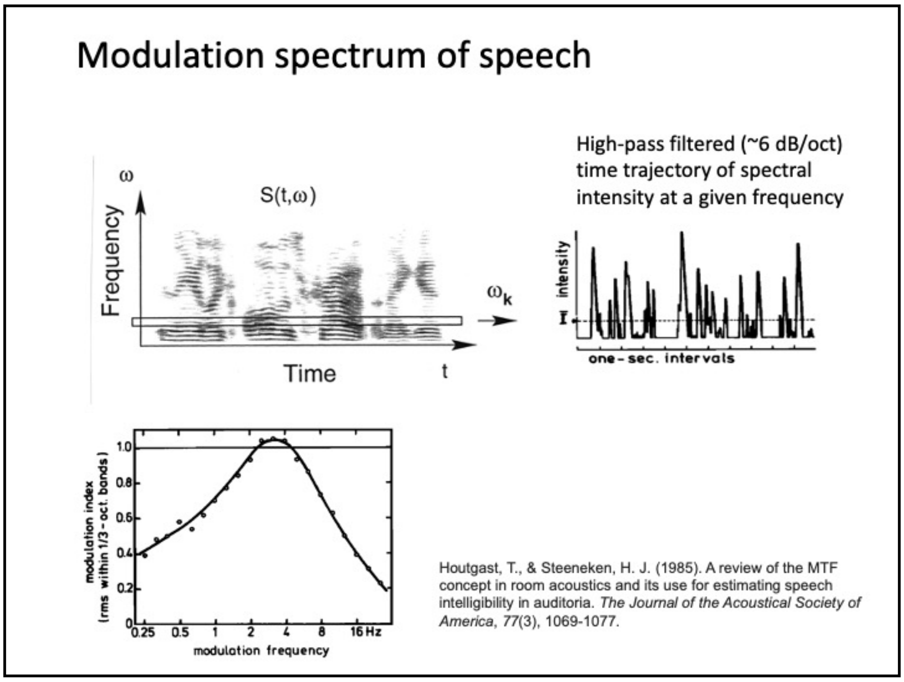
- *Clean read speech training and test data*
  - ***Model gain included***

	REVERB (WER %)		
	8 channel	single channel	Weighted Prediction Error de-reverberated single channel
Guo et al. ICASSP 21	14.3	-	-
Our mel baseline	9.2	23.2	20.7
FDLP	**7.2**	**19.4**	**18.0**

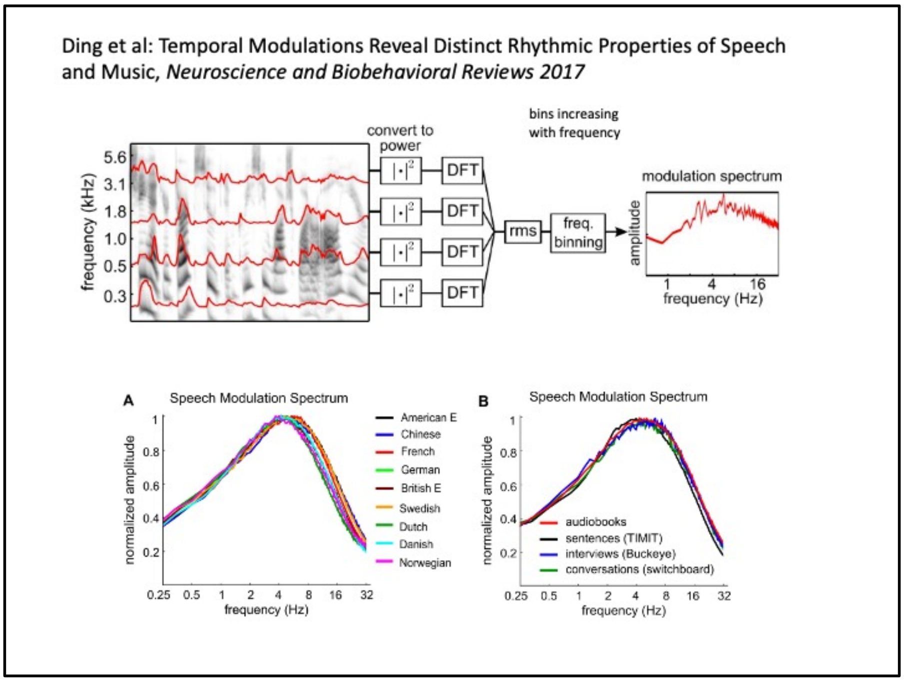
- *Simulated reverbeation in training, real reverberated test data*
  - ***Model gain left out***



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.

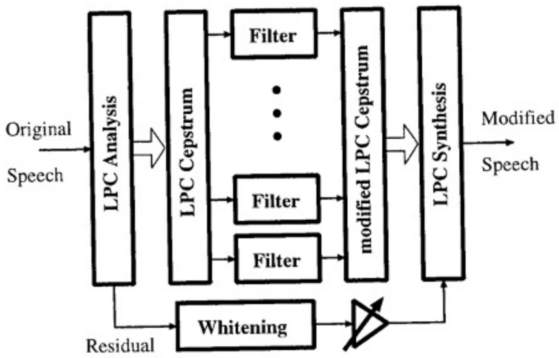


Modulation spectrum of speech can be computed by computing spectrum of temporal evolution of spectral envelope at a given frequency. Modulation frequency has a typical decline at about -6 dB/oct. Therefore when computing the modulation spectrum, this decline is being compensated for by differentating 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).

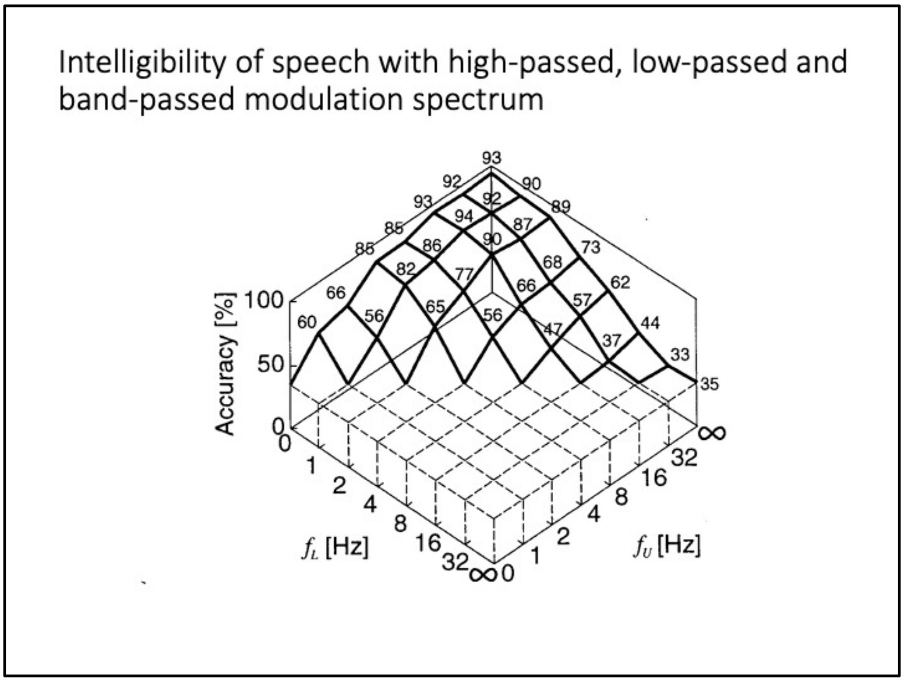


This is one of the more recent results where the overal modulation frequency of speech data from different languagaes and different American English speech databases are shown. It is striking how similar are the modulation spectra from these different databases.

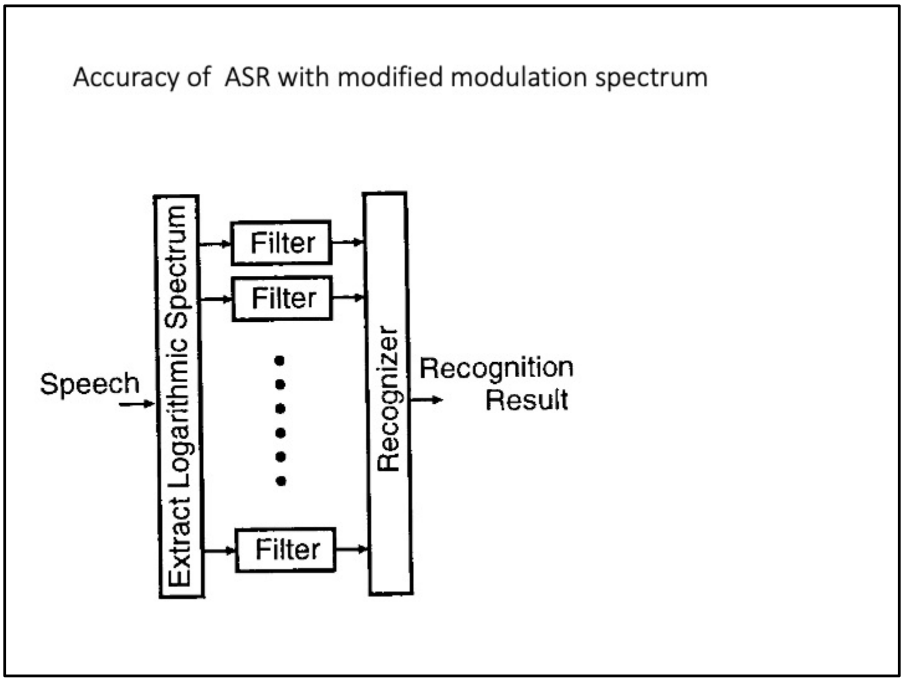
Intelligibility of speech with modified modulation spectrum



Tjis system allows for limiting the rate of changes of spectral envelopes. Residual exciting vocoder separates contributions of the voice source and of the spectral envelope. WIthout any filter, the original speech is reconstructed. The bank of filters control the rate with which the spectral envelope is changing. Perceptual experiments estimate speech intelligibilies with limited rate of modulations frequencies in the re-systhesized signal.

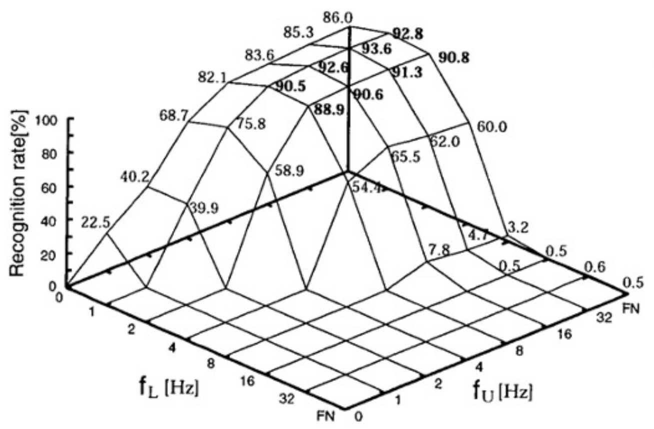


The experiment was run for different combinations of the hight-pass and the low-pass filtering of the modulation spectrum, resulting in the matrix of recognbtion accuraties for different filter combinations. The highest accuracy was obtrained for the full modulation spectrum. The accuracy gradually decreased as the filters were cutting ont the modulation spectrum range. The most noticeable decrease in accuracies was seen when the range between 2 and 8 Hz was eliminates from eny of the combinations. Eventually, the accuracy decreased ti around 35 % when all the changes in the spectral envelope were aliminated and the recogntion relied only on the prosody from the source signal.

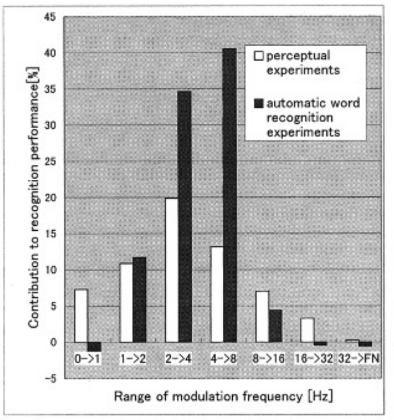


SImilar system but this time without resonstruction the speech is used in automatic speech recognition. Speech recogntion accuracies with modified modulation spectra are evaluated.

Accuracy of ASR with modified modulation spectrum

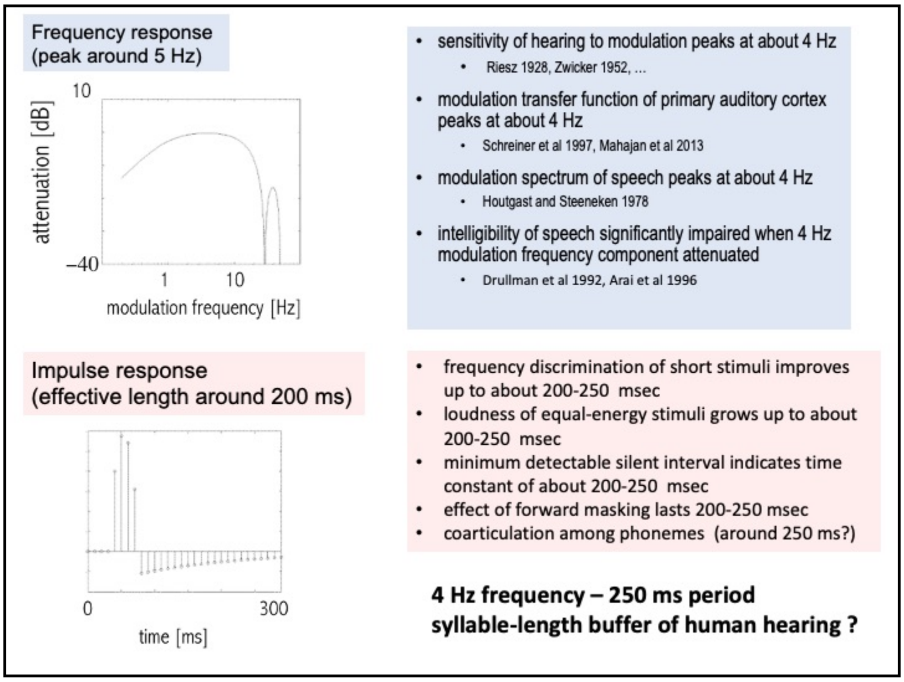


Relative importance of various components of modulation spectrum of speech for speech intelligibility and for ASR



On the relative importance of various components of the modulation spectrum for automatic speech recognition
N Kanedera, T Arai, H Hermansky, M Pavel - Speech Communication, 1999

Results of both the perceptual evaluations and the ASR accuracies indicate dominance of modulation components within the 1-16 Hz range. NOticethat the modulation frequency components between 0 and 1 Hz contribute negatively in ASR, i.e., they reduce the ASR accuracy.

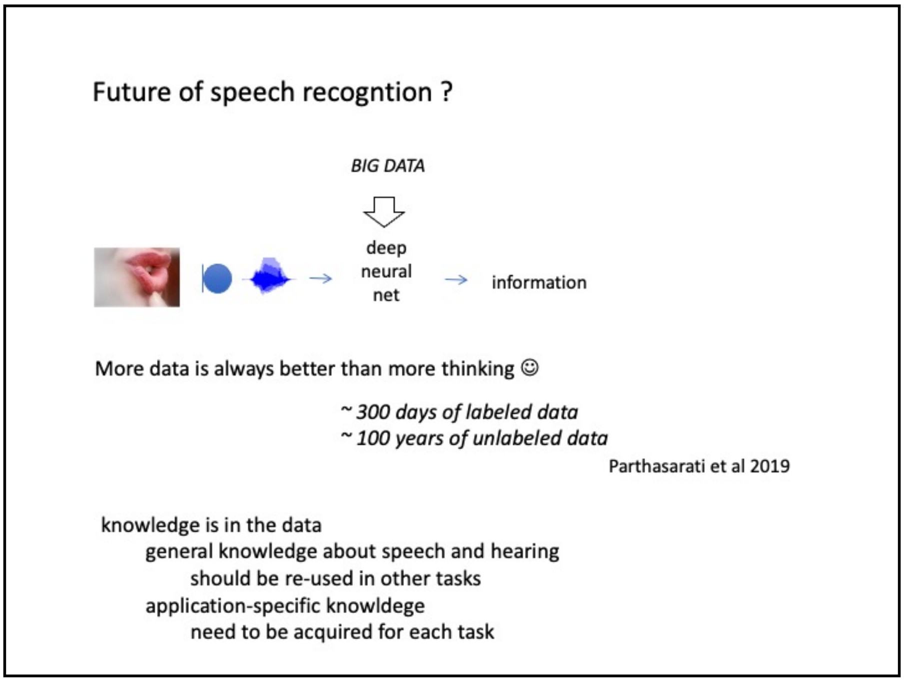


To summarize: The RASTA processing, which involves band-pass filtering of temporal trajectories of spectral energies between two nonlinearities, one being compressive (logarihm) and another expansive (exponential), intially developed to address linear distortions in speech, turmned out to be consistent witth several phenomena observed in human hearing.

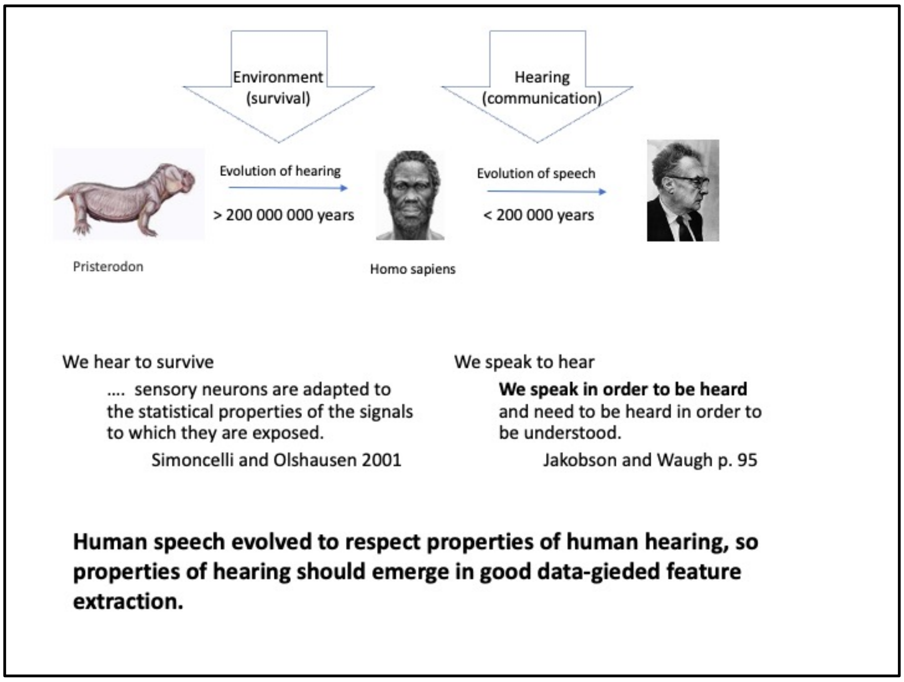
RASTA processing enhances modulations beween 1 and 12 Hz, which is consistent with human sensitivy to modulations (known since 1923). It is also consitent with the estimated temporal properties of hearing, seen in estimated impulse responses of auditory cortical receptive fields. It is also known that in this region, the energy in modulations in speech is the highest. Finally, it can be shown that when this region of speech modulation is removed from the signa, the intelligibility of speech as well as the accuracies of its machine recognition, decrease.

The effective length of the impulse response of the RASTA filter agrees with the concept of temporal buffer in hearing, observed, e.g., in frequency discrimination of short stimuli and in their loudness increase, and with the implied hearing inertia, seen in detection of the gaps in noise and in the temprla forward maskin. RASTA also provides a reasonable model of the forward masking.

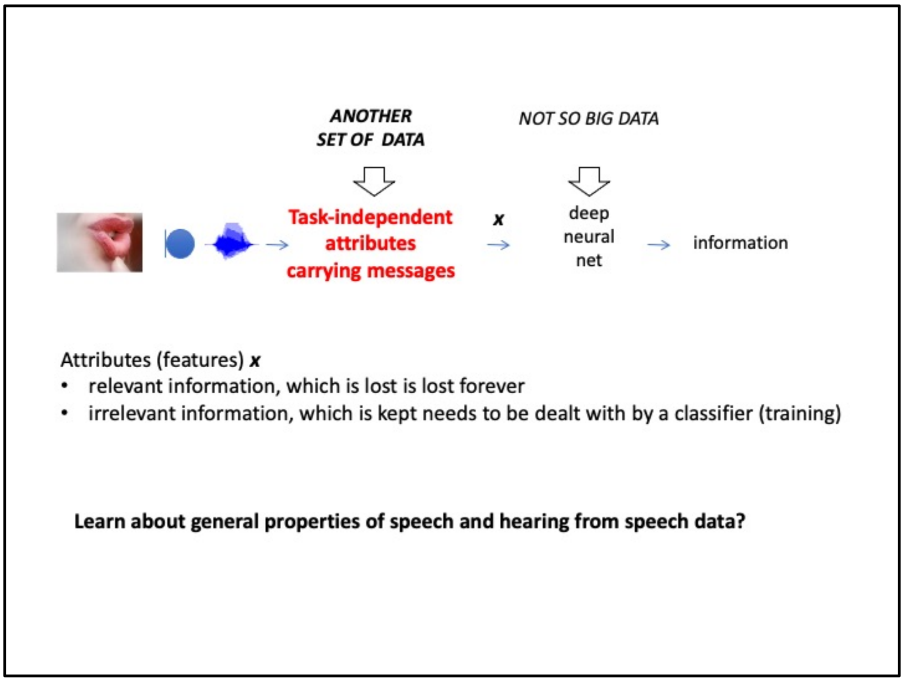
DATA-GUIDED SIGNAL PROCESSING



Current trends in ASR are to learn as much as possible from speech data. One extreme would be to take speech waveform as an input to a large machine learning system which would be trained in one shot to derive the required knowledge from speech. Maybe it is the future of ASR. However, why should we always re-learn the same general knowledgeabout speech and hearing for every new task? If some general knowledge is already hardwired into the sysem, it is possible that the current exterme needs for data may be reduced.

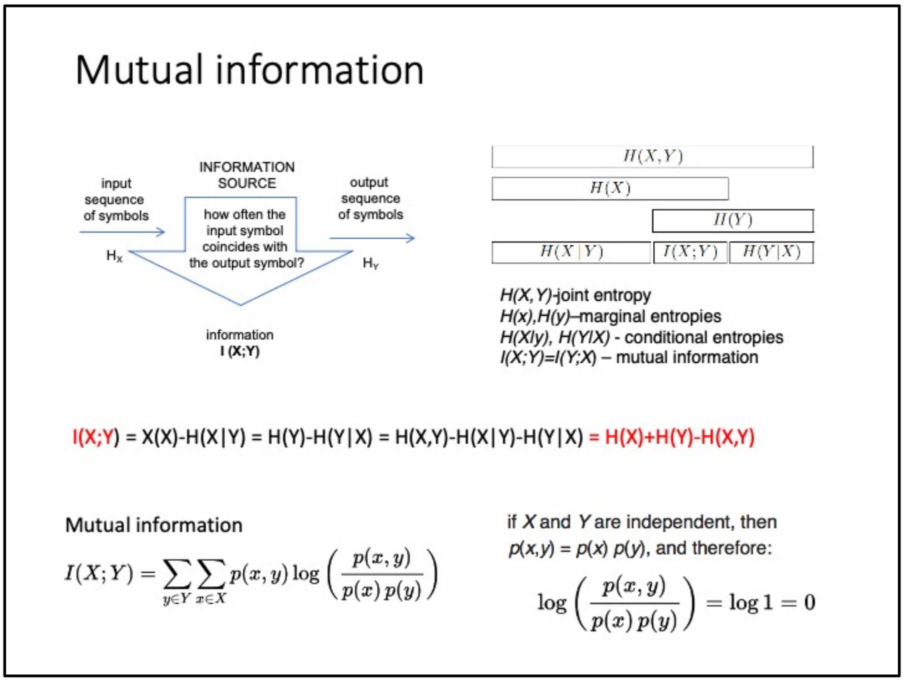


When life emerged from the waters of the primeval ocean, the suddenly earth-bounded animals had to evolve hearing functioning in the air. It took more than 200 million years of adapting the hearing so that it could well process sounds which were of critical interest to animals up to the point that the hearing of *homo sapiens* was very similar to what we have now. The *homo sapiens* is hypothesized that it started communicatimg by signals, which resembled speech. It had lessthat 200 000 years to evolve the speech code which could be well perceived by already existing mammalian hearing. So hearing was much earlier in evolution than speech, so speech had to evolve to respect properties of human hearing.
Subsequently, hearing properties should be found in speech.



Originally, speech signal was used to derive some features ***x***, typically based on short-time fourier spectrum. The feature xtraction is tricky. What is needed there is to alleviate information which is irrelevant for the task (typically recogntion of message in speech) but keep the informatiom which carries the message. Any useful relevant information which is alleviated during the feature extraction is lost forever, any irrelevant information kept needs to dealt with in the subsequent stages of the recogntion, typically by extensive training of the system over sources of the irrelevant information.

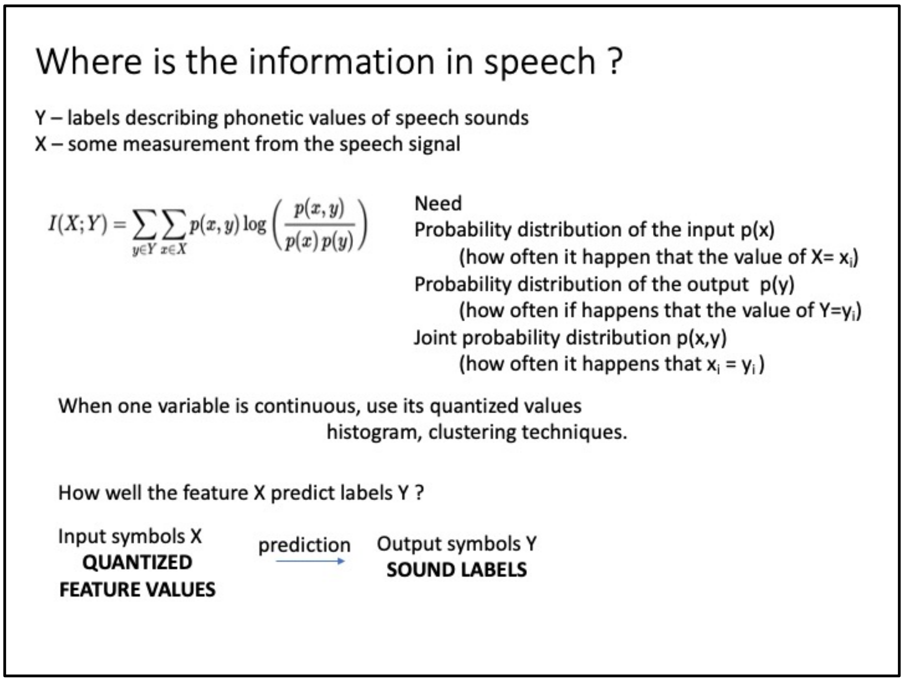
One way of finding out what is relevant and what is irrelevant if to derive the features as a part of the system training. If we are after speech specific but not the task specific information, this can be done on data which are not evenn directly relevant for the final task, therefore a latge abount of task-independent data can be used. Since some informatiopn reduction is already done using another set of data, there is a chance that the final system training may use less data than otherwise required.



When starting to study information transfer using speech (and getting to be interested in recognizing speech by machine) it is reasonable to ask wher is the information about speech sounds in the speech signal?
We recall the information transfer through the system. In one form, it requires to compute the true joint entropy of two variables, X and Y, derived from the confusion matrix, and the maximum joint entrope between these variables under the assumprion that the X and Y are independent. It is related to the concept of the mutual information, which we will be using here.

The ame way, mutual information between two variables, Y and Y, uses similarity of the true joint distribution p(X,Y) and this joint distribution under the assumptin that X and Y are independent, i.e. p(X)p(Y). It requires to compute probability distributions of the input variable, the output variable, and the joint probability distribution of the two variables.
These can be approximated by histograms. To use histograms for approximating probability distributions of continuous varaibles, we will quantize the continuous variables.
The difference between the two (computed useng the so called Kullback-Leibler divergence) is the mutual information between X and Y.

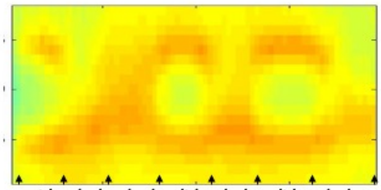
The joint probability distribution p(X,Y) of the input and the output variables can be derived from the confusion matrix.



While the principle is simple, the practice of estimating the multidimensional probability distributions can be difficult for more than one dimensional problems. So let’s decide that we will be primarily interested in finding mutual information between the speech sound labels and **one mesurement** from the time-feature matrix (let’s imagine spectrogram). While recognizing speech sounds from a single signal variable is int the usual practice in speech recogntion, where the whole feature vectors are used for the classification, it is reasonable starting point for discovering where the information is.

First, for discrete variables, the distributions are typically estimated from so called histograms, which cluster the discrete varable into bins, and the counts in the bins are used as estimates of probabilties. Clearly, the shape of the histogram depends on the sizes of bins, which is the art of its own and needs to be decided on first.
The second issue is the dealing with continuois variables, which for the use of histograms as the probability disctribution estimates need to be converted to discrete values through the quantization or clustering.

How well the features predict labels?



Input symbols X **QUANTIZED FEATURE VALUES** $\xrightarrow{\text{prediction}}$ Output symbols Y **SOUND LABELS**

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$$

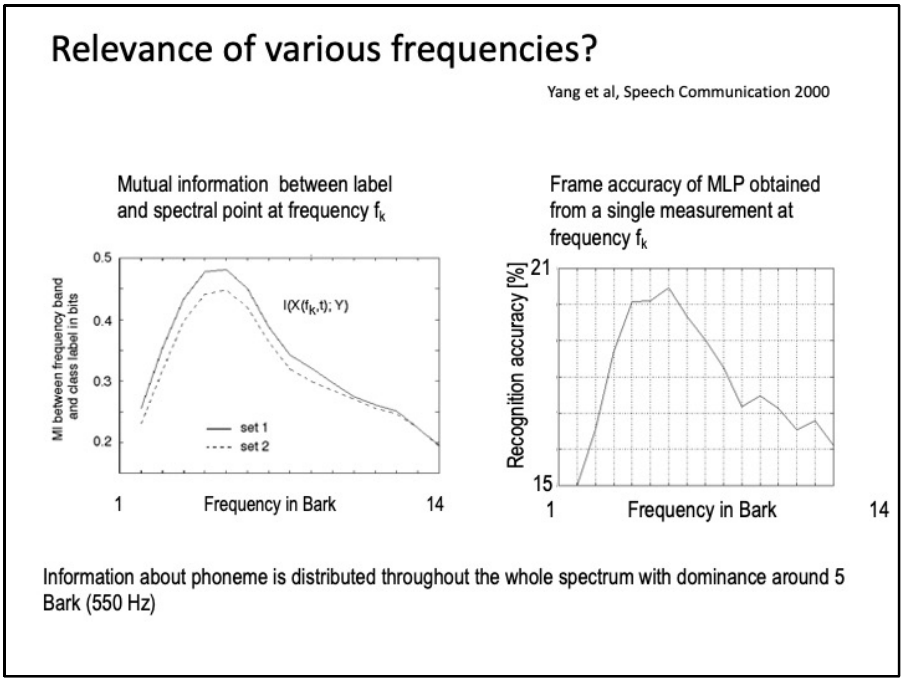
Need p(x) – feature
p(y) – phoneme labels
p(x,y) when x and y coincide

p(x)
Histogram of spectral values at a given frequency band from the whole database
(need for quantization of the continuous variable X – about 50-100 bins?)
p(y)
Histogram of label indices from the whole database
(number of bins = number of phonemes)

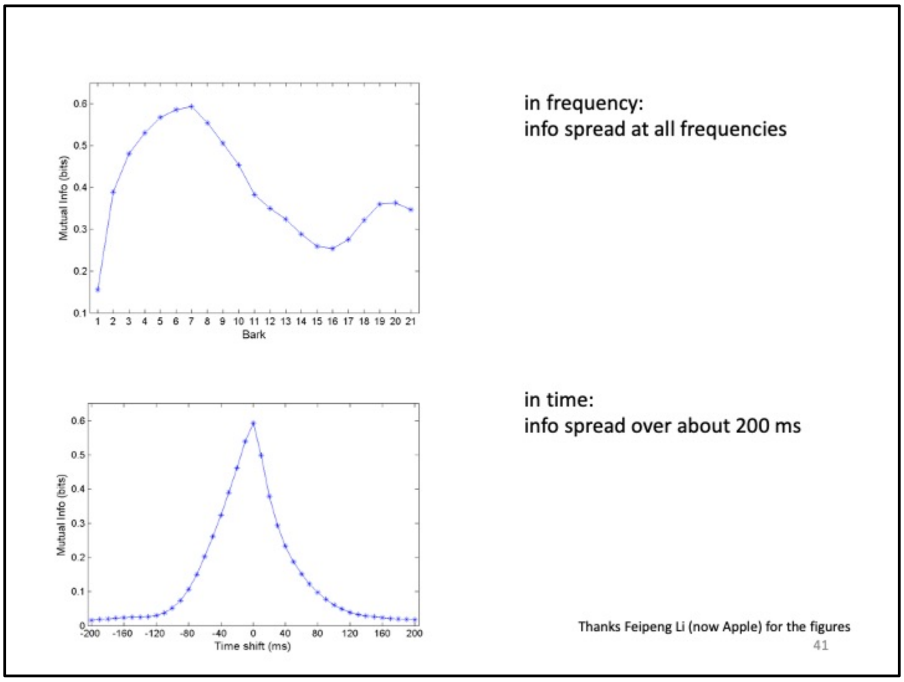
p(x,y)
Histogram (probability confusion matrix) of instances when the given phoneme and the given quantized spectral value coincide)
number of bins = (5-100) x number of phonemes

So after all the simplifications (decision about a single input variable, discretization of the continuous one-dimansional input x, decision about the size of bins in deriving the histograms,..) we can start. The question is:
"How well the input value X predicts the oulput value of the speech class Y" . We have the two-dimensional labeled speech representation (think spectrogram but can be anything else) from which for one pair X,Y one measurement X will be selected for each output label value Y. When we are interested in distribution of the information in frequency, we will form histogram of X at a given frequency for all the data, histogram of Y (will be the same for all X,Y pairs), and the probability confusion matrix between the X at the given frequency and Y. This allows us to compute the mutual information (information transfer) between X an Y. When this is repeated for all frequencies of interest, we get the distribution of mutual information between sound labels and spectral values in frequency.

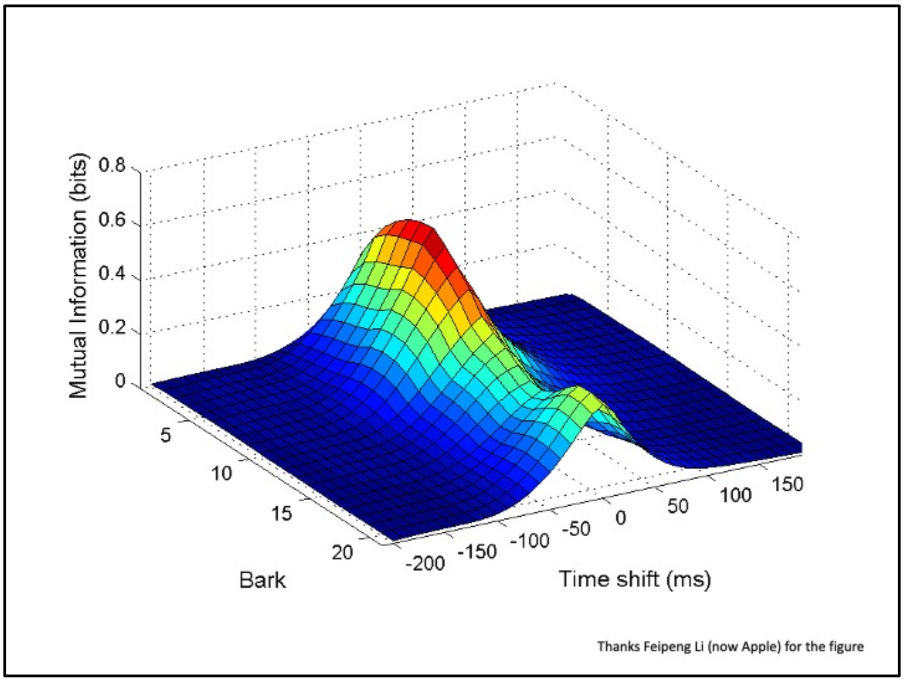
Equally, we may be interested in distribution of the mutual information at each frequency in time.



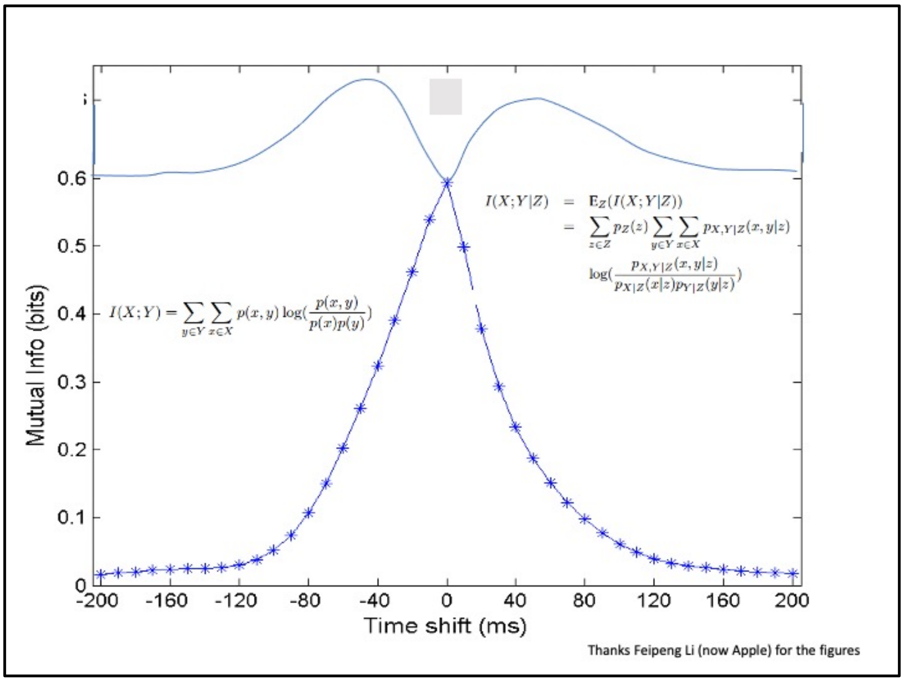
Results of evaluation mutual information between sound labels and measurements in frequency. It indicates that most information in frequency is around 5 Bark (the frequency here is in the auditory-like units mesured by sizes of critical bands). As a "sanity check" ASR recognizing phonems using **one frequency measurement** is show in the right part of the slide, showing that the mutual information evaluation is relevant for ASR.



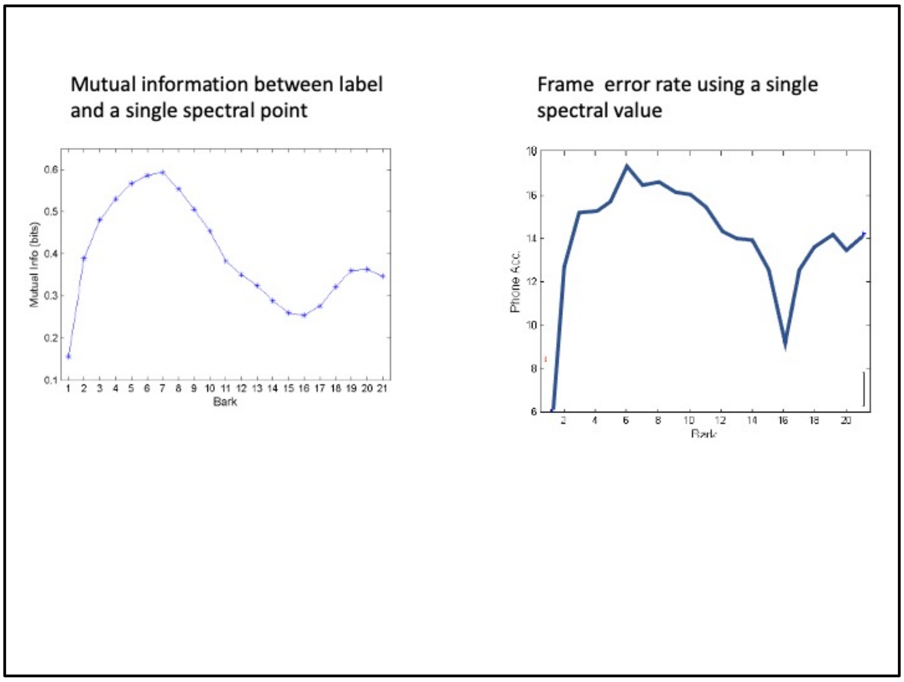
Evaluating information spread in time was done by ofsetting the measurement X with respect to labels. It shows that the highest information is obtained when the measurements and labels coincide. However ofsets as large as 10 msec in each direction still provide some information about the phoneme classe, indication the extent of the coarticulation.



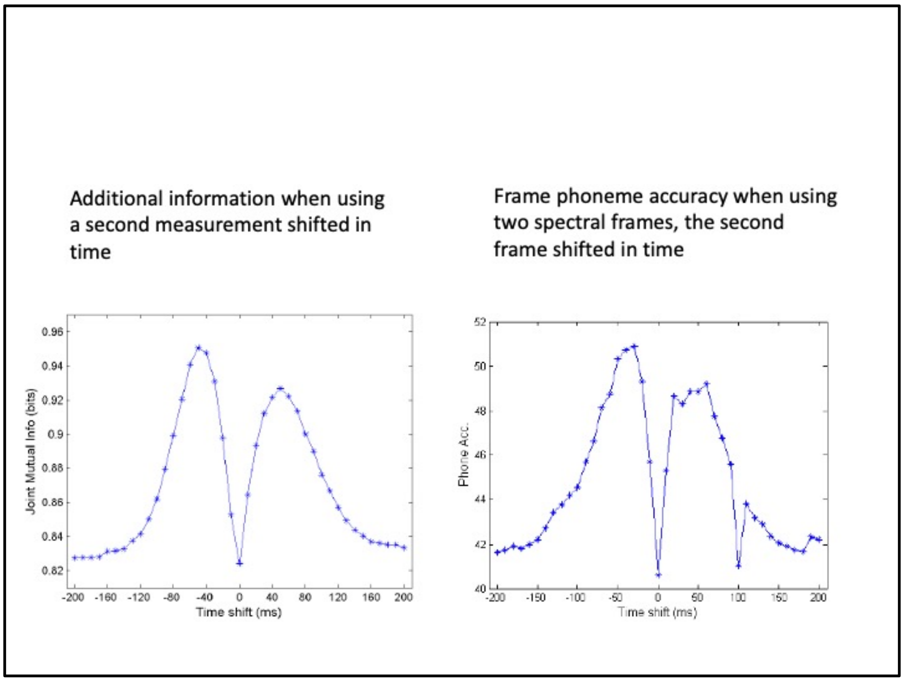
The global view of the information about the sound classes.



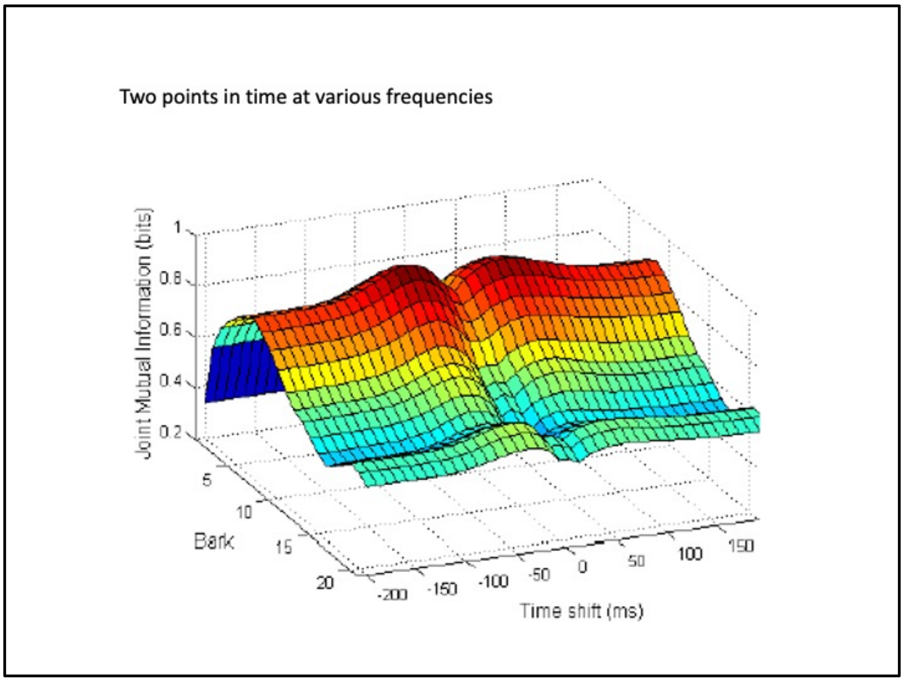
Using mutual information principles, it is possible to evaluate mutual information about sound labels and **two** points in time-frequency plane. This requires to use conditional probabilities, conditionad on one value of X (typically the value which gave the highest MI in the one point experiment. Going for more than two using the current technique would require lartger database of asome altrentaive methos of deriving the multidimensional probability distributions other than histograms. When adding one more measurement in time, the best is to move about 70 ms (to the neighbouring sound) from the current label in both directions. It further support the information spread in time over roughly 200 msec.



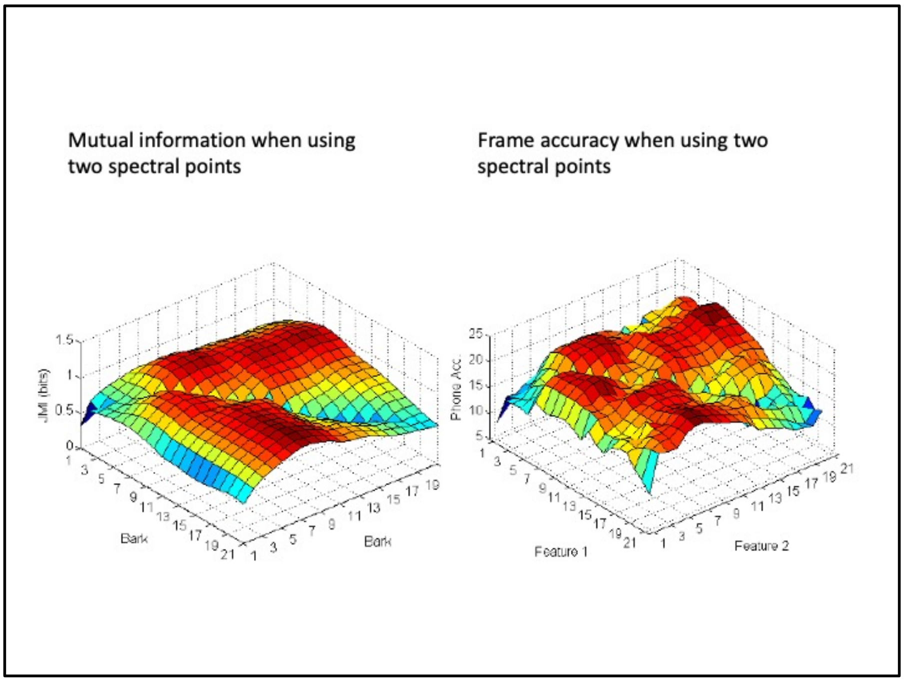
In frequency, the most mutual information (MI) is found around 5-8 Bark. THis is confirmed by results of recogntion experiments using one dimensional spectral vector coming from different frequenciesl of the spectrum. This supports relevanct of the MI measurements for ASR.



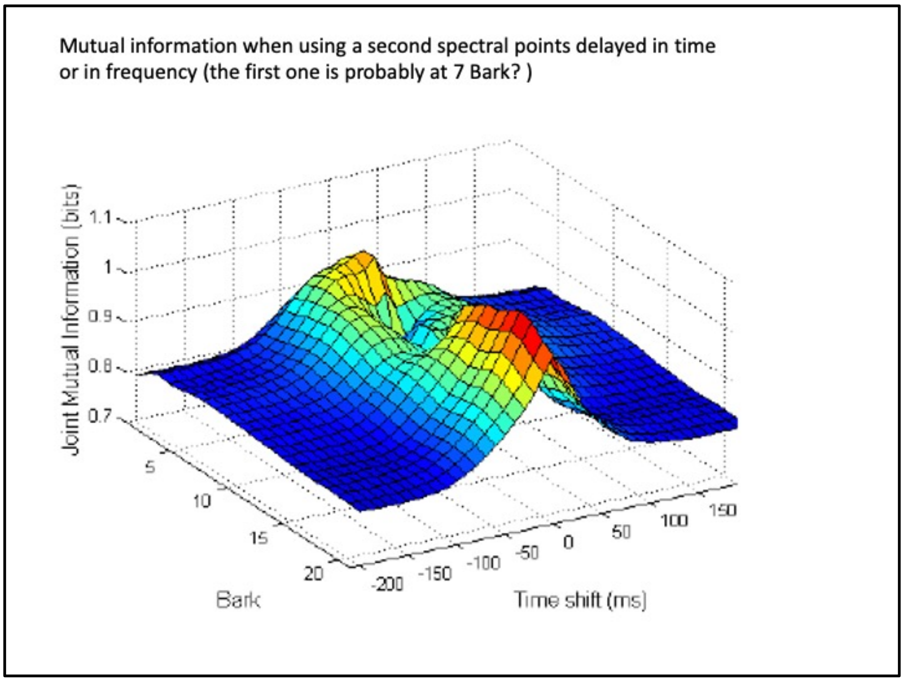
Mutual information from wro measurements in time suggests that for two measurements, the maximum information is available when the second measurement comes from about 70 msec following the first measuresm (or 70 msec before the second measurement), this is from the neighbouring speech sound. This ioservation can be further supported by running speech recogntion experiment using two spectral vectors and evaluate recogntion accuracies as a function of this distance, As seen, the accuracy closely follows the mutual informatio results, further confirming the relevance of the mutual information evaluations fpr ASR.



The two-measutement mutual information evaluations are summarized here in 2-D plot. Interestingly. the most effective second measurement in frequency should be coming from about 3-7 critical bands (Barks) from the first measurement. This would support the 3-4 bank specttral integration in perceptrion of speech whis we have discussesd earlier.



The left part of the slide shown the data of MI evaluation s using two measuremenst spectral mesurements. Recogntion experiments using two spectral measuremenst show similar tendencies.



Two-measurements differing in frequency and in time. The firt measurement was probably around 7 Bark (the accurate description of the experiment done bt Dr. Feipeng Li – now ate Apple) was lpst. The time dealy of the first measurement was zero.